Georges Dionne *Editor*

Handbook of Insurance

Volume I

Third Edition

THE GENEVA ASSOCIATION
INSURANCE FOR A BETTER WORLD

Springer

Handbook of Insurance

Georges Dionne
Editor

Handbook of Insurance

Volume I

Third Edition

Springer

THE GENEVA ASSOCIATION
INSURANCE FOR A BETTER WORLD

Editor
Georges Dionne
Canada Research Chair in Risk Management
and Finance Department
HEC Montréal
Montreal, QC, Canada

ISBN 978-3-031-69560-5 ISBN 978-3-031-69561-2 (eBook)
https://doi.org/10.1007/978-3-031-69561-2

1st edition: © Springer Science+Business Media New York 2000
2nd edition: © Springer Science+Business Media New York 2013
3rd edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

This first volume of the handbook is dedicated to the late Denis Kessler, who made an outstanding contribution to research in insurance economics.

Academic Preface

It is a *tour de force* to provide to the insurance industry and its stakeholders a structured, complete, intelligent, and critical synthesis of insurance economics in the twenty-first century. This is what you have in your hands. This reminds me that when Jacques Drèze, my PhD supervisor, suggested in 1984 that I could write my dissertation on insurance, I scratched my head thinking about who could tell me what was the state of the art in the field, and what papers I could read, at the time when there was no internet, and no Dionne's Handbook! I spent two months in the university's library to go across all issues of the *Journal of Risk and Insurance* that were available there. At that time, papers on insurance with a full-fledged economic modeling were scarce, limited by a few old names like Kenneth Arrow, Jo Stiglitz, Jan Mossin, Agnar Sandmo, or Karl Borch, and by emerging youngsters like Harris Schlesinger, Georges Dionne, Louis Eeckhoudt, Peter Zweifel, Denis Kessler, and Henri Loubergé, scientists that I was lucky enough to meet at the Geneva Association during my first year as a graduate student. At least on the European side of the Atlantic, these were among the people who would revolutionize the field later on, as well documented in this book.

The insurance industry used to be, and in many dimensions still is, perceived as a rusty activity, at best considered as a subsidiary actor of the banking and financial sectors. In academia, insurance economics is rarely structured as a serious research department. It is not always taught to students in business schools, and it is not always present in economics curricula. This is in spite of the crucial role that insurance markets play in our economies. Our world could not exist as we know it without mechanisms to reallocate risks across people and institutions. These mechanisms create value from washing out diversifiable risks, and from transferring those risks that cannot be diversified away to "the market," where less risk-averse participants accept to bear them in exchange for a premium. These risk transfers raise fascinating practical and managerial issues that are now much better understood by my fellow scientists around the world. If Pierre de Fermat and Blaise Pascal paved the way to probability theory more than three centuries ago, the measurement of insurable risk remains a challenge, in particular in the case of infrequent events or of uncertain correlations, as illustrated by the recent financial crises or pandemics. Scientific progresses in stochastic dominance theory and dependence modeling have already percolated to insurance practice and regulation. Beyond this measurement problem the insurance profession also faces issues related to the pricing, the selection, and the prevention of risks. These are problems where the interaction between insurance practices and regulation on one side, and scientific research on the other side, is very lively and active, for the benefit of the common good. This book illustrates this perfectly. New technologies such as artificial intelligence and the digital transformation create new business models for the sector, and a new research playground for insurance economists.

As an active observer of the field over the last four decades, I am indeed amazed by the many intellectual revolutions that hit our community during the period. Although insurance data remain a scarce resource for economists, many breakthroughs have been obtained from bringing our theories to the data. We now much better understand the role that adverse selection and moral hazard have in the

functioning of various insurance markets, such as the critically important health and unemployment insurance. This new knowledge has been key to shape better market regulations around the world. These scientific progresses have been also useful to push the frontier of insurability, for example in the domain of cyber risks, old-age dependence, or climate change. Beyond this data revolution, the insurance economics profession has also experienced progresses by dismantling walls that protected it from other fields, such as asset-pricing and corporate finance and public economics for example. This has enriched the field by reducing its niche atmosphere and by reinforcing its ability to mobilize other competences to answer complex theoretical and policy questions. Insurance economics has also been a core battlefield to apply the revolutionary concepts and methodology invented in contract theory, with its myriad of asymmetric information and incentive problems.

This third edition of the *Handbook of Insurance* should be the bible to anyone who wants to have a deep understanding of the complex challenges faced by insurance and reinsurance markets to create the large social value of risk sharing and risk diversification. The efforts of its editor, Georges Dionne, and of the authors of the chapters, together with those of the generations of economists who have contributed to this literature, must be acknowledged for providing the scientific and professional competences that are summarized in this unique document. In the current context of the emergence of new global and existential risks, the accomplishments contained in this handbook is an asset whose societal value should not be underestimated. From this viewpoint, it is a great honor for me to succeed the late Denis Kessler who wrote the preface of the first two editions of this handbook.

Toulouse, France

Christian Gollier
Director of the Toulouse School of Economics

Industry Preface

It is a pleasure to contribute, on behalf of The Geneva Association, the preface of this third edition of the *Handbook of Insurance*. This collective work not only offers a remarkable synthesis of cutting-edge research in insurance economics but also provides a rare resource, both comprehensive and authoritative, for professionals seeking a deeper understanding of insurance industry fundamentals and emerging trends.

The content of the Handbook reflects the richness and dynamics of the field and underlines the many facets involved in better understanding how insurance works and contributes to society. This handbook delves into the core principles of insurance, the ways it operates, is purchased, distributed, regulated, influenced by asymmetric information, influences behaviors, and much more. The opus also embraces contemporary trends and gives ample room to two major phenomena of utmost relevance for insurance, namely climate change and the development of new digital technologies. Climate risk is challenging the insurance industry. It calls for ways to extend the limits of insurability, for instance through catastrophe bonds and parametric insurance, and to enlarge the value proposition of insurance not only by offering risk transfer services but also risk mitigation and adaptation solutions. Such innovations will be vital for insurers to maintain their relevance for society in the face of increasingly frequent and severe weather-related events. New technologies, and digital technologies in particular, have created new risks which present both challenges and opportunities to insurers. Cyber risk is the most prominent example. But the relevance of new technologies for insurers goes beyond risk taking. Equally important, they have allowed insurers to transform their business model, via insurtech and decentralized finance in particular, and to enhance risk assessment and management, for instance through artificial intelligence and telematics. Finally, this handbook also highlights the importance of data for insurers and how they can harness it to refine the pricing of risks, bridging the gap between theory and practice, and offering timely knowledge.

The insurance landscape continues to evolve and the need for a comprehensive guide to navigate its complexities becomes increasingly vital. The *Handbook of Insurance* pools together the collective expertise of eminent academics who have explored every facet of insurance economics. For industry professionals, this handbook serves as a toolkit, presenting the foundational principles of insurance and state-of-the-art risk management practices. Policymakers and regulators will find a wealth of information to inform decisions, and scholars will discover a rich repository of knowledge that inspires further exploration and research.

The Handbook is also being published at a time when The Geneva Association has just celebrated its 50th anniversary. It is worth remembering that The Geneva Association was established for the purpose of promoting economic research in the sector of risk and insurance. It has been performing this role over the past 50 years, acting as a catalyst for initiating and disseminating research and analysis to advance the important role of insurance in modern societies. In doing so The Geneva Association has made a significant contribution to the evolution of insurance economics, helping

make it into the vast research field it is today. A tangible reflection of this evolution is the fact that two tomes of the Handbook are now necessary to reflect on the richness of insurance economics.

On behalf of The Geneva Association, I extend my gratitude to the contributors. Their intellectual rigor and dedication have shaped this handbook into a cornerstone of insurance knowledge. May it serve as a source of inspiration, fostering ongoing dialogue and innovation within the insurance community and shaping the trajectory of insurance economics for years to come.

Zürich, Switzerland

Jad Ariss
Managing Director, The Geneva Association

Introduction

Overview

This new version of the *Handbook of Insurance* updates this reference document on risk and insurance for professors, researchers, graduate students, regulators, consultants, and practitioners. It surveys the current research, providing references to the main contributions in different fields. The updated version contains 37 chapters in two volumes, written by 66 contributors who have published significant research in their respective fields of expertise. Twenty-four new chapters have been added. Many cover new research subjects developed since 2013, such as climate risk protection (five chapters), cyber risk and cyber insurance, digital insurance, pandemic insurance, insurance fraud detection with machine learning, telematics information and machine learning in insurance pricing, causality in applied research, efficiency of the insurance industry, behavioral insurance, the comparative statics of self-protection, new stochastic dominance rules, enterprise risk management, financial strength rating of insurance firms, financial pricing of insurance, regulation of the US insurance industry, solvency regulation of the European insurance market, insurance in China, the US health insurance market, microinsurance, and private equity in the life insurance market. All other chapters have been updated by the original authors, often with new collaborators.

Each chapter can be read independently of the others. The chapters have been (with very few exceptions) reviewed by at least two anonymous referees.

Background

It was the article "Uncertainty and the Welfare Economics of Medical Care" by Kenneth Arrow (*American Economic Review*, 1963) that first drew my research attention to risk, uncertainty, insurance, and information problems. Arrow's article proposed the first economic theorem showing that full insurance above a deductible is optimal when the premium contains a fixed-percentage loading in the absence of information problems. It also suggested economic definitions of moral hazard and adverse selection and discussed the applicability of these concepts to the medical care market.

I learned about Karl Borch's (*Econometrica*, 1962) contribution, "Equilibrium in a Reinsurance Market," when I started studying equilibrium models. The main result of the Borch theorem is that only aggregate risks matter in a population of risk-averse insurers. Individual risks can be diversified away with reinsurance markets (the reinsurance pool). Aggregate risks—those affecting the economy at large—cannot be fully diversified and must be shared among insurers of the pool. Climate risk management could be a current application of the theorem.

During the 1970s, researchers proposed results relating to optimal insurance coverage, security design, moral hazard, adverse selection, and equilibrium concepts for markets with imperfect information. The 1980s were characterized by several theoretical developments, such as the comparative statics of both insurance demand and prevention models, the consideration of more than one insurance contracting period, commitment in repeated insurance contracts, the presence of many contracting agents, multiple risks, nonexpected utility, and several information problems simultaneously. Other economic and financial issues, such as underwriting cycles, the financial pricing of insurance, insurance distribution, the US liability insurance crisis, insurance regulation, and market retention capacity, were addressed by academics and practitioners during that period. Hierarchical relationships in firms and organizations and organizational forms of insurers were also studied, along with the measurement of insurers' efficiency and the pricing and design of insurance contracts in the presence of many risks in a portfolio setting.

The empirical study of information problems became a real issue in the 1990s and advanced rapidly in the early 2000s. These years were also marked by the development of financial derivative products for risk management in the insurance sector and for the coverage of large losses due to catastrophic events. Alternatives to insurance and reinsurance coverage for these losses were emerging in the financial markets. During the early 2000s, new forms of risk financing and risk engineering products were offered by the financial markets to cover the growing exposure to catastrophic risk, to losses from legal liability, and to other dependent and significant risks. New sources of risk capital were developed, including loss warranties, sidecars, and risk-linked securities like catastrophe bonds. Alternative securitization designs were proposed to expand the market for insurance coverage. Risk engineering and actuarial modeling techniques were developed to improve the solvency measurement and risk management of insurance and reinsurance firms.

New theoretical developments were also published to better understand decision-making under risk and uncertainty. Today, higher-order risk attitudes play a pivotal role in understanding how decisions are made by consumers and managers. Portfolio and insurance models that incorporate background risk offered solutions to different puzzles, partly due to the inadequacy of standard models limited to risk aversion, for complex risky situations. Longevity risk, long-term care insurance, and corporate governance were some of the challenges facing researchers and managers in the insurance industry. Finally, the early 2000s saw the rapid growth of liquidity, credit, and operational risks. Modern society's second major financial crisis was marked not only by systemic risk in the financial markets but also by speculation and lack of transparency in different markets, including the insurance industry, albeit to a lesser extent than the banking industry.

Over the last 10 years, many new developments have been observed in the insurance literature. The regulation of solvency in insurance firms became an important activity after the 2007–2009 financial crisis. Insurers must now revise their risk appetite statements every year, with oversight from new risk committees that are related to the boards of directors of insurers and reinsurers. These committees are made up entirely of independent directors, with at least one member having sufficient knowledge about risk management. New risk management models, such as Own Risk and Solvency Assessment (ORSA), which are more related to market data than to accounting data, are now implemented to compute optimal economic and regulated capital.

Natural catastrophe risk is omnipresent, with a yearly average of 100 billion in insured losses over recent years. These losses have caused some insurers to leave the market, and important questions about the international coverage of these losses via reinsurance have been raised because global warming limits international diversification. The development of alternative financial coverage is a positive step, but it is not very significant, at about 10% of the total reinsurance capital. Moreover, the long run commitment of financial funds is an issue. What will happen if some funds are hit by a "big one"? Will they reinvest? Economic inflation in wages, damages and care costs, social inflation

costs, cyber risk costs, and pandemic costs are becoming the new sources of sizable losses. It is not clear that insurers' traditional risk management tools will be efficient for limiting these losses and expenses, which are often correlated.

The insurance market faces many new innovations that affect the management of insurance and reinsurance companies. Internet and digital insurance are becoming significant management tools for insurance policy management. Machine learning and telematics information are used to reduce moral hazard by improving safety monitoring and fraud auditing. In life insurance, unit-linked and equity-indexed contracts are important innovative products that appeared in recent years. Private equity in life insurance markets is also trending upwards.

These new aspects of the insurance literature are presented in various chapters of this third edition of the Handbook. Below, we outline what is covered in this new edition, following the order of the table of contents. The new edition is split into two volumes.

Content of Volume 1

History

The first chapter "Developments in Risk and Insurance Economics: The Past 50 years", by H. Loubergé and G. Dionne, reviews the development of insurance economics over the past 50 years, first presenting the situation in 1973, and then discussing the developments and new approaches that have flourished since then. Particular attention is given to contributions that appeared after the second edition of the *Handbook of Insurance*, in 2013. The authors argue that these developments were possible because, first, steady advances were made in the economics of risk and uncertainty and in financial theory, and second, the international risk environment has evolved significantly.

Climate Risk Protection

Five chapters on climate risk protection are included. In chapter "Fifty Years of US Natural Disaster Insurance Policy", K. Marcoux and K. R. H. Wagner review current challenges in both public and private natural disaster insurance markets in the USA and how the nature of these challenges has changed over the past 50 years. They discuss how the infrequent and spatially correlated extreme events that distinguish these markets complicate both the supply of and demand for natural disaster insurance.

According to W. J. W. Botzen and P. J. Robinson, the trend of rising costs for extreme weather events poses multiple challenges for the insurance sector. Their chapter "Climate Risk and Insurance", summarizes insurance supply arrangements that currently exist for natural disasters, with a focus on the European market. W. J. W. Botzen and P. J. Robinson discuss how reforms of the supply of insurance should cope with climate change. Empirical evidence of demand for insurance against climate risk is reviewed, with a focus on possible explanations for why this demand is often low in practice. Based on the outcomes of their investigation, key implications for insurers and policymakers are drawn, and an agenda for future research is outlined.

B. Collier and C. Kousky analyze households affected by climate events in chapter "Household Financial Resilience After Severe Climate Events: The Role of Insurance". These households can face significant financing challenges as personal disaster expenses and damages may exceed tens or even hundreds of thousands of dollars. The authors review the existing literature to describe how households fund their recovery, focusing on the role of insurance. While insurance is a key financial

tool for managing such large losses, many households are uninsured and rely on savings, credit, federal assistance programs, and help from family and friends. Their contribution highlights the importance of disaster insurance and motivates a comprehensive view of financial recovery, where income and race play critical roles.

M. Kelly and Z. Peng provide, in chapter "Insuring Wildfires Globally", an overview of the increasing risks posed by wildfires globally and the role of insurance in protecting against wildfire losses, including to agriculture and timber, structures and infrastructures, revenue due to business interruptions, tourism dollars, as well as significant indirect losses. Wildfire smoke also has important health impacts. Since wildfires are difficult to avoid, investments in suppression as well as mitigation by communities and individual property owners are essential, and insurance can play a role in incentivizing mitigation activities. Given the recent global increase in losses, significant concerns have been raised about the insurability of such losses in the future, especially in fire-prone California and Australia. Some recommendations to expand insurance market capacity and provide sufficient compensation after a wildfire loss are provided.

In chapter "Catastrophe Bonds", P. Barrieu, A. Braun, and D. Makariou describe catastrophe bonds as one of the most significant innovations in the risk transfer of catastrophic risks, including climate risks, over the last three decades. They contextualize catastrophe bonds within the broader realms of alternative risk transfer and alternative capital, introduce the relevant terminology, explain the main structural and market characteristics, discuss the state of the art in catastrophe bond pricing, and provide an outlook for further developments in this area.

New Challenges

M. Eling provides a comprehensive analysis of the insurability of cyber risks by examining both the academic literature and industry studies in chapter "Cyber Risk and Cyber Insurance". He addresses five key research questions, including the measurement of cyber risk, the economic impact of cyber risk, comparisons with other types of risk, an analysis of the current cyber insurance market, and a discussion of insurability limitations that are currently hindering market growth. The underdevelopment of the cyber insurance market can be attributed to four distinct characteristics of cyber risk: heavy tail distributions, nonlinear dependencies leading to potential accumulation of total risk, model/parameter risk, and asymmetric information. There are also several demand-side factors impeding the growth of the market.

A. Braun and N. Haeusle explore the digital transformation in the insurance industry in chapter "Digital Insurance and InsurTech". They examine the transformation of the insurance value chain, driven by advanced technologies and the emergence of innovative business models, including digital attackers, ecosystem orchestrators, prevention partners, online mutual aid, and decentralized insurance. The chapter also highlights the crucial role of InsurTech startups in accelerating the digital transformation and fostering innovation in the sector.

G. Kraut, P. La Bonté, and A. Richter provide an introduction to pandemic risk, explore the challenges associated with risk transfer solutions, and address other risk management strategies in chapter "Pandemic Risk Management and Insurance". After a general overview of outbreaks and predictions, they analyze the insurability of pandemic risks. The impact of COVID-19 on the insurance industry is discussed for different lines of business. The authors propose stylized paths for insurance market developments. Based on experience from previous pandemics, this chapter outlines a possible path forward for the management of pandemic risk, addressing the capacity and affordability issues of risk transfer solutions.

Life Insurance and Social Insurance

N. Gatzert and H. Schmeiser give an overview of life insurance financial products in chapter "Life Insurance Financial Products". After an outline of the global life insurance market, they present different forms of selected traditional and innovative life insurance financial products and their main characteristics. Unit-linked and equity-indexed contracts have been the basis for most innovative products in recent years. Basic aspects of the modeling, valuation, and risk management of unit-linked life insurance contracts are presented, using two forms of investment guarantees (interest guarantees and look-back guarantees) and different underlying investment strategies for illustrative purposes. They also outline the advantages and disadvantages of life insurance products.

In chapter "Private Equity Investments in the Life Insurance Industry", A. Kartasheva summarizes the latest trends in private equity investment in the life insurance industry, reviews the emerging research on the role of private equity in insurance, and discusses several new regulatory aspects. The author is of the opinion that the impact of transaction forms (namely, portfolio transfer, legal entity sale, or funded reinsurance) on the stakeholders and policyholders is rather unclear. She concludes that the empirical assessment of the outcomes of various forms is a promising avenue for future research. Another relevant topic for future research is the role of the regulatory environment for the development of private equity ownership in life insurance in the USA and Europe.

For P. Zweifel, in his chapter "Insurance: Private or Social?", the last few decades have seen a substantial change in the importance of private (PI) relative to social insurance (SI), with the latter gaining the advantages. The efficiency view of SI explains the market failures of PI. A benevolent government is introduced in the discussion to determine the optimal division of labor between PI and SI. The exposition turns to the public choice view, which emphasizes the interests of risk-averse voters even with above-average wealth in redistribution through SI. This view predicts a crowding out of PI by SI. The author turns to normative issues by proposing a test that indicates whether, and in which insurance lines, the division of labor between PI and SI could be improved.

Empirical Models

G. Dionne discusses the difficult question of measuring causality effects in empirical analyses, with applications to asymmetric information and risk management, in chapter "Causality in Empirical Analyses with Emphasis on Asymmetric Information and Risk Management". He claims that the measurement of policy effects must be causal. After presenting the main frameworks for causality analysis, including instrumental variable, difference-in-differences, and the generalized method of moments, he analyzes the following questions: Does risk management affect firm value and risk? Do we face a moral hazard problem in insurance data? How can we separate moral hazard from adverse selection and asymmetric learning? Is liquidity creation a causal factor for reinsurance demand? He shows that residual information problems are often present in different markets, while risk management may increase firm value when appropriate methodologies are applied. Liquidity creation is a causal factor for reinsurance demand by small insurers.

In chapter "Econometrics of Insurance Based on Telematics Information and Machine Learning", M. Guillén discusses how telematics data provide real-time information that can be used to develop new insurance pricing models in all lines of insurance. She then focuses on automobile insurance, where telematics data capture information on such variables as driver speed, acceleration, braking, and cornering. The preprocessing of telematics data and their synchronization with claims records are discussed, as well as how machine learning algorithms can identify patterns in telematics data that are not otherwise immediately apparent and serve to create risk scores. Regulatory constraints,

contract choice, ethical issues, and practical insights regarding the adoption of telematics technology in motor insurance are presented.

D. Banulescu-Radu and Y. Kougblenou provide in chapter "Data Science for Insurance Fraud Detection: A Review" an overview of the main models used to prevent and detect insurance fraud, as well as the main challenges faced by modelers as part of this process. Special attention is given to evaluating gains in terms of statistical predictive performance when using machine learning models over traditional econometric models. The authors illustrate empirically how different estimation models or algorithms and their cost-sensitive counterparts can improve efficiency. Results show that machine learning techniques perform better statistically and can also be more effective than standard approaches in reducing fraud-related costs. However, these techniques must be accompanied by expert knowledge and human analysis to ensure accurate and reliable fraud detection in the insurance industry.

J. Pinquet analyzes in chapter "Nonnegative Second-Order Semiparametric Analysis and Experience Rating in Non-life Insurance" the role of memory in longitudinal non-life insurance data, and especially in automobile insurance. Memory in bonus-malus scales, in real-world rating structures, and in statistical models on insurance data is related to the magnitude of no-claim discounts. The chapter also studies point-record driving licenses based on traffic violations, which are applied to create incentives for safe driving and reduce moral hazard. The incentive properties are discussed, as the future of point-record driving licenses in a context where the driver's behavior would be recorded in real time.

In chapter "Efficiency and Productivity in the Insurance Industry Using Frontier Methodologies", M. Rubio-Misas reviews the research on efficiency and productivity in the insurance industry utilizing frontier methodologies that have been published from 2012 to 2022. Forty-six empirical papers are surveyed. In these analyses, econometric and nonparametric approaches have been used to estimate frontiers. Data envelopment analysis (DEA) is the most frequently utilized method. The reviewed papers are grouped into 12 different application areas where the main analyzed issues and/or tested hypotheses as well as the principal findings are discussed.

Acknowledgments

I want to thank Claire Boisvert for her immense contribution to the preparation of this third edition. Without her dedication to the project, the production of this version would have been less dynamic and less pleasant. I also want to thank all the collaborators and referees for their competent and timely input. Their enthusiasm for participating in the book helped me to maintain focus on the different steps of the project. The Geneva Association and Social Sciences and Humanities Research Council (SSHRC) Canada provided financial support to the project. The professional collaboration of Saveetha Balasundaram, Lorraine Klimowich, Carolyn Kousky, Sinduja Ravi, and Kasthuri G, from Springer, is also acknowledged, as well as the cooperation of Christophe Courbage for managing important contacts.

I cannot end this introduction without mentioning the great contribution made to the academic research in insurance economics by the late Denis Kessler, a former CEO and president of SCOR. At various stages of his great career, he created and founded research chairs in insurance economics, financed research projects, and established the SCOR Foundation for Science. He also participated in varied important debates, including the development of optimal pension plans, the link between savings and economic development, and the significance of systemic risk in the insurance industry during the 2007–2009 financial crisis. He authored many books and wrote the preface of the first two editions of this Handbook. More information on his momentous career and generosity to the academic world is available at https://www.risksresearch.com.

Georges Dionne, HEC Montréal
7 February 2024

Contents

Part I History

Developments in Risk and Insurance Economics: The Past 50 Years 3
Henri Loubergé and Georges Dionne

Part II Climate Risk Protection

Fifty Years of U.S. Natural Disaster Insurance Policy 55
Kendra Marcoux and Katherine R. H. Wagner

Climate Risk and Insurance 81
W. J. Wouter Botzen and Peter J. Robinson

Household Financial Resilience After Severe Climate Events: The Role of Insurance 99
Benjamin Collier and Carolyn Kousky

Insuring Wildfires Globally 123
Mary Kelly and Zhe Peng

Catastrophe Bonds 169
Pauline Barrieu, Alexander Braun, and Despoina Makariou

Part III New Challenges

Cyber Risk and Cyber Insurance 199
Martin Eling

Digital Insurance and InsurTech 225
Alexander Braun and Niklas Haeusle

Pandemic Risk Management and Insurance 251
Gunther Kraut, Paulina La Bonté, and Andreas Richter

Part IV Life Insurance and Social Insurance

Life Insurance Financial Products 289
Nadine Gatzert and Hato Schmeiser

Private Equity Investments in the Life Insurance Industry 311
Anastasia Kartasheva

Insurance: Private or Social? 337
Peter Zweifel

Part V Empirical Models

Causality in Empirical Analyses with Emphasis on Asymmetric Information and Risk Management 361
Georges Dionne

Econometrics of Insurance Based on Telematics Information and Machine Learning 401
Montserrat Guillén

Data Science for Insurance Fraud Detection: a Review 417
Denisa Banulescu-Radu and Yannick Kougblenou

Nonnegative Second-Order Semiparametric Analysis and Experience Rating in Non-life Insurance 447
Jean Pinquet

Efficiency and Productivity in the Insurance Industry Using Frontier Methodologies 463
María Rubio-Misas

Index 505

Contributors of the Third Edition

Denisa Banulescu-Radu Université d'Orléans, Orleans, France

Etti Baranoff Virginia Commonwealth University, Richmond, VA, USA

Pauline Barrieu London School of Economics and Political Science (LSE), London, UK

Daniel Bauer University of Wisconsin-Madison, Madison, WI, USA

Carole Bernard Grenoble Ecole de Management, Grenoble, France

W. J. Wouter Botzen Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Narjess Boubakri American University of Sharjah, Sharjah, United Arab Emirates

Alexander Braun University of St. Gallen, St. Gallen, Switzerland

Patrick Brockett University of Texas at Austin, Austin, TX, USA

Benjamin Collier Temple University, Philadelphia, PA, USA

Christophe Courbage Geneva School of Business Administration, Geneva, Switzerland

Keith J. Crocker Pennsylvania State University, University Park, PA, USA

Georges Dionne Canada Research Chair in Risk Management and Finance Department, HEC Montréal, Montreal, QC, Canada

René Doff University of Amsterdam, Amsterdam, The Netherlands

David M. Dror Micro Insurance Academy, New Delhi, India

David L. Eckles University of Georgia, Athens, GA, USA

Martin Eling University of St. Gallen, St. Gallen, Switzerland

Paul Embrechts ETH Zürich, Zurich, Switzerland

Hanming Fang University of Pennsylvania, Philadelphia, PA, USA

Nathalie Fombaron Université Paris Nanterre, Nanterre, France

Nadine Gatzert Friedrich Alexander Universität, Erlangen, Germany

Linda Golden University of Texas at Austin, Austin, TX, USA

Martin F. Grace University of Iowa, Iowa, IA, USA

Montserrat Guillén Universitat de Barcelona, Barcelona, Spain

Martin Halek University of Calgary, Calgary, AB, Canada

Niklas Haeusle Leipzig University, Leipzig, Germany

Glenn W. Harrison Georgia State University, Atlanta, GA, USA

James I. Hilliard Temple University, Philadelphia, PA, USA

Marius Hofert University of Hong Kong, Pokfulam, Hong Kong

Robert E. Hoyt University of Georgia, Athens, GA, USA

Rachel J. Huang National Central University, Taoyuan, Taiwan

Anastasia Kartasheva School of Finance, University of St. Gallen, and Swiss Finance Institute, St. Gallen, Switzerland

Mary Kelly Wilfrid Laurier University, Waterloo, ON, Canada

Yannick Kougblenou University d'Orléans, Orleans, France

Carolyn Kousky Environmental Defense Fund, New York, NY, USA

Gunther Kraut Munich Re, Munich, Germany

Paulina La Bonté Ludwig-Maximilians-Universität München, Munich, Germany

André P. Liebenberg University of Mississippi, Oxford, MS, USA

Henri Loubergé University of Geneva, Geneva, Switzerland

Despoina Makariou University of St. Gallen, St. Gallen, Switzerland

Kendra Marcoux University of California, Berkeley, Berkeley, CA, USA

Wanda Mimra ESCP Business School, Paris, France

Alvaro Parra University of British Columbia, Vancouver, BC, Canada

Zhe Peng Property and Casualty Insurance Compensation Corporation, Toronto, ON, Canada

Richard Peter University of Iowa, Iowa, IA, USA

Richard D. Phillips Georgia State University, Atlanta, GA, USA

Pierre Picard École Polytechnique, Paris, France

Jean Pinquet Université Paris Nanterre, Nanterre, France

Béatrice Rey Université Lumière Lyon 2, Lyon, France

Andreas Richter Ludwig-Maximilians-Universität München, Munich, Germany

Peter J. Robinson Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Casey Rothschild Wellesley College of Business, Wellesley, MA, USA

María Rubio-Misas Universidad de Málaga, Malaga, Spain

Hato Schmeiser University of St. Gallen, St. Gallen, Switzerland

Arthur Snow University of Georgia, Athens, GA, USA

Sharon Tennyson Cornell University, Ithaca, NY, USA

Nicolas Treich Toulouse School of Economics, Toulouse, France

Larry Y. Tzeng National Taiwan University, Taipei, Taiwan

Pascale Valéry HEC Montréal, Montreal, QC, Canada

Katherine R. H. Wagner University of British Columbia, Vancouver, BC, Canada

Ralph A. Winter University of British Columbia, Vancouver, BC, Canada

Xian Xu Fudan University, Shanghai, China

Charles C. Yang Florida Atlantic University, Boca Raton, FL, USA

George Zanjani University of Alabama, Tuscaloosa, AL, USA

Boyi Zhuang University of Alabama, Tuscaloosa, AL, USA

Peter Zweifel University of Zurich, Zurich, Switzerland

Referees of the Third Edition

Carlo Acerbi École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Andrew Austin Congressional Research Service, Washington, DC, USA

Denisa Banulescu-Radu Université d'Orléans, Orleans, France

Levon Barseghyan Cornell University, Ithaca, NY, USA

Antoine Bommier ETH Zürich, Zurich, Switzerland

Patricia Born Florida State University, Tallahassee, FL, USA

Alexander Braun University of St. Gallen, St. Gallen, Switzerland

Marielle Brunette National Research Institute for Agriculture, Food and Environment (INRAE), BETA, Nancy, France

James Carson University of Georgia, Athens, GA, USA

Arthur Charpentier Université du Québec à Montréal, Montreal, QC, Canada

Benjamin Collier Temple University, Philadelphia, PA, USA

Christophe Courbage Geneva School of Business Administration, Geneva, Switzerland

Pamela Drake James Marison University, Harrisonburg, VA, USA

Evan Eastman Florida State University, Tallahassee, FL, USA

Hanming Fang University of Pennsylvania, Philadelphia, PA, USA

Guangyuan Gao Renmin University of China, Beijing, China

Christian Genest McGill University, Montreal, QC, Canada

Marc Guertler Technische Universität Braunschweig, Braunschweig, Germany

Montserrat Guillén Universitat de Barcelona, Barcelona, Spain

Adlane Haffar University of Science and Technology Houari Boumediene, Algiers, Algeria

Robert Hartwig University of South Carolina, Columbia, SC, USA

Peter Hieber University of Lausanne, Lausanne, Switzerland

Annette Hofmann University of Cincinnati, Cincinnati, OH, USA

Michael Hoy University of Guelph, Guelph, ON, Canada

Ruo (Alex) Jia Peking University, Beijing, China

Kwangmin Jung Pohang University of Science and Technology (POSTECH), Pohang, South Korea

Kyeonghee Kim Florida State University, Tallahassee, FL, USA

Tyler Leverty University of Wisconsin, Madison, WI, USA

Jingyuan Li Lingnan University, Tuen Mun, Hong Kong

André P. Liebenberg University of Mississippi, Oxford, MS, USA

Tong Liu MIT Sloan School of Management, Cambridge, MA, USA

Ying Liu Shandong University, Jinan, China

Pingyi Lou Fudan University, Shanghai, China

Alexis Louaas École Polytechnique, Paris, France

Yang Lu Concordia University, Montreal, QC, Canada

James R. Meldrum United States Geological Survey, Fort Collins, CO, USA

Mario Menegatti Università di Parma, Parma, Italy

Andreas Milidonis University of Cyprus, Nicosia, Cyprus

Steve Miller University of South Florida, Sarasota, FL, USA

Thorsten Moenig Temple University, Philadelphia, PA, USA

Donald Pagach North Carolina State University, Raleigh, NC, USA

Ahyan Panjwani Federal Reserve Board, Washington, DC, USA

Sojung Park Seoul National University, Seoul, South Korea

Pierre Picard École Polytechnique, Paris, France

Thierry Post Nazarbayev University, Astana, Kazakhstan

Dick Rae Fellow of the Institute of Actuaries, London, UK

Eva Raiber Aix Marseille Université, Aix en Provence, France

María Rubio-Misas Universidad de Málaga, Malaga, Spain

Bernard Salanié Columbia University, New York, NY, USA

François Salanié Toulouse School of Economics, Toulouse, France

Evan Saltzman Emory University, Atlanta, GA, USA

Tess Scharlemann Federal Reserve Board, Washington, DC, USA

Jörg Schiller University of Hohenheim, Stuttgart, Germany

Arthur Snow University of Georgia, Athens, GA, USA

Didier Sornette SUSTech, Shenzhen, China

Ivan Sosa Universidad Camilo José Cela, Madrid, Spain

Sharon Tennyson Cornell University, Ithaca, NY, USA

Andrea Tracogna University of Trieste, Trieste, Italy

Stéphane Verani Federal Reserve Board, Washington, DC, USA

Stéphane Villeneuve Toulouse School of Economics, Toulouse, France

Katherine R. H. Wagner University of British Columbia, Vancouver, BC, Canada

Po-Lin Wang University of South Florida, Sarasota, FL, USA

Joakim Weill Federal Reserve Board, Washington, DC, USA

Mario Valentin Wüthrich ETH Zürich, Zurich, Switzerland

Meryem Yankol Schalck IPAG Business School, Paris, France

Raphael Zerbato HEC Montréal, Montreal, QC, Canada

Kai Zhao University of Connecticut, Storrs, CT, USA

Wei Zheng Peking University, Beijing, China

Part I

History

Developments in Risk and Insurance Economics: The Past 50 Years

Henri Loubergé and Georges Dionne

1 Introduction

In the early 1970s, some 50 years ago, the economics of risk and insurance was still embryonic. Indeed, when the International Association for the Study of Insurance Economics (known as the "Geneva Association") was founded in 1973, one of its promoters' main goals was to foster the development of risk and insurance education in economics curricula. In particular, there was a clear need to develop an understanding of risk and insurance issues among the future partners of the insurance industry. It also seemed necessary to attract the attention of economists to risk and insurance as a stimulating and promising research field.

At that time, some attempts to link insurance to general economic theory had already been made, but they were still scarce. The books written by Pfeffer (1956), Mahr (1964), Greene (1971), and Carter (1972), or the one edited by Hammond (1968), tried to bridge this gap. Corporate risk management started to be considered seriously as a branch of study, at least in the United States. See Mehr and Hedges (1963) and Greene (1973) for early references. The main obstacle was obvious: traditional economic theory was based on the assumption of perfect knowledge—with some ad hoc departures from this assumption, as in the theory of imperfect competition or in Keynesian macroeconomics. In order for risk and insurance issues to be integrated into general economics, the

This survey is the revised and updated version of earlier surveys published by Henri Loubergé as "Risk and Insurance Economics 25 Years After" and "Developments in Risk and Insurance Economics: The Past 40 Years," respectively, in *The Geneva Papers on Risk and Insurance—Issues and Practices* (No. 89, October 1998, pp. 540–567) and in the *Handbook of Insurance*, Second edition, G. Dionne (Ed.), Springer, 2013, Chapter 1, pp. 1–40. We thank Louis Eeckhoudt, Harris Schlesinger, and Richard Peter for their comments on successive versions and Claire Boisvert for the preparation of this version. The usual disclaimer applies.

H. Loubergé (✉)
GSEM – Geneva School of Economics and Management, Geneva, Switzerland
e-mail: henri.louberge@unige.ch

G. Dionne
Canada Research Chair in Risk Management and Finance Department, HEC Montréal, Montreal, QC, Canada
e-mail: georges.dionne@hec.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_1

theory of risk had to develop and gain a position at the heart of economic theory. The foundations were already at hand: the Von Neumann and Morgenstern (1947) and Savage (1954) theories of behavior under uncertainty, the Friedman and Savage (1948) application to risk attitudes, Pratt's (1964) analysis of risk aversion, Rothschild and Stiglitz's (1970) characterization of increases in risk, and the Arrow (1953) and Debreu (1959) model of general equilibrium under uncertainty. These approaches had already started to bring about a first revolution in the study of finance, with the Markowitz (1959) model of portfolio selection and the Sharpe (1964)–Lintner (1965)–Mossin (1966) model of equilibrium capital asset pricing (the CAPM). With the benefit of hindsight, we now know that they provided the starting point for the accomplishment of one of the Geneva Association's long-term objectives: the integration of risk and insurance research into mainstream economic theory. The papers presented for the 50th anniversary of the European Group of Risk and Insurance Economists (EGRIE), founded by the Geneva Association, provide evidence of this achievement. See, in particular, Eling et al. (2024), Gollier (2024), Peter (2024), Rothschild (2024), and Harrison (2024).

The purpose of this chapter is to first remind the reader of the situation of insurance economics in 1973 (Sect. 2) and then to summarize its major developments since then in the three main areas of investigation that could be defined at that time: optimal insurance and protection (Sect. 3), market equilibrium under asymmetric information (Sect. 4), and insurance market structure (Sect. 5). The sixth section reviews financial economics by focusing on the new approaches that resulted from the growing integration of insurance and finance. The seventh section concludes by discussing recent developments in insurance markets. The discussion is mainly concentrated on risk and insurance economics issues, as they relate to property-liability insurance. Due to limitations of space and time, two important related topics were omitted from this survey: health economics and social security. Additionally, life insurance is covered only partially.

2 Insurance Economics in 1973

In 1973, the economic theory of insurance had already started to develop from six seminal papers: Borch (1962), Arrow (1963), Mossin (1968), Akerlof (1970), Ehrlich and Becker (1972), and Joskow (1973).[1] All these articles were based on the expected utility paradigm. Following these articles, a whole series of important contributions were published, signaling that the development of an economic theory of risk and insurance was underway.

2.1 Borch (1962)

In his 1962 *Econometrica* article, "Equilibrium in Reinsurance Markets," Karl Borch showed how Arrow's (1953) model of general equilibrium under uncertainty could be applied to the problem of risk sharing among reinsurers. But generations of economists later learned that this insurance application had far-reaching implications for the general economy.[2] In 1953, Arrow showed that financial markets provide an efficient tool to reach a Pareto optimal allocation of risks in the economy.

[1] Note that three of these seven authors, George Akerlof, Kenneth Arrow, and Gary Becker, later received the Nobel Prize in Economic Sciences.

[2] See Gollier (1992) for a review of the economic theory of risk exchanges, Drèze (1979) for an application to human capital, and Drèze (1990) for an application to securities and labor markets.

Nine years later, Borch's theorem[3] demonstrated how the mechanism could be organized in the practice of insurance.

The main argument is the following. In a population of risk-averse agents, only social risks matter. Individual risks do not really matter because they can be diversified away using the insurance markets (the reinsurance pool of Borch's paper). But social risks—those affecting the economy at large—cannot be diversified; they must be shared among participants to the pool.[4] Borch's theorem on Pareto optimal risk exchanges implies that the sharing rule is based on participants' risk tolerances (Wilson 1968). Each individual (reinsurer) gets a share of the social risk (the reinsurance pool) in proportion to his/her absolute risk tolerance, which is the inverse of absolute risk aversion. If all individual utility functions belong to a certain class (later known as the HARA[5] class and including the most widely used utility functions), the sharing rule is linear. The CAPM, which has long been the dominant paradigm in finance theory, represents a special case of this general result.

Borch's paper is the cornerstone of insurance economics. It can conveniently be used to show how the insurance mechanism of risk pooling is part of a more global financial mechanism of risk allocation and how a distinction can nevertheless be made between insurance institutions and other financial institutions.[6] For this reason, it can also be used to clarify ideas on a hotly debated issue: the links between finance and insurance (see Sect. 6 below).

Prior to 1973, Borch's seminal contribution found its main insurance economics extensions in the papers by Arrow (1970), Kihlstrom and Pauly (1971), and Marshall (1974).[7] Arrow (1970) explicitly defined insurance contracts as conditional claims, i.e., an exchange of present money against conditional future money. Kihlstrom and Pauly (1971) introduced information costs in the risk-sharing model, arguing that economies of scale in the processing of information explain why insurance companies exist. In 1974, Marshall further extended this analysis by introducing a distinction between two modes of insurance operations: reserves and mutualization. Under the reserve mode, aggregate risk is transferred to external risk bearers (investors). With mutualization, external transfer does not apply or cannot apply: aggregate losses are shared among insureds.

2.2 Arrow (1963)

The article published in 1963 by Kenneth Arrow in *The American Economic Review*, under the title "Uncertainty and the Welfare Economics of Medical Care," represents the second point of departure for risk and insurance economics. This work may be credited with at least three contributions. First, the article provided, for the first time, what has now become the most famous result in the theory of insurance demand: if the insurance premium is loaded, using a fixed percentage loading above the actuarial value of the policy, then it is optimal for an expected utility-maximizing insured to remain partially at risk, i.e., to purchase incomplete insurance coverage. More specifically, Arrow (1963) proved that full insurance coverage above a deductible is optimal in this case. Second, Arrow (1963) also proved that when the insured and insurer are both risk-averse expected-utility

[3] Actually, Borch's theorem was already present in Borch (1960), but the 1960 article was primarily written for actuaries, whereas the 1962 *Econometrica* paper was addressed to economists.

[4] Climate risk could be an example of a social risk that cannot be totally diversified. See Cummins et al. (2002) and Dionne and Desjardins (2022) for applications of the Borch model to climate risk management.

[5] HARA = Hyperbolic Absolute Risk Aversion. As noted by Drèze (1990), the linearity of the sharing rule follows from the linearity of the absolute risk tolerance implied by hyperbolic absolute risk aversion.

[6] The question of whether or not "institutions" are needed to allocate risks in the economy was tackled later in the finance literature.

[7] The applications of Borch's theorem in the actuarial literature are reviewed by Lemaire (1990).

maximizers, Borch's theorem applies: the Pareto optimal contract involves both a deductible and coinsurance of the risk above the deductible. This result was later extended by Moffet (1979) and Raviv (1979) and generalized by Gollier and Schlesinger (1996) and by Schlesinger (1997) under the less restrictive assumption of risk aversion.[8] Third, the Arrow article was also seminal in that it introduced asymmetric information into the picture. Arrow (1963) noted that transaction costs and risk aversion on the insurer's side explained the incomplete risk transfer, but he also emphasized that moral hazard and adverse selection represented major obstacles to the smooth running of the insurance mechanism. By attracting the attention of economists to these problems, he paved the way for more focused work by Pauly (1968) and Spence and Zeckhauser (1971) on moral hazard and by Akerlof (1970) on adverse selection.

2.3 Mossin (1968)

The paper by Jan Mossin, "Aspects of Rational Insurance Purchasing," published in 1968 in *The Journal of Political Economy*, is generally considered the seminal paper on the theory of insurance demand, although some of Mossin's results were also implicit in Arrow (1963) and explicit in another paper on insurance demand published earlier the same year in the same journal (Smith 1968).[9] Mossin's paper is mainly famous for having shown that (1) partial insurance coverage is optimal for a risk-averse expected-utility maximizer when the insurance premium is such that a positive proportional loading applies to the actuarial value of the policy[10] and (2) insurance is an inferior good if the individual has decreasing absolute risk aversion (DARA). It was later pointed out (see below) that these strong results are, respectively, based on the implicit assumptions that the individual faces only one risk and that the amount of risk is fixed (unrelated to wealth or income).

2.4 Akerlof (1970)

His 1970 contribution in the *Quarterly Journal of Economics*, "The Market for Lemons: Quality Uncertainty and the Market Mechanism," presented the first market equilibrium result under asymmetric information. Akerlof (1970) showed how the presence of agents with different nonobservable qualities in a given market can decrease the average quality in the market and reduce the size of the market.

His most famous example was a used car market where buyers cannot distinguish the different qualities of used cars, and all cars are sold at the same average price. The presence of bad used cars (lemons) can drive the good used cars out of the market when the difference in quality is high because the average price under asymmetric information is too low for sellers of good used cars. This example explains the development of various institutions to counteract the reduction in average quality, such as guarantees and brand-name goods.

Akerlof's main example of insurance was the medical insurance market for people over 65 years of age. Good risks do not buy this insurance because the average price is too high. He described other

[8] More precisely, Schlesinger (1997) considers one version of Arrow's theorem: the case where the insurer is risk neutral and the insured is risk averse (risk aversion being defined by Schlesinger as preferences consistent with second-degree stochastic dominance). In this case, a straight deductible policy is optimal whenever the insurer's management costs are proportional to the indemnity payment.

[9] Optimal insurance coverage using a deductible was also analyzed by Pashigian et al. (1966) and Gould (1969).

[10] Incomplete insurance may be obtained using a deductible or coinsurance (or both).

examples from the insurance literature and concluded that "This principle of adverse selection is potentially present in all lines of insurance" (p. 493). He then discussed different market mechanisms put in place to reduce asymmetric information between insurance companies and insureds.

This article generated many contributions in the insurance literature. Rothschild and Stiglitz (1976) and Wilson (1977) presented conditions for obtaining a separating equilibrium, and Hoy (1982) and Crocker and Snow (1986) showed how risk classification, based on observable characteristics, can reduce adverse selection. Other significant contributions include Riley (1979) and Spence (1978).

2.5 Ehrlich and Becker (1972)

In the modern theory of risk management, insurance is seen as only one of the tools available to manage risk. The whole set of tools may be decomposed into subsets according to the different steps of the risk management process. Insurance belongs to the set of risk-transfer tools and represents a very powerful financial mechanism to transfer risk to the market. Another subset corresponds to risk prevention. Broadly, risk-prevention mechanisms may be classified under two headings: mechanisms intended to modify the probability of an event and mechanisms intended to mitigate the consequences of an event. Ehrlich and Becker (1972), in their article "Market Insurance, Self-Insurance, and Self-Protection," published in the *Journal of Political Economy*, were the first to propose a rigorous economic analysis of risk prevention. They coined the terms *self-protection* and *self-insurance* to designate two kinds of activities, and they studied their relationship with market insurance. For this reason, their paper may be seen as the first theoretical paper on risk management. Briefly, the paper provides three main results:

1. In the absence of market insurance, a risk-averse expected-utility maximizer will engage in self-protection and self-insurance activities, but the optimal investment in these activities depends on their cost. As usual, marginal benefit in terms of higher marginal expected utility has to be weighted against the marginal disutility brought about by additional costs, so complete elimination of the risk is not optimal in general.
2. Self-insurance and market insurance are substitutes: an increase in the degree of protection provided by the insurer induces a rational individual to reduce their investment in activities (or behaviors) aimed at reducing the consequences of the insured event. Of course, this result is also important for the theory of moral hazard (see Sect. 4), but Ehrlich and Becker (1972) did not cover asymmetric information.
3. Self-protection and market insurance may be complements or substitutes, depending on the sensitivity of the insurance premium to the effects of self-protection. Thus, the insurer can give the insured an incentive to engage in self-protection activities (which reduce the likelihood of a loss) by introducing a link between the premium rate and the observation of such activities. This result is also important for the theory of moral hazard and, more generally, for agency theory (the theory of relationships between an informed agent and an uninformed principal).

2.6 Joskow (1973)

The article published by Paul Joskow in the *Bell Journal of Economics and Management Science*, under the title "Cartels, Competition and Regulation in the Property-Liability Insurance Industry," was the first successful attempt to submit the insurance sector to an economic evaluation. The paper assesses competition by analyzing market concentration and barriers to entry. It measures returns to

scale and discusses insurance distribution systems and rate regulation. By providing empirical results on these issues, it became a reference point for subsequent empirical research on the insurance sector. Briefly, Joskow (1973) found that the insurance industry was approximately competitive, that constant returns to scale could not be excluded, and that the direct writer system was more efficient than the independent agency system.

The six seminal contributions presented above prepared the ground for numerous developments, which can be grouped under three main headings: the demand for insurance and protection, economic equilibrium under asymmetric information, and insurance market structure. They are addressed in Sects. 3, 4, and 5. It is striking to realize that many of these developments are not limited to insurance economics per se. They occurred within the wider field of general economics, with insurance providing, in some cases, an illustration of general results and, in other cases, stimulation to search for general results.[11]

3 Developments: Optimal Insurance and Protection

3.1 The Demand for Insurance

Observations of economic life show that individuals generally do not insist on getting partial coverage when they take out an insurance policy. As the insurance premiums are generally loaded (at least to cover management costs), this is, however, the behavior that would be expected from the insured, according to Mossin's (1968) results. Moreover, insurance does not seem to be empirically an inferior good, as would be implied by Mossin's results, under DARA. If it was, insurance companies would be flourishing in the world's poorer nations and would be among the declining industries in the richer ones. Recent empirical research on individual demand for insurance suggests that "higher income is (. . .) positively associated with insurance purchases of all kinds" (Cohen and Siegelman 2010). This, again, contradicts Mossin's analysis (given that absolute risk aversion is, indeed, empirically decreasing). One of the seminal papers at the root of insurance economics thus led to two paradoxes. It is interesting to observe how theory was reconciled with factual observations.[12]

The second paradox (insurance is an inferior good) did not stimulate much research effort. Some scholars tried to dig into the idea by exploring the conditions under which insurance would be not only an inferior good but also a Giffen good: see Hoy and Robson (1981) and Briys et al. (1989). But interest remained limited. There are probably two reasons for this. First, following Arrow (1970), it was quickly recognized among economists that insurance is a financial claim. Thus, it does not really seem appropriate to apply to insurance concepts derived to categorize consumption goods. Second, it is likely that most scholars noticed that the condition under which Mossin's result is obtained is not generally met in practice. Mossin assumes that the individual's wealth increases but that the risky component of wealth remains unchanged. In reality, changes in wealth generally imply changes in the portion of wealth exposed to a risk of loss, and this is sufficient to resolve the paradox (Chesney and Loubergé 1986).

The first paradox (partial coverage is optimal) has stimulated much more research effort. It was first noticed that the result is not robust to changes in the pricing assumptions: for example, full

[11] The survey of developments presented in the next three sections draws on the survey of insurance economics originally proposed by Dionne and Harrington (1992).

[12] Other strange results were observed later on; for example, an increased loss probability has an ambiguous impact on insurance purchasing if the insured has DARA preferences and the insurer adjusts the premium to take the increased loss probability into account (Jang and Hadar 1995).

insurance is optimal if the loading is a lump sum.[13] Some researchers pointed out that the result was either strengthened or did not hold if the behavioral assumptions were modified: see Razin (1976) and Briys and Loubergé (1985) or the nonexpected utility developments (Machina 1995). But the most interesting breakthrough came from broadening the scope of the analysis. This was done in the early 1980s by deriving the logical conclusion from the observation that insurance is a financial claim. It had long been recognized (Markowitz 1959) that the demand for financial assets should take place in a portfolio context, taking into consideration correlations across random asset returns. The same kind of reasoning was applied to insurance by Mayers and Smith (1983), Doherty and Schlesinger (1983a, 1983b), Turnbull (1983), and Doherty (1984). In this portfolio approach, which was soon accepted as an important improvement, the demand for insurance coverage for one risk should not be analyzed in isolation from the other risks faced by the decision-maker: insurance demand is not separable, even when the risks are independent (Eeckhoudt and Kimball 1992). When considering the insurance demand for one risk, one has to take into account the other risks, their stochastic dependence on the first risk, whether they are insurable or not and under what conditions, whether some insurance is compulsory or subsidized, whether a riskless asset is traded, and more: see, e.g., Schlesinger and Doherty (1985), von Schulenburg (1986), Kahane and Kroll (1985), Briys et al. (1988), and Gollier and Scarmure (1994).[14] Thus, assuming that correlation is a sufficient measure of dependence,[15] it may be optimal to partially insure a risk that is negatively correlated with another risk, even if the premium is actuarial. Conversely, it may be optimal to fully insure a risk despite unfair pricing if this risk is positively correlated with another uninsurable risk. In a portfolio context, incomplete markets for insurance provide a rationale for full insurance of the insurable risks. Mossin's paradox can thus be resolved by changing the perspective rather than changing the analytical model (the expected utility model).[16]

Eeckhoudt and Kimball (1992) introduced the concept of prudence into the analysis of optimal insurance purchasing under background risk and pointed out that the demand for insurance for one risk was not independent of the background risk, even when the two risks are independent. Building on this premise, several papers derived the conditions under which optimal insurance demand under background risk has desirable comparative statics properties, such as an increase in optimal insurance coverage when the insured or uninsured risks increase or whether a deductible policy remains optimal under background risk: see Meyer (1992), Dionne and Gollier (1992), Eeckhoudt et al. (1991, 1996), Gollier and Schlesinger (1995), Gollier (1995), Gollier and Pratt (1996), Tibiletti (1995), Guiso and Jappelli (1998), Meyer and Meyer (1998), Mahul (2000), and Gollier and Schlee (2006).

[13] It is obvious that the paradox can be resolved if one introduces differential risk perceptions. If the insured overestimates the probability (or the amount) of loss, full insurance may be optimal even when the premium is loaded with a fixed proportional factor.

[14] On a related theme, see also Doherty and Schlesinger (1990) for the case where the insurance contract itself is risky, due to a nonzero probability of insurer default. The paper shows that full insurance is not optimal under fair insurance pricing and that the usual comparative statics resulting from the single-risk model do not carry over to the model with default risk. Their work was extended by Cummins and Mahul (2003) to the case where the insurer and policyholder have divergent beliefs about the insurer default risk.

[15] Hong et al. (2011) argued that correlation is not an adequate measure of stochastic dependence when expected utility is used. Turning to more general notions of positive and negative dependence, and focusing on coinsurance, they showed that the individual will purchase less than full (more than full) insurance if and only if the insurable risk is positively (negatively) expectation dependent with random initial wealth.

[16] These theoretical advances closely followed similar advances in the theory of risk premiums under multiple sources of risk: Kihlstrom et al. (1981), Ross (1981), and Doherty et al. (1987). This literature on optimal insurance in presence of a background risk is closely linked to the literature on the demand for a risky asset, which was pioneered by Arrow (1963) in a single-risk setting and developed later to consider the impact of background risks: see, e.g., Tsetlin and Winkler (2005) and Li (2011).

The insurance model with background risk has been extended to the case where the uninsurable risk is nonpecuniary. This is the case, for example, if the background risk represents a state of health. The problem may be analyzed using state-dependent utility functions or by introducing a second argument in the decision-maker's utility function, besides wealth. Using the second approach, Rey (2003) demonstrated that the impact of nonfinancial risk on insurance demand depends not only on the relationship between the two risks but also on the impact of background risk on the marginal utility of wealth. For instance, if the marginal utility of wealth increases under the occurrence of a nonfinancial loss (some degree of disability, for example)[17] and if the two risks are positively correlated, then insurance demand will increase. Full insurance becomes possible even with a loaded premium. But full insurance with a loaded premium may also be obtained if the two risks are negatively correlated and if the marginal utility of wealth is lower under the occurrence of a nonfinancial loss.[18] The range of possibilities for contradicting Mossin's first proposition (in Sect. 2.3 above) expands.

3.2 Insurance, Consumption, and Saving

Research integrating joint optimal decisions on consumption, saving, and insurance represents a different research program, which was addressed by Moffet (1977) and Dionne and Eeckhoudt (1984). The latter paper showed that investing in a riskless asset is a substitute for purchasing insurance. This work was generalized by Briys (1988) using a continuous-time model. Gollier (2003) considered the impact of time diversification on optimal insurance demand in a dynamic framework under the assumption of no serial correlation in risks. He showed that liquidity-constrained individuals will insure widely. Wealthy individuals will take advantage of time diversification to accumulate buffer stock wealth and avoid costs due to loaded insurance prices.

A related avenue of research concerns the joint determination of insurable asset purchases and optimal insurance coverage: see Meyer and Ormiston (1995), Eeckhoudt et al. (1997), Meyer and Meyer (2004), and Loubergé and Watt (2008) for work along this line.[19] In the first of these papers, the individual is endowed with riskless wealth and with a risky insurable asset, but insurance can only be paid for by selling a portion of the risky asset. Hence, the model differs from that of Mossin (1968), where insurance is paid out of riskless wealth. However, because an increase in wealth impacts only riskless wealth, insurance remains an inferior asset under DARA. Eeckhoudt et al. (1997) generalized the previous work by considering an individual who allocates some nonrandom wealth to the purchase of a safe asset, a risky asset, and an insurance contract to cover the risky asset. The optimal demands for the risky asset and for insurance coverage are determined jointly and paid for out of risk-free wealth. As the insurance indemnity is not linear,[20] it may be optimal to keep some wealth in the risk-free asset. As expected, it turns out that insurance and holding a riskless asset are substitutes. However, the generality of the model did not allow the authors to derive clear comparative statics results. In particular, "increases in initial wealth can lead to increases or decreases in the insurance level and to increases or decreases in the holding of the risky asset, (. . .) even when the decision-

[17] This case corresponds to a negative cross-derivative of the two-attribute utility function ($u_{12} \leq 0$). Eeckhoudt et al. (2007) showed that this is equivalent to "correlation aversion," i.e., the aversion to losses simultaneously affecting the two attributes of utility (health and wealth, for example).

[18] Following Eeckhoudt et al. (2007), the individual is then "correlation loving." For that individual, in this case, purchasing more insurance against a loss in wealth helps to mitigate the adverse impact of a negative correlation between the two risks.

[19] The following developments on this topic borrow from the literature review in Loubergé and Watt (2008).

[20] Coinsurance must be excluded to avoid corner solutions of either no holding of the riskless asset or zero demand for insurance. Meyer and Meyer (1998) address the specific case of deductible insurance.

maker is decreasingly risk averse" (Eeckhoudt et al. 1997, p. 26). Thus, Mossin's second paradox is not confirmed: insurance is not necessarily inferior under DARA. Furthermore, Meyer and Meyer (2004) considered the case where the individual is endowed with a composite portfolio of risky and riskless assets in a fixed proportion. The risky asset may be insured without changing the proportions of the two assets held. Under these peculiar circumstances, insurance turns out to be normal whenever relative risk aversion is nondecreasing. In addition, the authors proved that insurance is ordinary (not Giffen) if relative risk aversion is less than or equal to one, a condition already derived in previous work based on the standard model (Hoy and Robson 1981). The same restrictive condition is also pivotal—as in Meyer and Ormiston (1995) and in Eeckhoudt et al. (1997)—to determine whether or not insurance demand will increase in reaction to an increase in the size of the loss. Loubergé and Watt (2008) addressed the same issues by focusing on the case where the riskless asset is dominated and all available wealth is invested in a risky and insurable asset (an investment opportunity), partly to purchase the asset, partly to finance the purchasing of insurance. In their setting, coinsurance is allowed, and partial insurance is optimal when the premium is loaded and may be optimal when insurance is fair. The fraction of the investment subject to a loss is a very important parameter in the model, along with risk aversion. If the fraction is low enough, no insurance is optimal. With a larger fraction and positive insurance, insurance is normal if relative risk aversion is nondecreasing. But even with decreasing relative risk aversion, insurance is still normal if the proportion of the investment subject to a loss is higher than the rate at which relative risk aversion decreases.[21] Insurance demand increases unambiguously if the percentage loss increases. But, when the loss probability increases and the insurer simultaneously adjusts the premium rate to take this change into account, the results are less clear-cut: it turns out that the demand for insurance increases if relative risk aversion is constant and less than or equal to 1, but it may also increase for values of constant relative risk aversion larger than 1, depending on the relationship between this value and the value of the percentage loss to which the asset is exposed.[22] Obviously, changes in risk aversion complicate the analysis when purchasing an insurable risky asset and purchasing insurance are considered simultaneously instead of separately.[23]

3.3 Self-Protection and Self-Insurance

During the 1980s, research on risk prevention (self-protection and self-insurance activities, in the sense of Ehrlich and Becker (1972)) developed more slowly than research on insurance demand, but it received increased attention thereafter. The earlier contributions came from Boyer and Dionne (1989a), who noted that self-insurance leads to stronger changes in risk than self-protection (see also Chang and Ehrlich 1985), and from Dionne and Eeckhoudt (1985), who showed that increased risk aversion leads to more self-insurance but obtained the surprising result that an increase in risk aversion does not necessarily result in higher self-protection, everything else being constant. Briys and Schlesinger (1990) later proved that this result is quite robust to a change in the model setting, e.g., introducing state-dependent utility functions or a random initial wealth. As written by Briys and Schlesinger (1990), "an increase in the level of self-protection causes both a mean-preserving spread

[21] In this case, with increasing wealth, the rate of increase of the possible loss amount is higher than the rate of decrease of relative risk aversion.

[22] Using Mossin's (1968) approach, Jang and Hadar (1995) obtained that the effect of an increase in the probability of loss is, in this case, indeterminate if the utility function displays DARA, and that the demand for insurance decreases with a CARA or IARA utility.

[23] See Schlesinger (2013) for a survey on insurance demand.

and a mean-preserving contraction in the wealth distribution, with the spread occurring at lower wealth levels and the contraction at higher wealth levels" (p. 465). For this reason, a more risk-averse individual does not necessarily invest more in self-protection. Eeckhoudt and Gollier (2005) showed that introducing prudence into the picture does not lead to a more intuitive result.[24] More prudence (in the sense of Kimball 1990) does not lead to more self-protection. In particular, if the optimal self-protection expenditure of a risk-neutral agent is such that his/her probability of loss is $p_n \geq 0.5$, a risk-averse and prudent agent spends less on self-protection than does the risk-neutral agent (see also Dachraoui et al. (2004) and Dionne and Li (2011), for comparable results). The reason is that prevention has a current monetary cost (defined as an expenditure), and a more prudent individual wants to increase savings to hedge against future contingencies.[25]

Courbage et al. (2017) replaced the analysis of prevention in the broader multiple-risk context. Steps in that direction had already been taken by Briys and Schlesinger (1990) and by Briys et al. (1991) with their analysis of risky risk management. Later, Dachraoui et al. (2004) noted that their analysis of self-protection for a mixed risk-averse agent à la Caballé and Pomansky (1996) also applies when the agent faces a background risk. An important additional step on this issue was also taken by Lee (2012) in his analysis of self-protection under background risk. He considered two kinds of self-protection: self-protection effort and monetary investment in self-protection devices. He obtained that an individual facing background risk will exert more self-protection effort than the same individual without background risk. Concerning monetary investment in self-protection, Lee was also able to show that the presence of a background risk will increase self-protection if the self-protection expenditure is paid out of income, not wealth, and if wealth and consumption are complements.[26]

Following the lead of Menegatti (2009), who introduced a two-period model to consider optimal prevention, Courbage and Rey (2012) investigated the impact of background risks on optimal self-protection expenditures in a two-period model. They showed that a prudent individual does not necessarily exert more effort in the presence of background risk. The results differ depending on whether the background risk is introduced in the first or second period and depending on whether the background risk arises in the loss or no-loss state of nature. Eeckhoudt et al. (2012) also used a two-period model to determine the impact of background risk on protection expenditure. In their model, the background risk appears in the second period and it increases prevention under prudence. Moreover, a deterioration of the background risk generates still more prevention under mixed risk aversion.[27]

The model proposed by Courbage et al. (2017) is more general than the previous works because it does not assume a "background risk."[28] There are, at least, two risks, and the risks may be positively or negatively correlated or independent. The authors "wonder how the characteristics of the multiple risks and the efficiency of alternative prevention instruments influence the decision to invest

[24] See also Courbage and Rey (2006) for an extension of this result to the case of two-argument utility functions, wealth and health.

[25] The recent models are becoming very complex without generating less ambiguous results (Peter 2021a). A probability of loss equal to (greater than, lower than) 0.5 remains a crucial threshold for comparative statics analysis because it affects the direction of the variation of the variance and other statistical moments when we change the probability of loss in a two-state model. Moreover, it is not dependent on decision-makers' risk behavior.

[26] Note that all this literature on self-insurance and self-protection has been driving away from the study of the links between insurance demand and prevention. In addition, except for the paper by Lee (2012), it has focused on the case where prevention implies a monetary cost (prevention expenditures), instead of the case where prevention implies an effort producing a direct loss in utility, because the analysis of the latter case is more straightforward even if it is often applied in principal-agent models.

[27] See also Wang and Li (2015).

[28] But, like previous work, it deals with self-protection, leaving self-insurance aside.

in a portfolio of prevention measures" (p. 900). They showed that introducing a second, possibly correlated, risk increases prevention for the first risk, independently of the sign of correlation. The prevention expenditure for the first risk also increases if the two risks become more interdependent or if the severity of the second risk increases (higher expected loss). When the two risks are endogenous, a substitution effect arises in the optimal prevention for each risk, with important public policy implications: exogenously raising the required prevention level for one risk leads to a drop in the optimal prevention for the other risk.

In addition, Hofmann and Peter (2015, 2016) found that the presence of savings also matters for the effect of utility curvature on self-insurance and self-protection. As shown by Peter (2017), in general, the presence of savings matters for the comparative statics results of two-period self-protection.

Peter (2024, 2025) presented two detailed surveys that summarize the different technical difficulties in obtaining intuitive results in existing models on self-protection research. He emphasized the role of probability thresholds in obtaining more precise comparative statics results and reviewed extensions to nonexpected utility and recent contributions in two-period models.[29]

Ehrlich and Becker (1972) had already provided the main intuition to understand the self-protection puzzle. They noticed that, unlike insurance, self-protection does not redistribute income among world states. Rather, the amount spent to decrease the probability of loss reduces final wealth in all states, leaving the size of the loss unchanged. Peter (2024) presented an example demonstrating the importance of this feature of the self-protection puzzle. The main result is that the comparative statics of self-protection does not correspond to the single-crossing property proposed by Jewitt (1989) for comparative statics analysis under uncertainty.

Jullien et al. (1999) presented a method to understand this (see also Lee 1998). Considering two individuals with different levels of risk aversion, they showed that an increase in risk aversion raises the demand for self-protection if and only if the loss probability of the less risk-averse decision-maker (DM) is small enough. The use of probability threshold results such as the one discovered by Jullien et al. (1999) became common in the economic analysis of self-protection. However, the optimal loss probability is endogenous because it depends on the risk preferences of the reference DM's utility function. The probability threshold itself is also an endogenous quantity because it depends on the DM's utility function and the risk preferences of the reference DM. The main question is how to determine the magnitude of the threshold in a given application (Dachraoui et al. 2004). How to measure this threshold empirically is not obvious!

Turning to a different question, Dionne and Eeckhoudt (1988) analyzed the effect of increasing risk on self-protection. In particular, they studied a marginal increase in the loss amount that is compensated by an increase in initial wealth such that the expected final wealth stays constant (mean-preserving spread). Counterintuitively, such a change again has no definitive effect on self-protection and may well lead to a decrease in its optimal level. Peter (2024) extended the problem raised by Dionne and Eeckhoudt (1988) and showed how the use of threshold analysis generates definitive predictions.

Sweeney and Beard (1992) studied the effects of wealth on self-protection. Unlike insurance, DARA is neither necessary nor sufficient for an increase in wealth to reduce the optimal demand for self-protection. The effect depends on the loss probability relative to a threshold value. Peter (2022) revisited the question of wealth effects and presented two main findings. First, the probability threshold can be calibrated for various classes of utility functions. These parametric results suggest that self-protection is an inferior good. Self-protection can even be Giffen, which is, however, not

[29] See also the Geneva Risk Economics Lecture presented by Bleichrodt (2022) for another review with an emphasis on nonexpected utility. He also discusses how empirical research may help to understand the prevention puzzle.

empirically plausible, according to Peter (2021b). Second, when considering that the value of assets subject to prevention is increasing in income, self-protection is more plausibly a normal good.

One of the biggest shortcomings of this literature is the lack of empirical work on self-protection. Courbage et al. (2013) made this point 10 years ago, and little progress has been made since then. For example, prevention in health should be studied further. Courbage and Rey (2006) and Peter (2021a) considered self-protection against a health risk, which is often referred to as *primary prevention* in health economics. Another field of application should be climate risk. Insurance companies are starting to leave this market, which is considered too risky. How private and social self-protection activities can be used for population resilience will become a very important issue in the coming years. However, it is not clear that the theoretical results in the actual self-protection literature would be very useful. More theoretical research is still needed, with careful attention given to applicability.

Another problem with this literature is the use of different endogenous thresholds, which depend on the nature of the utility function, to obtain intuitive comparative statics results. This is quite problematic for tests in empirical studies. It is thus very difficult for a public authority to implement preventive projects for a population with different nonobservable risk attitudes and substitution effects between prevention efforts for different risks.

3.4 The Demand for Liability Insurance

Liability risk raises particular issues that have been addressed in a specific branch of the insurance economics literature at the interface between law and economics. An economic agent (the injurer) may be made liable for the monetary and nonmonetary losses they impose on another agent (the victim). The losses are random but are generally influenced by the injurer's decisions regarding his/her level of potentially harmful activity and level of prevention. The injurer can contract liability insurance to cover the risk of being sued by the victim(s). The availability of insurance is not without influence on the injurer's decisions regarding the levels of activity and care, as has been known since Ehrlich and Becker (1972), but the specificity of liability insurance arises from the possibility that the losses imposed on the victim(s) exceed the injurer's wealth. In this case, the injurer is judgment proof and cannot be forced to bear the full monetary consequences of the losses resulting from his/her activity. This influences the injurer's optimal level of care and insurance demand.

The impact of the judgment-proof problem on the demand for liability insurance was first analyzed by Sinn (1982). He remarked that when injurers are socially guaranteed a minimum subsistence level of wealth, a kink appears in their utility function. The kink breaks the overall concavity of the function, with the result that a risk-averse injurer may rationally choose not to purchase any liability insurance even if the insurance premium is actuarially fair. Whether or not insurance will be purchased depends on the injurer's initial wealth, on the socially guaranteed minimum subsistence level, and on the size of the loss, among other usual influences such as risk aversion and insurance price. If insurance is purchased, it will be full insurance if the premium is actuarially fair (no proportional loading applies), but partial insurance if the premium entails proportional loading—as would be expected from Mossin's (1968) results. However, a noteworthy implication of the analysis is that insurance demand for liability insurance is an increasing function of the injurer's initial wealth, even if the injurer's preferences are DARA and the possible loss is fixed—a result opposite from the one obtained by Mossin in the property insurance case.

Huberman et al. (1983) emphasized that the reluctance of injurers to purchase liability insurance derives from the fact that the insurance premium takes into account a range of (high) losses to which injurers are not exposed if such losses exceed the value of their assets (or the difference between their assets and the socially guaranteed minimum subsistence level). Using the example of a three-state

model and assuming actuarially fair insurance pricing, they showed that when a possible large loss exceeds the injurer's assets, it is preferable for the injurer to set a limit on the insurance indemnity and remain partially covered, even for losses that do not imply bankruptcy. The risk-spreading function of insurance is hampered.

Liability insurance and the judgment-proof problem were comprehensively analyzed by Shavell (1986) in a model where the potential injurer decides simultaneously on whether to purchase insurance and on the optimal level of care under two possible legal frameworks: strict liability and the negligence rule.[30] Under strict liability, the results depend on whether or not the insurer can observe the insured's level of care. With perfect information, the insurer can adjust the premium to the level of care. In this case, if the injurer's initial wealth exceeds some threshold, full insurance and an efficient level of care are both optimal. As care is observable and impacts the insurance premium, the insured gains from adopting the efficient level of care, whereas in the absence of insurance, a lower level of care would have been optimally chosen due to the judgment-proof problem. If the injurer's initial wealth is below the threshold, no insurance is purchased and the level of care is zero or reduced below the efficient level.

These results make a strong case for imposing compulsory liability insurance, but they are conditioned on the assumption of perfect observability. If care exercised by the injurer is not observable by the insurer (moral hazard), either no insurance is chosen (in a wider range of injurer's wealth) or insurance is partial, with a sufficient level of care. In this situation, the case for compulsory liability insurance is not made. Tension arises between risk-spreading and appropriate incentives to avoid losses.

When the negligence rule holds, care is assumed to be observable. In this case, it is optimal for the injurer not to purchase any insurance and to exercise the optimal no-insurance level of care when the judgment-proof problem arises: no care at all if initial wealth is below a first threshold; an increasing level of care if initial wealth is beyond this threshold but below a second threshold; or an efficient level of care if initial wealth is beyond the second threshold. The reason is that it is useless to purchase insurance if the injurer applies for the appropriate level of care. Not applying this level and purchasing coverage for the risk of being liable would be more costly.[31] Of course, the result under the negligence rule hinges on the belief that the injurer will not be judged liable for the loss imposed on the victim(s) if the efficient level of care is chosen. Judicial uncertainty is ruled out (see Shavell 2000, for more information on this issue). More generally, the results presented above rely on the assumption that the legal tort liability/liability insurance system is efficient, which has been hotly debated, particularly in the wake of the US liability insurance crisis of the mid-1980s. (See Danzon and Harrington (1992), for an early survey of liability insurance issues; and Ambrose et al. (2013), Baker and Siegelman (2013), Faure (2014), and Shavell (2018), for more recent surveys.) Parra and Winter (2025) discussed the interaction of protection against risk through limited liability and protection through insurance. Incentives from tort law, in the presence of moral hazard, were also addressed. The literature on liability insurance has not been especially dynamic in recent years. This may be explained by a few changes in laws and regulations.

[30] Under the negligence rule, an injurer cannot be made liable for the losses imposed on a victim if the injurer has applied the appropriate level of care. In theoretical work, the appropriate level of care is the socially efficient level of care optimally chosen by a risk-averse potential injurer if the judgment-proof problem does not arise. This level balances the marginal benefits and marginal cost of care.

[31] In addition, the injurer would run the risk of being denied indemnification by the insurer if it turned out *ex post* that the level of care had been inappropriate.

3.5 Other Contributions

Other work in the theory of optimal insurance has focused on the following:

1. The specific issues raised by the corporate demand for insurance: these issues will be considered in Sect. 6 below.
2. The focus is on anomalies observed in actual insurance purchasing behavior (Kunreuther and Pauly 1985).
3. The extension of the expected utility model to take into account state-dependent utility functions. One can thus introduce into the analysis important observations from reality. For example, the observation that the indemnity paid by the insurer cannot provide complete compensation for a nonmonetary loss, such as the loss of a child, or the observation that the marginal utility of wealth is different under good health than under disability: see Arrow (1974), Cook and Graham (1977), and Schlesinger (1984) for important papers along this line. This line of research is related to, and generalized by, the literature on insurance purchasing under multiattribute utility functions (Rey 2003).
4. The replacement of the expected utility model with recent generalizations, grouped under the heading "nonexpected utility analysis." This research program has produced several interesting results. Using the distinction between risk aversion of order 1 (first-order risk aversion) and risk aversion of order 2 (second-order risk aversion),[32] Segal and Spivak (1990) showed that Mossin's (1968) result on the optimality of partial coverage under a loaded insurance premium does not necessarily hold if risk aversion is of order 1 (see also Schlesinger 1997). Now, risk aversion of order 1 may occur under the expected utility model (if the utility is not differentiable at the endowment point) or under some generalizations of this model, such as Yaari's (1987) dual theory or Quiggin's (1982) rank-dependent expected-utility theory, and in the presence of a dependent background risk (Dionne and Li 2014). In particular, using Yaari's model, Doherty and Eeckhoudt (1995) showed that only full insurance or no insurance (corner solutions) are optimal with proportional insurance when the premium is loaded.[33] Karni (1992) showed that Arrow's (1963) result on the optimality of a deductible policy is robust to a change in behavioral assumptions if the modified model satisfies some differentiability conditions, which are met by Yaari's (1987) and Quiggin's (1982) models. Konrad and Skaperdas (1993) applied Ehrlich and Becker's (1972) analysis of self-insurance and self-protection to the rank-dependent expected utility model. Schlee (1995) confronted the comparative statics of deductible insurance in the two classes of models. To date, the most comprehensive attempt to submit classical results in insurance economics to a robustness test by shifting from expected utility to nonexpected utility can be found in Machina (1995). Using his generalized expected utility analysis (Machina 1982), he concluded that most of the results are quite robust to dropping the expected utility hypothesis. However, the generality of his conclusion was challenged by Karni (1995) since Segal and Spivak (1990) showed that Machina's generalized expected utility theory is characterized by risk aversion of order 2.

The demand for insurance under background risk in a nonexpected utility setting was analyzed by Doherty and Eeckhoudt (1995) using Yaari's (1987) dual choice theory. They showed that partial insurance may be obtained under proportional coverage and a loaded insurance premium if an

[32] The orders of risk aversion, as defined by Segal and Spivak (1990), rest on the behavior of the risk premium in the limit, as the risk tends toward zero.

[33] This result is reminiscent of the same result obtained under Hurwicz's model of choice under risk: see Briys and Loubergé (1985).

independent background risk is present (full insurance remains optimal if the premium is fair). Dropping the independence assumption, they noted that the likelihood of getting a corner solution increases. But, qualitatively, the effects of introducing positively or negatively correlated background risks are the same as those under expected utility. More generally, Schlesinger (1997) showed that introducing an independent background risk in a decision model with risk aversion does not change the predictions obtained under a single source of risk: full insurance is optimal under a fair premium; partial or full insurance may be optimal under a loaded premium; and a deductible policy remains optimal.

4 Developments: Markets Under Asymmetric Information

The Arrow (1953) model shows that a market economy leads to a general and efficient[34] economic equilibrium—even under risk and uncertainty—if the financial market is complete, i.e., provided the traded securities and insurance contracts make it possible to optimally cover any future contingency. This is an important result since it extends to the case of risk and uncertainty the classical result on the viability and efficiency of a competitive market economy.

However, as Arrow himself noticed in his 1963 article, full insurance coverage is not always available in insurance markets for various reasons. Among these reasons, asymmetric information has received much attention in the economic literature and has been generally discussed under two main headings: moral hazard and adverse selection. Moral hazard exists when (1) the contract outcome is partly under the influence of the insured and (2) the insurer is unable to observe, without costs, to what extent the realized losses are attributable to the insured's behavior. Adverse selection occurs when (1) the prospective insureds are heterogeneous and (2) the risk class to which they belong cannot be determined a priori by the insurer (at least not without costs) so that every insured is charged the average premium rate.[35] Clearly, asymmetric information is a source of incompleteness in insurance markets, e.g., a student cannot be insured against the risk of failing an exam; a healthy elderly person may not find medical insurance coverage at an acceptable premium, etc. For this reason, a competitive market economy may not be first-best efficient, and this may justify second-best resource allocation mechanisms and even government intervention.

4.1 Moral Hazard

Economists make a distinction between two kinds of moral hazard, depending on the timing of the insured's action. If the unobservable action occurs before the realization of the insured event, one has

[34] An economic equilibrium is first-best efficient if it is Pareto optimal: it is impossible to organize a reallocation of resources that would increase the satisfaction of one individual without hurting at least one other individual. The first theorem of welfare economics states that any competitive equilibrium is Pareto optimal, and the second theorem states that a particular Pareto optimum may be reached by combining lump sum transfers among agents within a competitive economic system. In an efficient equilibrium, market prices reflect social opportunity costs.

[35] In an interesting article on the history of the term "moral hazard," Rowell and Connelly (2012) noted that the concepts of moral hazard and adverse selection have often been confused in the traditional insurance literature. They also remarked that this literature tends to attribute a pejorative meaning to "moral hazard," often associating it with fraud, in contrast to the economic literature, which focuses on incentives and maintains that "moral hazard has in fact little to do with morality" (Pauly 1968).

ex ante moral hazard, while *ex post* moral hazard exists when the insured's action is taken after the insured event.[36]

Ex ante moral hazard has been studied by Pauly (1974), Marshall (1976), Holmstrom (1979), and Shavell (1979), among others. They showed that insurance reduces the incentive to take care when the insurer is unable to monitor the insured's actions.[37] More generally, partial provision of insurance is optimal under moral hazard. More specifically, it was demonstrated that actuarial pricing is not optimal when the insured's behavior affects the probability of a loss. The equilibrium premium rate is an increasing function of the amount of coverage purchased (nonlinear pricing, as in Pauly (1974)). In addition, under moral hazard in loss reduction, the optimal contract is conceived such as to make the degree of coverage a nonincreasing function of the amount of losses, with large losses signaling careless behavior by the insured. Small losses are fully covered, but losses exceeding a limit are partially covered (Parra and Winter 2025). In a security design analysis, Holmstrom (1979) presented conditions under which a deductible or a coinsurance rate above a deductible is optimal (see also Shavell 1979). Shavell (1982, 1986) extended the study of moral hazard to the case of liability insurance. He showed that making liability insurance compulsory results in less than optimal care.[38]

The existence of long-term (multiperiod) contracts does not necessarily mitigate the effect of moral hazard. Under the infinite-period case, Rubinstein and Yaari (1983) proved that the insurer can eliminate the moral hazard problem by choosing an appropriate experience rating scheme that provides full incentives to take care. But the result does not, in general, carry over to the finite-period case (Parra and Winter 2025). In addition, the possibility for the insured to switch to another insurer makes a penalty scheme difficult to enforce in competitive insurance markets, where insurers do not share information about prospective insureds.[39]

Ex post moral hazard was first pointed out by Spence and Zeckhauser (1971) and later studied by Townsend (1979). In this case, the nature of the accident is not observable by the insurer, who has to rely on the insured's report or engage in costly verification.[40] Mookherjee and Png (1989) showed that random auditing is the appropriate response by the insurer in this situation. Their work was extended by Fagart and Picard (1999), who investigated the characteristics of optimal insurance under random auditing. Using a deterministic auditing policy, Bond and Crocker (1997) obtained that the optimal insurance contract includes generous payments of easily monitored losses and undercompensation for claims exhibiting higher verification costs.

The consequences of moral hazard for the efficiency of a market economy were studied by Helpman and Laffont (1975), Stiglitz (1983), Arnott and Stiglitz (1990), and Arnott (1992), among others. They showed that a competitive equilibrium may not exist under moral hazard and that the failure to get complete insurance coverage results at best in a subefficient equilibrium. This is due to the fact that moral hazard involves a trade-off between the goal of efficient risk bearing, which

[36] *Ex post* moral hazard is particularly important in medical insurance, where claimed expenses are dependent on decisions made by the patient and the physician once illness has occurred. It is also important in workers' compensation and in auto bodily injury insurance.

[37] Dionne (1982) pointed out that moral hazard is also present when the insured event results in nonmonetary losses, for example the loss of an irreplaceable commodity.

[38] Note that moral hazard is also present in the insurer–reinsurer relationship (see Jean-Baptiste and Santomero 2000). The success of index products in insurance securitization is partly due to the fact that they remove the moral hazard from the relationship between insurers and providers of reinsurance coverage (Doherty and Richter 2002).

[39] The situation is of course different in regulated markets where information on retrospective rating is shared by the market.

[40] At some point, this moral hazard problem becomes a fraud problem. See Picard (1996), Crocker and Morgan (1998), the special issue on fraud in *The Journal of Risk and Insurance*, September 2002 (Derrig 2002), and more recently Dionne et al. (2009). See Picard (2025) for a survey of the literature.

is met by allocating the risk to a less risk-averse insurer, and the goal of efficient incentives, which requires leaving the consequences of decisions about care to the insured (Parra and Winter 2025). However, government intervention does not necessarily improve welfare in this case. This depends on government information rather than the information at the disposal of private insurers. Arguments may be put forward in favor of a taxation and subsidization policy providing incentives to avoid and reduce losses, but public provision of insurance does not solve the moral hazard problem (Arnott and Stiglitz 1990).

Moral hazard has become a popular theme in economics, not only because its presence in insurance markets results in less than optimal functioning of any economic system but also because it is a widespread phenomenon. As Parra and Winter (2025) noted, moral hazard can be broadly defined as a conflict of interest between an individual (behaving rationally) in an organization and the collective interest of the organization. Insurance markets provide the best illustration for the effect of moral hazard, but this effect is also observed in labor relationships, in finance contracts, and quite generally in all circumstances where the final wealth of a principal is both uncertain and partially dependent upon the behavior of an agent whose actions are imperfectly observable. For example, in a corporation, the wealth of the firm's owners (stockholders) is partly dependent upon the actions of the manager; in judicial procedures, the final outcome is partly dependent upon the efforts of the lawyers; in a team, the success of the team is partly dependent on the individual effort of the members; and, in a bank loan, the default state is dependent on the borrower's behavior. All these situations have been studied in the economic and financial literature under the headings of principal-agent relationships or agency theory, with close connections to the literature on moral hazard in insurance. In both cases, the objectives are to define the optimal incentive contract to mitigate the effect of asymmetric information and to study the consequences of different arrangements on deviations from efficiency: see Ross (1973), Radner (1981), Lambert (1983), and Grossman and Hart (1983) for canonical references. See also Allen (1985), Fudenberg and Tirole (1990), and Chiappori et al. (1994) for research introducing credit markets and savings into the analysis. Similarly, the consequences for general economic equilibrium of market incompleteness brought about by moral hazard, among other causes, has become a central theme of research in economics: see, e.g., Polemarchakis (1990). Moral hazard developments in insurance economics are closely related to developments in general economic theory.

Turning to empirical work, and focusing on moral hazard in insurance contracts, evidence for moral hazard impacting insured losses has been documented in several studies, taking advantage of natural experiments provided by changes in legislation, starting with Dionne and St-Michel (1991) for workers' compensation, and continuing with Boyer and Dionne (1989b) and Cohen and Dehejia (2004) for automobile insurance and with Chiappori et al. (1998) and Klick and Stratmann (2007) for health insurance. More recently, Dionne and Liu (2021) applied a difference-in-differences methodology to separate moral hazard from adverse selection in an experiment in China where the bonus-malus was modified in a city and not modified in another similar city. They did not reject the presence of moral hazard while controlling for adverse selection. In contrast, Abbring et al. (2003) did not find any evidence of moral hazard in multiperiod data provided by the French system of bonus-malus in automobile insurance. A malus, which results from prior accident history and increases the cost of insurance for the policyholder, did not lead to a significant reduction in insured losses in this study.

Dionne et al. (2013) extended this literature by proposing a dynamic methodology to separate moral hazard from adverse selection (and asymmetric learning). Since there are two potential information problems in the data, a simple correlation between risk and insurance coverage is not sufficient to isolate the causality between risk and insurance coverage. They added a contract choice

equation to the standard claim equation in their dynamic econometric model. They did not reject moral hazard and asymmetric learning for certain groups of insured.

In addition, even if a drop in insurance claims is observed following the introduction of an experience rating, evidence from Canada suggests that much of the decline is due to an increased incentive not to report small claims to avoid premium increases (see also Robinson and Zheng 2010). In this case, policy changes introduced to address the *ex ante* moral hazard issue stimulate the development of a kind of *ex post* moral hazard. In the case of fraud—an extreme version of moral hazard—Hoyt et al. (2006) reported that antifraud laws introduced in the United States between 1988 and 1999 had mixed effects on automobile insurance fraud. Some laws had no statistically significant effects while others actually increased fraud. More recently, Dionne et al. (2009) found significant *ex post* moral hazard in the portfolio of a European insurer.

In summary, although there are strong theoretical reasons to believe that moral hazard is a major issue in insurance and that experience rating is a powerful tool for dealing with it, the empirical evidence so far is not compelling. This is probably due to the difficulty for econometricians to set up tests that isolate moral hazard from other influences on accidents, given the information limits on actual incentives and behavior among insureds. As pointed out by Parra and Winter (2025), the real incentives of repeated insurance contracts through the threat of higher future premiums still deserve more thinking.

4.2 Adverse Selection

A central development in the study of adverse selection was the paper by Rothschild and Stiglitz (1976). This paper assumed two classes in the insured population: good risks and bad risks. The two classes differ only with respect to their accident probability. The authors showed that a competitive insurance market does not reach a pooling equilibrium under adverse selection. Moreover, under the assumptions of the model, including the assumption of a pure Cournot-Nash strategy by insurers, a separating equilibrium is obtained if the proportion of good risks in the economy is not too large. The equilibrium situation involves the supply of discriminating contracts providing full insurance at a high price to the bad risks and partial coverage at a low price to the good risks.[41] Compared with the symmetric information case, the bad risks get the same expected utility, but the good risks suffer a welfare loss. The policy implication of the model is that, in some circumstances, insurance markets may fail, and monopolistic insurance (under government supervision) or compulsory insurance may be justified as a second best.[42]

Extensions of the basic Rothschild-Stiglitz model are due to Wilson (1977), Spence (1978), and Riley (1979), who dropped the assumption of a Cournot-Nash strategy by insurers. An equilibrium always exists, either as a separating equilibrium (Riley, Wilson) or as a pooling equilibrium (Wilson). Moreover, Spence (1978) showed that this equilibrium is efficient if discriminating insurance contracts are combined with cross-subsidization among risk classes, with low risk subsidizing high risk.[43] Extensions have concerned the case where individuals face a random loss distribution (Doherty and Jung 1993; Doherty and Schlesinger 1995; Landsberger and Meilijson 1996; Young and Browne

[41] Insurance contracts are defined in terms of price and quantity, instead of price for any quantity. Insureds reveal their class by their choice from the menu of contracts, such as the choice of a deductible.

[42] Stiglitz (1977) studied the monopolistic insurance case. Under asymmetric information, the monopolist insurer maximizes profit by supplying a menu of discriminating contracts. At the equilibrium situation, the high risks get some consumer surplus, but the low risks are restricted to partial insurance and do not get any surplus.

[43] See Crocker and Snow (1985) for a review of these models, and Dionne et al. (2025) for a survey of adverse selection theory.

1997), the case where they differ with respect to both accident probability and degree of risk aversion (Smart 2000; Villeneuve 2003), the case where some of them are overconfident (Sandroni and Squintani 2007), and the case where they are exposed to multiple risks or background risk (Fluet and Pannequin 1997; Crocker and Snow 2008). Allard et al. (1997) have also shown that the Rothschild-Stiglitz results are not robust to the introduction of transaction costs: for arbitrary small fixed setup costs, pooling equilibria may exist in a competitive insurance market, and high-risk individuals (rather than low-risk individuals) are rationed. In addition, it is important to note that a separating equilibrium may be invalidated if insureds have the opportunity to purchase coverage for the same risk from different insurers. For this reason, Hellwig (1987, 1988) presented models to take into account the sharing of information by insurers about policyholders. Mimra and Wambach (2019) extended these models to introduce dynamic strategic interactions in competitive markets.

Adverse selection was empirically tested by Dahlby (1983, 1992) for the Canadian automobile insurance market and by Puelz and Snow (1994), who used individual data provided by an automobile insurer in the US state of Georgia. Both studies reported strong evidence for adverse selection and provided empirical support for the separating equilibrium outcome; in addition, the former study found evidence of cross-subsidization among risk classes, whereas the latter found no such result. Cohen (2005), with data from an insurer in Israel, found some evidence of adverse selection for drivers with more than 3 years of driving experience.

However, more recent studies have reported less clear results. For instance, focusing on automobile insurance, Chiappori and Salanié (2000) showed that drivers with comprehensive insurance do not have a statistically different accident frequency than drivers with minimum coverage, controlling for all observable characteristics. Richaudeau (1999) and Saito (2006) obtained similar conclusions.[44] Dionne et al. (2001) verified that the results of Puelz and Snow (1994) could be explained by an improper econometric specification. They concluded that there is no residual asymmetric information on risk type in their data from an insurer in Canada once the information provided by an efficient risk classification policy has been taken into account in the econometric specification. This led to fundamental questions being raised about the proper tests for adverse selection (see Cohen and Siegelman (2010), for a review of empirical tests). More precisely, it seems that adverse selection plays a significant role in some insurance markets (annuities,[45] crop insurance[46]), but not in others (automobile, life insurance[47]), and that the evidence is mixed for other markets (health[48]). This is due partly to the inability of individuals to use their information advantage. This is also due to the difficulty of dissociating adverse selection from moral hazard in claim observations when dynamic information is not available to researchers. Evidence of adverse selection requires that individuals with comprehensive insurance coverage report higher average claims than individuals with partial coverage or uninsured individuals. But higher average claims for fully insured individuals may also be due to moral hazard. For this reason, a positive correlation observed between risk and insurance coverage does not necessarily signal the presence of adverse selection, although a null correlation signals that adverse selection, as well as moral hazard, does not play a role. Moreover, as soon as empirical tests are considered, a simplifying assumption of the original Rothschild and Stiglitz (1976) model is enhanced, namely, of identical individuals, except for their accident probability. In particular, they all have the same degree of risk aversion. Of course, in reality, attitudes toward risk differ. As

[44] Other recent studies include those by Kim et al. (2009), Olivella and Vera-Hernández (2013), Dardanoni et al. (2018), and Geyer et al. (2020).

[45] See Finkelstein and Poterba (2002, 2004).

[46] See Makki and Somwaru (2001).

[47] See Cawley and Philipson (1999) and Hendel and Lizzeri (2003).

[48] See Cutler and Reber (1998) and Cardon and Hendel (2001).

the degree of risk aversion is one factor influencing the demand for insurance coverage, it becomes difficult to conclude whether or not individuals who get more coverage from their insurers can be considered high risk. It may also happen that they belong to the good-risk group but demand more insurance simply because they are more risk averse (Smart 2000; Villeneuve 2003; Cohen and Einav 2007; Chiappori and Salanié 2013). The problem gets worse if, as Einav and Finkelstein (2011, p. 124) remark, "in many instances individuals who value insurance more may also take action to lower their expected cost: drive more carefully, invest in preventive health care, and so on." Such a remark opens the door to the possibility of "advantageous selection": the insureds with a high degree of coverage are those with the lowest accident probability. They demand more insurance, even at a high price, not because they are high risks but because they are, on average, more risk averse than high risks and because the heterogeneity in risk-aversion coefficients exceeds the heterogeneity in endowed riskiness.[49] Advantageous selection is not only a textbook curiosity. It has been documented in several markets, such as long-term care insurance (Finkelstein and McGarry 2006) and supplemental insurance (Fang et al. 2008). In these markets, the correlation between observed risk and insurance coverage is negative rather than positive.

To overcome these pitfalls in testing for adverse selection, Einav et al. (2010) used identifying variations in the price of health insurance provided by one specific insurer to estimate the demand for insurance. The resulting variations in quantity, together with cost data, were then used to estimate the marginal cost of additional policies. This information allowed them to test for adverse selection (the marginal cost of contracts should be decreasing) and for the associated welfare loss. Their study provides evidence of adverse selection, but the welfare impact of this inefficiency seems to be small in both absolute and relative terms.

When adverse selection is present, other insurance devices, used to deal with it, are experience rating and risk categorizations. They may be used as substitutes or complements to separating contracts. Dionne (1983) and Dionne and Lasserre (1985), on the one hand, and Cooper and Hayes (1987), on the other hand, extended Stiglitz's (1977) monopoly model to multiperiod contracts, respectively, with an infinite horizon and a finite horizon and with full commitment by the insurer to the terms of the contract. Hosios and Peters (1989) extended the finite-horizon case to limited commitment. In this case, contract renegotiation becomes relevant, as information on the risk types increases over time. In addition, the strategic use of accident underreporting becomes an issue.

Cooper and Hayes (1987) also extended the Rothschild and Stiglitz (1976) model to a two-period framework. They were able to demonstrate the beneficial effect of experience ratings under full commitment by insurers, even when the insureds had the opportunity to switch to a different insurer in the second period (semicommitment).[50] At equilibrium, the competitive insurer earns a profit on good risks in the first period, compensated for by a loss in the second period on those good risks who do not report an accident. This temporal profit pattern was termed *highballing* by D'Arcy and Doherty (1990). A different model, without any commitment and assuming myopic behavior by insureds, was proposed by Kunreuther and Pauly (1985). The nonenforceability of contracts implies that sequences of one-period contracts are written. Insurers' private information about their customers' accident experience allows for negative expected profits in the first period and positive expected profits on

[49] Advantageous selection can lead to too much insurance being purchased if there are transaction costs and competition among insurers drives profits to zero. In equilibrium, the marginal cost of insurance exceeds the market price (see Einav and Finkelstein 2011). The possibility of advantageous selection was first introduced by Hemenway (1990), who termed it "propitious" selection, and it was later analyzed by De Meza and Webb (2001).

[50] Note, however, that there exist alternative views on the welfare effect of asymmetric information. Using a two-period model where insureds have the option to switch insurers in the second period, de Garidel-Thoron (2005) showed that information sharing among insurers is welfare-decreasing. This is because it reduces the set of viable long-term contracts available to individuals in the first-period competition game.

the policies they renew in subsequent periods (*lowballing*).[51] Later, Dionne and Doherty (1994) proposed a model assuming private information among insurers about the loss experience of their customers and semicommitment with renegotiation. Then, the insured has the option to renew the contract on prespecified conditions (future premiums are conditional on prior loss experience). This assumption seems to come closer to actual practices used in insurance markets. The authors derived an equilibrium with first-period semipooling[52] and second-period separation. Their model predicts highballing since a positive rent must be paid in the second period to the high-risk individuals who experienced no loss in the first period, and this is compensated for by a positive expected profit on the pooling contract in the first period.[53] Their empirical test, based on data from Californian automobile insurers, provides some support for this prediction. They conclude that some (but not all) insurers use semicommitment strategies to attract a portfolio of predominantly low-risk drivers.

More recently, Crocker and Snow (2011) drew attention to the fact that multidimensional screening is routinely used by insurers to cope with adverse selection. With n mutually exclusive perils, insurers "can now exploit n signaling dimensions to screen insurance applicants" (p. 293). The good risks tend to accept higher deductibles for perils they are less likely to be exposed to, for instance, theft. This allows insurance markets to circumvent the nonexistence problem identified by Rothschild and Stiglitz (1976). As the authors themselves remarked, using multidimensional screening at a point in time presents an analogy with (and may be a substitute for) using repeated insurance contracts in a dynamic framework.[54]

Risk categorization, which uses statistical information on correlations between risk classes and observable variables (such as age, sex, and domicile), was first studied by Hoy (1982), Crocker and Snow (1986), and Rea (1992).[55] Their work shows that risk categorization enhances efficiency when classification is costless but that its effect is ambiguous when statistical information is costly (Bond and Crocker 1991). This last result was recently challenged by Rothschild (2011), who showed that a ban on risk categorization is always suboptimal—even when categorization is costly. He introduced a distinction between a regime where categorization is used by insurers (as in Crocker and Snow 1986) and a regime where categorization is allowed but may or may not be used in equilibrium. He then showed that a ban on risk categorization is always (regardless of the insurance market regime) Pareto dominated by having the government introduce partial social insurance and simultaneously lifting the ban on risk categorization for private supplemental coverage.[56]

Here is a quote from Rothschild (2011, p. 269):

> The intuition behind the effectiveness of social insurance for preventing the negative consequences of lifting categorical pricing bans is simple. Categorical pricing bans are potentially desirable insofar as they implicitly transfer resources from individuals in low-risk categories to individuals in high-risk categories. (. . .) Providing partial social insurance effectively socializes the provision of this cross-subsidy. Lifting a categorical pricing ban then allows the market to employ categorical information to improve efficiency without risking undoing the cross-subsidy.

[51] In Kunreuther and Pauly (1985), the insurers have no information about other contracts their customers might write. For this reason, price-quantity contracts are unavailable. The equilibrium is a pooling equilibrium with partial insurance for the good risks, as in Pauly (1974).

[52] In the first period, insureds may choose either a pooling contract with partial coverage and possible renegotiation in the second year, or the Rothschild-Stiglitz contract designed for high risks.

[53] For good risks who do not file a claim in the first period, the reward takes the form of additional coverage in the second period.

[54] See also Bonato and Zweifel (2002) on the use of multiple risks to improve the assessment of loss probability.

[55] See Crocker et al. (2025) for a recent survey.

[56] See Dionne and Rothschild (2014) for a framework to evaluate the economic consequences of legalized and banned risk classification systems, both in a static environment and a dynamic environment with learning.

These results are of the utmost political importance, given the ethical critiques on the use of observable personal attributes, such as sex and race, in insurance ratings. The problem of risk categorization is even more acute when personal attributes are not observable a priori but may be revealed to the insurer and/or the insured after some information-providing steps have been taken, as in the case of genetic diseases. Rothschild and Stiglitz (1997) pointed out that this results in a conflict between the social value of insurance and competition among insurers. If valuable information can be made available about the probability (or certainty) that an insured will suffer from a particular genetic disease, then insurers will want that information. But this will result in less insurance coverage since the insureds who are virtually certain to get the disease will not be able to obtain insurance, whereas those who are revealed to be immune to the disease will no longer need insurance.[57] For ethical reasons, society prohibits the use of genetic information by insurers to categorize risks. But this means that adverse selection problems are enhanced, at least in medical insurance. As Doherty and Posey (1998) showed, private testing is encouraged when test results are confidential and there is a treatment option available,[58] but insurers are unable to charge different prices to different customers with private information about their genetic background. Combining partial social insurance with supplemental private insurance, as suggested by Rothschild (2011), could be a way out of this conflict between the efficiency of insurance pricing and the mutuality principle.

Like moral hazard, adverse selection is an important problem that extends beyond the field of insurance. It is mainly encountered in labor markets, where employers are uninformed about the productivity of prospective employees, and in financial markets, where banks and finance companies lack information on the reimbursement prospects of borrowers. The insurance economics literature on adverse selection reviewed above has thus led to applications in other areas of economics: see, e.g., Miyazaki (1977) for an application to the labor market and Stiglitz and Weiss (1981) for an application to credit markets. Note, however, that in these cases, quality signaling by the informed agents represents a feasible strategy for circumventing the asymmetric information problem (Spence 1973). For example, education and dividend payments find additional justification in these circumstances. In contrast, signaling does not generally occur in insurance markets, as insureds do not engage in specific activities to signal that they are good risks.

4.3 Moral Hazard and Adverse Selection

Progress in analyzing moral hazard and adverse selection together has remained very limited. This was already noted by Arnott (1992) in the early 1990s, and the situation has not significantly changed since then. This has long limited the significance of empirical investigations in the economics of insurance since both problems come together in actual insurance markets. A positive correlation between insurance indemnities and insurance coverage may be interpreted as signaling the presence of adverse selection, moral hazard, or both. The first attempts to address these two problems jointly were made by Dionne and Lasserre (1987) in the monopoly case and by Eisen (1990) in the competitive case, but neither was echoed in the literature. In the same period, Bond and Crocker (1991) pointed out that risk categorization may be endogenous if it is based on information on consumption goods that are statistically correlated with an individual's risk (correlative products). Thus, adverse selection and moral hazard become related in their framework. If individual consumption is not observable, then the taxation of correlative products by the government may be used to limit moral hazard, which would

[57] This is an example of the well-known result that additional public information may have adverse welfare consequences (see, e.g., Arrow 1978).

[58] In contrast, Doherty and Thistle (1996) found that additional private information has no value if there is no treatment option conditional on this information.

reduce the need for self-selection mechanisms as an instrument for dealing with adverse selection. However, they did not provide a general model.

Advances on this research front currently seem more promising at the empirical level, using material provided by longitudinal data on insurance purchasing and loss experience. Adverse selection is due to differences in the dynamics of learning about the insured's true risk type for the insured themselves and for the insurer. Once adverse selection has been identified (or not), then, with the use of longitudinal data, the residual effect of moral hazard may be tested. This approach was followed by Dionne et al. (2013) using data on automobile insurance and car accidents in France. They first calibrated a simulation model for the optimal behavior of car owners, using the specific features of auto insurance in France, and showed that adverse selection and moral hazard should be expected. They then tested the model with real data for the presence of asymmetric learning on the one hand and moral hazard on the other hand. Asymmetric learning is about accidents, not claims, and insurers observe only claims while insureds observe both. In the long run, asymmetric learning can be shown to become adverse selection. The statistical results differ according to the experience of the insureds. For insureds with less than 15 years of experience, they found strong evidence of moral hazard but little evidence of asymmetric learning. The latter occurs only for insureds with less than 5 years of experience. In contrast, for insureds with more than 15 years of experience, there is no evidence of moral hazard or adverse selection. These results should stimulate further research along the same line in other insurance contexts or markets.

5 Developments: Insurance Market Structure

Numerous studies on the insurance sector have followed the lead provided by Joskow (1973). The fact that most of these studies pertain to the US market can be explained by the availability of data and the size of the insurance market.

5.1 Insurance Distribution

Insurance distribution systems have been analyzed by several researchers, more particularly Cummins and VanDerhei (1979) and Berger et al. (1997).[59] In line with Joskow (1973), direct writing has generally been found to be more cost-efficient than independent agents. However, the differences in efficiency are not significant, which may be interpreted as an indication that independent agents provide valuable services. This interpretation has received support in several articles, e.g., Barrese et al. (1995) for the United States and Eckart and Räthke-Döppner (2010) for Germany. A comprehensive study on this topic is that of Cummins and Doherty (2006). The authors showed that insurance intermediaries (brokers and independent agents) play a valuable role in improving the efficiency of the market. Over the past 30 years, insurance markets in several European and Asian countries have witnessed the marketing of insurance contracts through banking channels—bancassurance. The efficiency of this distribution system has been investigated, with mixed results so far; for instance, Chang et al. (2011) did not report efficiency gains for this system in Taiwan.

Hilliard and Tennyson (2025) presented a detailed overview of insurance distribution systems in the US and international markets. They observed an increase in the standardization of simpler insurance products in the life and nonlife markets over recent years. For these products, there is an emphasis on low-cost distribution and nontraditional methods of reaching clients, including

[59] See also Zweifel and Ghermi (1990) for a study using Swiss data.

the Internet. Customers pay their own premiums, and insurers use individual underwriting models, thereby significantly reducing distribution costs.

The industry is moving away from traditional relationships between insurer direct writing and independent agents. Professional agents are now more focused on selling more complex products such as commercial insurance. We observe fully integrated distribution systems without professional agents versus agency systems of distribution instead of the traditional separation between independent agents and a tied agency team. According to the authors, this can be explained by technological changes and increased competition in insurance markets. The regulation of insurance distribution should now focus on the market conduct of professionals in service-oriented sectors and on adequate disclosure in standardized insurance products.

5.2 Organizational Structure

The various forms of organizational structure in the insurance industry—stock companies, mutuals, and Lloyds' underwriters—were analyzed in an agency theory framework by Mayers and Smith in a series of papers (1981, 1986, 1988), among others. They verified that conflicts of interest between owners, managers, and policyholders affect the choice of organizational form for different insurance branches (see also Hansmann 1985; Cummins et al. 1999b). Mutuals tend to pevail when the relationship between owners and policyholders triggers substantial agency costs. However, mutuals are constrained by their lack of access to external capital, with the result that mutuals with strong growth choose to convert to the stock structure when the constraint on their expansion becomes too costly: see also Mayers and Smith (2002), Harrington and Niehaus (2002), and MacMinn and Ren (2011) for other references. Ho et al. (2013) showed that mutual insurers have lower total risk, underwriting risk, and investment risk than do stock insurers. This research subject has not significantly evolved per se over the last 10 years. However, organizational structure is often used as a control variable in empirical insurance analyses, such as corporate risk management, mergers and acquisitions, insurance pricing, and corporate governance.

5.3 Mergers and Acquisitions (M&A)

Chamberlain and Tennyson (1998) showed that targets with greater capital needs are more likely to obtain capital infusion at acquisition, particularly after 1985, when targets were poorly capitalized. Cummins et al. (1999a) showed that target life insurers experienced larger efficiency gains than did firms in a control group that were not involved in M&A deals. Cummins and Xie (2008) did a similar analysis in the property and liability sector for the 1994–2003 period. They obtained that targets achieve higher cost and allocative efficiency, as compared to acquirers. Being acquired is value-enhancing in their sample of firms. They also verified that poorly performing targets with low capitalization and poor underwriting performance are more likely to be targets, while large and rapidly profitable firms are more likely to be acquirers. Groups are more likely to be acquirers compared to unaffiliated single firms and mutuals. Shim (2011) showed that the performance of acquiring firms decreases during the gestation period after M&As and that more focused insurers outperform product-diversified insurers.

Boubakri et al. (2008) investigated whether M&A transactions created value for acquirers' shareholders during the 1995–2000 period, where acquirers are US property-liability insurers and where targets could be US or foreign insurers. They added a corporate governance dimension to their analysis by testing if the long-run performance of the acquirers is related to the percentage of

shares held by institutional investors, blockholders, and the CEO in the absence of entrenchment problems. Their results confirm positive abnormal returns in the long run for acquirers. The returns are higher when buying other property-liability insurers, meaning that focused M&A transactions are more value-enhancing. The results related to corporate governance indicate that the percentage of shares held by the CEO and CEO duality (CEO and board chair) are negatively related to the bidders' long-run performance, which is consistent with entrenchment theory.

More recently, Cummins et al. (2015) tested the hypothesis that M&As can be value-creating for both acquirers and targets. They also tested whether cross-border transactions are value-creating for targets and acquirers. They verified that the transactions are value-enhancing for both acquirers and targets, with a greater effect for targets. Moreover, for acquirers, cross-border deals provide greater value, while there is no statistical difference for targets.

Klumpes (2022), with European data, showed that technical efficiency gains were significantly higher for acquiring insurers than for nonacquiring insurers in the prefinancial crisis period. The results also indicate that larger insurers, with a higher premium/surplus ratio, realized the highest efficiency gains. Well-capitalized insurers, with high ratios of equity/invested capital and capital/total assets, realized a loss of efficiency for the subperiod of 1997–2007. Klumpes (2022) also estimated the probability that an insurer becomes an acquiring firm, using a dummy dependent variable, taking the value of zero for nonacquiring insurers and 1 for acquiring insurers. Acquiring firms are included in the logistic regression analysis only in the year of their acquisition deal, and non-M&A firms are included for all sample years for the precrisis and postcrisis subperiods. The logistic regression results show that more efficient insurers are more likely to be acquirers, which tends to support the previous results related to the technical efficiency improvements of M&A transactions. The results also indicate that the log of assets, life insurance dummy, percentage of equity investments, and UK dummy are significantly positively related to the probability of being an acquirer in both the precrisis and postcrisis periods.

5.4 Regulation

Following the lead provided by Joskow (1973), the effects of rate and solvency regulation have been scrutinized in a variety of studies, such as those of Borch (1974), Ippolito (1979), Munch and Smallwood (1980), Danzon (1983), Finsinger and Pauly (1984), Pauly et al. (1986), Harrington (1984, 1987), Cummins and Harrington (1987), D'Arcy (1988), Harrington and Danzon (1994), and Cummins et al. (2001). These studies were stimulated by the traditional government regulation of insurance activities, a general trend toward deregulation in the 1980s and 1990s,[60] and consumer pressures for reregulation (mainly in California and Florida) after major disasters such as Hurricane Andrew in 1992 and the Northridge earthquake in 1994. Dionne and Harrington (1992) concluded their survey of research on insurance regulation by noting the following. First, "not much is presently known about the magnitude of the effects of regulatory monitoring and guaranty funds on default risk" (p. 32). Second, rate regulation seems to have produced a variety of effects. It favored high-risk groups, increased market size, and encouraged insurers' exits but nonetheless reduced the ratio of premiums to losses and operating expenses. More recently, Klein et al. (2002) found that price regulation tends to increase leverage while Doherty and Phillips (2002) remarked that, during the 1990s, when there was a trend toward deregulation, the role of rating agencies was enhanced: stringency in the rating procedures provided an incentive to decrease leverage and seemed to

[60] Berry-Stölzle and Born (2012) provide an empirical account of the deregulation introduced in Germany in 1994. They find evidence of a significant price decrease in highly competitive lines, offset by higher prices in the other lines.

substitute for tight regulations. Rees et al. (1999), considering the focus of the European Commission on solvency regulation instead of rate regulation, suggested that "the role of regulation in insurance markets should be confined to providing customers with information about the default risk of insurers" (p. 55).

Debates on insurance regulation were exacerbated by the financial crisis in 2007–2008 and doubts about insurers' solvency following the collapse of Lehman Brothers and the rescue of AIG by the federal government. Eling and Schmeiser (2010) derived ten consequences of the crisis for insurance supervision, while Lehmann and Hofmann (2010) stressed the differences between insurance and banking. Harrington (2009) reviewed the AIG case and questioned the exposure of the insurance sector to systemic risk. He noted that this sector remained largely on the periphery of the crisis, in contrast to AIG (see also Kessler 2014).[61] He also noted that the crisis revealed the imperfect nature of the federal regulation of banks and related institutions. These considerations led him to reject the plans for creating a federal systemic risk regulator for insurers and other nonbank institutions designated as systemically significant. He also rejected the claim that the AIG crisis strengthens arguments for federal insurance regulation—be it optional or mandatory. In his view, "an overriding goal of any regulatory changes in response to the AIG anomaly should be to avoid further extension of explicit or implicit 'too big too fail' policies beyond banking" (p. 815). Notwithstanding these strong arguments against federal involvement in US insurance regulation, the debate continued and the US federal government revisited the regulatory system in the McCarran-Ferguson Act. The 2010 Dodd-Frank Wall Street Reform and Consumer Protection Act (Dodd-Frank) gave increased systemic risk regulatory authority to the Federal Reserve. In addition, Dodd-Frank created a Federal Insurance Office within the Department of the Treasury to establish greater uniformity among states with regard to insurance and reinsurance information reported to the federal government and to participate in international treaties (Grace 2025).

In a comprehensive review of insurance regulation procedures, Klein (2012)—see also Klein and Wang (2009)—spelled out the principles for insurance regulation and compared the traditional system of detailed state-by-state regulation in the United States with the principle-based approach introduced in EU member countries under Solvency II.[62] He concluded that, compared to the European developments, "the systems for solvency and market conduct regulation in the United States warrant significant improvement" and that "the (US) states should move forward with full deregulation of insurance prices" (p. 175).

5.5 Insurance Cycles

A related avenue of research, not considered by Joskow (1973), deals with cycles in the insurance industry. In the 1970s, insurance company profits seemed to be subject to more or less regular cycles, and this phenomenon was reflected in cyclical capacity and premium rates. The Geneva Association sponsored one of the first investigations in this area (Mormino 1979). The most frequently quoted papers were published a few years later by Venezian (1985), Cummins and Outreville (1987), and Doherty and Kang (1988). As pointed out by Weiss (2007, p. 31), "in tracking underwriting cycles,

[61] AIG's failure is mainly attributed to two causes. First, a subsidiary of AIG—AIG Financial Products—became heavily involved in the writing of credit default swaps. Second, another subsidiary, operating in the life branch, had engaged in securities-lending programs that were severely hurt by the onset of the subprime crisis. In either case, insurance operations were not concerned.

[62] The Solvency II regulation is presented and analyzed in Eling et al. (2007), Gatzert and Wesker (2012), and Doff (2025).

most of the attention tends to be directed at insurance pricing, or, conversely, insurance underwriting profits, rather than the amount of coverage available." The US insurance liability crisis of the mid-1980s and the overcapitalization of property-liability insurers during the 1990s spurred research on insurance cycles (Harrington 1988; Cummins and Doherty 2002). Haley (1993) and Grace and Hotchkiss (1995) documented the impact of external factors—mainly interest rates—on underwriting profits, but the latter authors found that unanticipated external real economic shocks have little effect on underwriting performance. Other research has suggested that delays in adjusting premiums to expected claims costs—due to regulation or structural causes, external shocks to supply capacity, and variations in insurer insolvency risk—are responsible for cyclical effects: see Winter (1994), Gron (1994), Cagle and Harrington (1995), Doherty and Garven (1995), and Cummins and Danzon (1997). Choi et al. (2002) compared six alternative insurance pricing models as theories of the underwriting cycle. They showed that two models are consistent with short-run and long-run data on underwriting profits: the capacity constraint model and the actuarial pricing model. Cummins and Nini (2002) provided empirical evidence that the overcapitalization of property-liability insurers during the 1990s was mainly due to capital gains during the stock market boom and to retained earnings. They interpreted their results as providing evidence that insurers tend to hoard capital in favorable periods as a hedge against future adverse shocks. This behavior negatively affects underwriting profits and is a source of profitability cycles.

Following the lead provided by Cummins and Outreville (1987), research on insurance cycles has also been conducted at the international level and has provided evidence that these cycles are not specific to the US market: see Lamm-Tennant and Weiss (1997), Chen et al. (1999), Leng and Meier (2006), and Meier and Outreville (2006).[63]

Researchers became more skeptical about the existence of pricing and profitability cycles in the property-liability insurance industry (Boyer et al. 2012; Henriet et al. 2016). One critical factor was the methodology used in previous studies. Grace (2023) proposed a new methodology to detect insurance profit cycles and used quarterly firm data, instead of yearly industry aggregate data, to isolate firm-level heterogeneity. He found evidence that firms can be in high- and low-loss ratio regimes by using a Markov switching model, but he did not observe cyclical changes in pricing and profitability over time, confirming the results of the above two contributions.

5.6 Insurance of Catastrophes

The insurance of catastrophes has become a major concern for the industry and is the subject of intensive academic research. Beyond the insurance industry, catastrophes have become an issue for the economy at large. Events like Hurricane Katrina on the Gulf Coast in 2005 or the Fukushima catastrophe in Japan in 2011, with its sequential earthquake, tsunami, and major nuclear accident, had repercussions for the international economy, not only locally. Given that major catastrophic events occur every year somewhere in our globalized world, it has become inappropriate to continue to define catastrophes as "low-probability/high-consequence events" (see Kunreuther and Michel-Kerjan 2009, p. 351). The major journals in the economics of insurance and some general economic journals have devoted special issues to this topic over the past two decades. Books and edited collections have also addressed this issue.[64] Researchers have tended to take a broad view of the subject, so the term catastrophe has been used to encompass different kinds of situations: not only natural disasters (like

[63] See Harrington et al. (2013) for a survey of the literature.

[64] See, in particular, Froot (1999a, 1999b), OECD (2005), Wharton Risk Management Center (2007), Kunreuther and Michel-Kerjan (2009), and Courbage and Stahel (2012).

earthquakes, tsunamis, floods, and hurricanes) and man-made catastrophes (such as Chernobyl or Bhopal) but also socioeconomic developments that result in a catastrophic accumulation of claims to insurers (Zeckhauser 1995). The prominent example of this last category is the liability crisis in the US due to the adoption of strict producers' liability and the evolution in court assessments of compensation to victims, as in the cases of asbestos, breast implants, pharmaceuticals, etc. (Viscusi 1995). To cope with the financial consequences of catastrophes, traditional insurance and reinsurance are often considered insufficient (Kunreuther 1996; Froot 1999a, 1999b; Cummins et al. 2002). Some researchers have invoked the difficulties individuals face in dealing with low-probability/high-loss events (Kunreuther and Pauly 2004). Others have invoked capital market imperfections and market failure in reinsurance supply (Froot 2001; Zanjani 2002; Froot and O'Connell 2008). Still others have pointed to US insurance price regulation in catastrophe-prone lines of business as a major source of inefficiency in the insurance and reinsurance markets (Cummins 2007). Several researchers have advocated for more government involvement (Lewis and Murdock 1996; Kunreuther and Pauly 2006),[65] but others have argued that the government has no comparative advantage over the market in providing coverage for catastrophic losses (Priest 1996) and have instead called for less government intervention by deregulating insurance prices (Cummins 2007). Alternative solutions may be found in financial innovation, either in the design of insurance contracts, by introducing a decomposition of insurance risk into a systemic and diversifiable component (Doherty and Dionne 1993; Schlesinger 1999; Doherty and Schlesinger 2002), or in the design of new financial securities, or both.

5.7 Climate Finance

Climate finance is defined by the United Nations Framework Convention on Climate Change (UNFCCC) as "local, national, or transnational financing—drawn from public, private, and alternative sources of financing—that seeks to support mitigation and adaptation actions that will address climate change" (Reported in Harrison et al. 2020). This concerns the financing that must be undertaken to change the world economy and build resilience to climate change.

Many financial sectors (banking, insurance, real estate) are directly impacted by the risks generated by carbon prices, tornadoes, wildfires, pandemics, and flooding. Difficult questions are raised: How should financial markets price and mitigate risks from global warming? How can capital markets raise so much financing? How should the distribution of damages from catastrophic events be managed? These issues were recently discussed in a special issue of the *Review of Financial Studies*, edited by Harrison et al. (2020). But there are no contributions available in finance and insurance journals on the causal effects of climate change on the insurance industry.

A crucial input for the analysis of climate change risk is the causal impact of climate events on economic activity, which is called the distribution of damages. This raises an important question about the modeling and sharing of extreme weather risks. Do extreme weather risks, such as the impact of Hurricane Sandy in 2012 or the 2018 California wildfires, have long-run causal effects on insurance markets? These distributions of damages depend on locational decisions by households and firms and on technology (self-protection and self-insurance) decisions in terms of preventing and mitigating the damage caused by disasters. They also depend on market insurance coverage (including moral hazard and adverse selection effects). In suitably modeling these loss distributions, the insurance industry should be able to play a critical role in facilitating risk sharing and extending insurance coverage for extreme weather events. In addition, public authorities must improve their prevention role.

[65] Monti (2011) provides a recent review of public–private arrangements that already exist in the OECD area.

The potential impacts of new climate patterns on damages from catastrophic risks must be better estimated by the insurance industry and public authorities. These potential impacts may have been underestimated in risk management for many years. Here are some worldwide statistics according to a Munich Re report (2019):

- Of all natural disasters worldwide, 83% were weather related during the 1980–2019 period, and 43% of the overall economic losses during that period occurred in Asia.
- Of the insured losses, 64% were incurred in North America (incl. Central America and the Caribbean) during this period, which represents about 30% of the overall losses in this region and around the world.
- According to the World Meteorological Organization, 11,000 climatic events were observed from 1970 to 2019, for a total cost of US$1400B.

In 2008, worldwide economic losses from climate disasters were evaluated at $130B, of which $44B was insured. Climate risk was rated number one of the top 10 risks facing the insurance sector (Ernst and Young 2008). The average total loss over the last 10 years was $187B ($340B in 2017 only). The year 2019 was below the last 10-year average, with a total loss of $150B but with 33 events with more than 1B in total losses each. Nine events cost the insurance industry more than $1B that year, and all of them were climate risk events (cyclones, storms with flooding, and tornadoes). And this is not the end. In April 2020, severe weather events in the US cost insurers billions of dollars, which had 14 tornadoes—the fifth worst month on record since 1950, according to an AON report (2023).

The increasing frequency and severity of extreme weather events highlight a dangerous link between insurance risk and climate change, even though less than 40% of total losses are covered. According to a recent survey by Deloitte (2020), most US state insurance regulators expect all types of insurance companies' climate change risk to increase over the medium-to-long term. More than half of the state regulators surveyed also indicated that climate change is likely to have a high impact on coverage availability and underwriting assumptions. US state regulators and lawmakers are concerned about the insurance industry's response to climate change. Two traditional mechanisms are usually used to reduce financial fragility. Insurers can increase premiums in the states or counties most affected, or they can increase reinsurance coverage (Grenier 2019). But these two alternatives may not be satisfactory for the long-run stability of the industry, even though they have been used by the US property and casualty insurance industry over the last 30 years (Dionne and Desjardins 2022).

Insurance-linked securities (ILSs) can lower the cost of risk transfer under hard (re)insurance market conditions. They help maintain the (re)insurance capacity and offer multiyear protection. They are free of credit risk by offering full collateralization of losses. ILS are noncorrelated with other market, liquidity, and credit risks, so they represent an important diversification asset. Alternative capacity (ILS) includes collateral reinsurance, sidecars, industry loss warranty (ILW), and CAT bonds. ILS penetration can reduce the price of insurance in the long run and increase the demand for insurance. But there is a long-run commitment issue as regards the financial participation of financial markets.

According to data in 2019, the total international reinsurance capital was about $625B, including $62B in ILS capacity. It should be mentioned that the total capital of the US insurance industry was about $1.1 trillion in 2020 (Dionne and Desjardins 2022). In addition to reinsurance, ILS accounted for only about 10% of global catastrophe reinsurance capital in 2019. We might think that there is sufficient capacity because the annual average long-run losses are below $150B, but there have been significant recent exceptions: 2011 ($160B) and 2017 ($170B), with an annual growth of 5–7%.

Mills (2009) analyzed different mechanisms to improve the insurance industry's capacity to cover insurable losses: new coverage products, a better understanding of climate change, and the financing of activities designed to reduce climate risk. Gollier (2005) insisted on the government playing a necessary role in reducing the fragility of the insurance industry when there are extreme events. The government should act as a reinsurer to reduce bankruptcies. This complements Mills (2009), who favored stronger private risk management (see also Michel-Kerjan 2012; Kunreuther 2018).

Jametti and von Ungern-Sternberg (2010) did not consider the observed risk selection between the private and public sectors to be optimal where the private sector keeps the acceptable or lower losses and where the public sector is limited to extreme losses. Born and Viscusi (2006) proposed a different approach to analyze the effects of natural disasters on the insurance industry. Using data from the Swiss Re Sigma Reports for the 1984–2004 period, they showed that small insurers are more likely to be affected because they are less diversified. Finally, Born and Klimaszewski-Blettner (2013) verified that some insurers tend to reduce their activities when they are subject to severe regulations or unanticipated large claims. This behavior is less frequent for large insurers that are better diversified but has become more common in recent years.

In their theoretical analysis, Cummins et al. (2002) studied the capacity of the US insurance industry to cover very large catastrophes. They proposed a sufficient condition for capacity maximization: all insurers must hold a net of reinsurance underwriting portfolios that are perfectly correlated with aggregate industry losses (Borch 1962, model). Estimating capacity using insurers' financial statement data, they found that the industry could adequately fund a $100B event, with data covering the 1983–1997 period, where the equity capital of US insurers was approximately equal to $350B in 2002. Cummins et al. (2002) also showed that the industry would be able to pay very high percentages of the losses. Using their parameter estimates, the US insurance industry would have been able to pay about 81% of a $200B loss in 1997. Dionne and Desjardins et al. (2022) showed that the capacity had increased to 98% in 2020.

5.8 Microinsurance

At the other end of the insurability spectrum, microinsurance has emerged as a new topic for research in insurance economics. The fact that a large fraction of the world's population has no access to the benefits of insurance coverage has motivated practical initiatives to remedy this situation, as well as interest among researchers and international organizations, for instance the International Labour Office (ILO). A recent study (Biener and Eling 2012) reviewed the current situation of microinsurance. The authors pointed out that the microinsurance industry has experienced strong growth in recent years (10% on average) but that much remains to be done and that further developments are constrained by well-known insurability problems: risk assessment, asymmetric information, and a lack of financial resources in the uninsured population. They also provided tentative solutions to overcome these problems. In a more recent analysis of microinsurance markets, Dror and Eling (2025) concluded that the offer of low-cost insurance to these populations is still not effective.

5.9 Corporate Governance

Corporate governance issues in the insurance industry have attracted more attention from researchers in the wake of the 2008 financial crisis and the collapse of AIG. The impacts of corporate governance and institutional ownership on efficiency (Huang et al. 2011), risk-taking (Cheng et al. 2011; Ho

et al. 2013), M&As (Boubakri et al. 2008), and CEO turnover (He and Sommer 2011) have been investigated, among other issues.[66]

New risk regulations were introduced after the 2008 financial crisis. Boards of large insurers had to establish a risk committee consisting of only independent members and with at least one member having a thorough understanding of risk management. The role of the committee is to monitor the company's risk appetite and oversee the risks the insurer is exposed to and the models used to measure, monitor, and mitigate them (Magee et al. 2019; Ames et al. 2018).

6 New Approaches: Finance and Insurance

Apart from the tremendous developments summarized in the three preceding sections, risk and insurance economics experienced a major refocus in the 1970s and 1980s, in which insurance was increasingly analyzed within the general framework of financial theory. This change in perspective was implicit in the definition of Arrow (1970): "insurance is an exchange of money for money." It was also foreshadowed by the recognition that insurers are financial intermediaries (Gurley and Shaw 1960). It soon became impossible to maintain the dichotomy in the analysis of insurance firms, with insurance operations on the one hand and financial investment on the other. As a result, insurance research has become deeply influenced by advances in finance theory. This was heightened by the fact that finance underwent a major revolution in the 1970s, with the development of option theory, and that this revolution stressed the similarity between insurance products and new concepts coming from financial innovations (e.g., portfolio insurance).[67]

6.1 Portfolio Theory and the CAPM

The influence of portfolio theory on the analysis of insurance demand is mentioned in Sect. 3. But this theory also has a profound influence on the theory of insurance supply. It was soon recognized that financial intermediaries could be analyzed as a joint portfolio of assets and liabilities (Michaelsen and Goshay 1967), and this global approach was applied to insurance company management. Under this view, insurers have to manage a portfolio of correlated insurance liabilities and investment assets, taking into account balance-sheet and solvency constraints, and there is no justification for separating the operations into two separate areas: what matters is the overall return on equity (see Kahane and Nye 1975; Kahane 1977).[68]

This way of looking at insurance operations led to a theory of insurance rating, reflecting the move observed a decade earlier in finance, from portfolio theory to the CAPM. Applying this model to insurance, it turns out that equilibrium insurance prices will reflect the undiversifiable risk of insurance operations. If insurance risks are statistically uncorrelated with financial market risk, then equilibrium insurance prices are given by the present value of expected claims costs (in the absence of transaction costs). If they are statistically correlated, a positive *or negative* loading is observed in equilibrium. The model was developed by Biger and Kahane (1978), Hill (1979), and Fairley (1979).

[66] See Boubakri (2011), Boubakri and Valéry (2025), Anderloni et al. (2020), and the September 2011 special issue of *The Journal of Risk and Insurance* for recent surveys of corporate governance in the insurance industry.

[67] The similarity between option contracts and insurance policies was stressed by Briys and Loubergé (1983).

[68] See also Loubergé (1983) for an application to international reinsurance operations, taking foreign exchange risk into account, and MacMinn and Witt (1987) for a related model.

It was empirically evaluated by Cummins and Harrington (1985). It was also applied to determine the "fair" regulation of insurance ratings in Massachusetts (Hill and Modigliani 1986).[69]

In their recent review of the literature, Bauer et al. (2025) discussed in detail the difficulties of reconciling theory and practice in insurance pricing models. Particular attention was given to the CAPM model. First, the CAPM model is a one-period model, while insurance cash flows cover many years in many lines of business. A partial solution proposed by Myers and Cohn (1987) is to use the discounted net present value of loss payments. This solution involves estimating a liability beta to obtain a discount rate. This introduces a new difficulty, as only accounting information is available for many insurance liabilities.

Another solution is to use the cost of capital measured by the internal rate of return (IRR) approach. As shown by Bauer et al. (2023), the IRR model can be reconciled with the CAPM model when the discount rates of assets, liabilities, and equity are comparable. In fact, there is no need to compute the discount rate of liabilities if we know the two other rates. However, the IRR approach is not devoid of technical problems. Bauer et al. (2025) discussed other approaches, such as the risk-adjusted discounted cash flow model and the present value of income over the present value of equity. In conclusion, they mentioned that it is still difficult to divide insurance policy premiums into general market risk changes from financial markets and frictional costs associated with holding capital.

6.2 Option Pricing Theory

Another limitation of the capital asset pricing model is that it does not take into account nonlinearities arising from features such as limited liability and asymmetric tax schedules. These aspects are best analyzed using option pricing theory since it is well known that optional clauses imply nonlinearities in portfolio returns. Doherty and Garven (1986) and Cummins (1988) analyzed the influence of limited liability and default risk on insurance prices, while Garven and Loubergé (1996) studied the effects of asymmetric taxes on equilibrium insurance prices and reinsurance trade among risk-neutral insurers. A major implication of these studies is that loaded premiums not only reflect transaction costs and asymmetric information, or insurers' risk aversion, they also reflect undiversifiable risk arising from institutional features, and they lead to prices implying risk-sharing in equilibrium, even when market participants are risk neutral.

The importance of option theory for the economics of insurance has also been observed in the area of life insurance. This resulted from the fact that competition between insurers and bankers, in order to attract savings, has led to the inclusion of numerous optional features (hidden options) in life insurance contracts. Advances in option theory have thus often been used to value life insurance contracts (see, e.g., Brennan and Schwartz 1976; Ekern and Persson 1996; Nielsen and Sandmann 1996) or to assess the effects of life insurance regulation (Briys and de Varenne 1994). It seems, however, that the option pricing of insurance has not attracted much research over the last 20 years.

6.3 Insurance and Corporate Finance

The portfolio approach to insurance demand led to a paradox when applied to corporations because these are owned by stockholders who are able to diversify risks in a stock portfolio. If insurance risks, such as accidents and fires, are diversifiable in the economy, the approach leads to the

[69] Myers and Cohn (1987) extended the model to multiperiod cash flows, while Kraus and Ross (1982) considered the application to insurance of the more general arbitrage pricing theory.

conclusion that corporations should not bother insuring them. They would increase shareholders' wealth by remaining uninsured instead of paying loaded premiums (Mayers and Smith 1982).[70] The paradox was solved using the modern theory of corporate finance, where the firm is considered a nexus of contracts between various players: managers, employees, suppliers, bondholders, banks, stockholders, consumers, etc. A reduction in contracting and bankruptcy costs provides an incentive to manage risk and to purchase insurance, even if the premium is loaded and shareholders are indifferent to insurance risk: see Main (1982), Mayers and Smith (1982, 1990), and Stulz (1984). In addition, increasing the marginal cost of external financing and convex tax schedules arising from progressive tax rates and incomplete loss offset offer other explanations for concern with insurance risk management in widely held corporations: see Froot et al. (1993), Smith and Stulz (1985), Smith et al. (1990), Tufano (1996), and Stulz (1996).

These considerations have changed the relationship of corporate managers to risk management in general and to insurance in particular. These tools are no longer used simply because risks arise. They must find a justification in the firm's overall objective of value maximization. Following these premises, several studies have addressed the relationship between corporate risk management and capital-structure decisions or dividend policies. For instance, Auñon-Nerin and Ehling (2008) found that higher leverage increases the demand for corporate insurance and the use of derivatives, while hedging generally has a significant positive effect on leverage (see also Zou and Adams 2008). This is in line with corporate concern for default and agency costs. They also found that corporate hedging is negatively related to the dividend payout ratio. This is related to the use of cash flows as a possible substitute for insurance (in this aspect, see also Rochet and Villeneuve (2011)).[71] In addition, as Doherty (1997) noted, the development of financial engineering in the 1980s challenged traditional insurance strategies in corporate risk management. Traditional insurance strategies often involve large transaction costs, and they fail if the risk is not diversifiable, as in the case of the US liability crisis. For this reason, innovative financial procedures, such as finite risk plans and financial reinsurance, represent alternative instruments for dealing with corporate risk. Of course, they also widen the competitive interface between banks and insurers.

The theory of corporate finance was also used by Garven (1987) to study the capital-structure decisions of insurance firms. His paper shows that redundant tax shields, default risk, bankruptcy costs, and the abovementioned agency costs influence the insurer's capital-structure decision. Plantin (2006) emphasized that reinsurance purchases and capital-structure decisions are linked. Professional reinsurers are used by insurers as a signal of credible monitoring sent to the financial market. Reinsurance is used not only as a device to mutualize risks, as in Borch (1962), but also to address agency problems, as in Mayers and Smith et al. (1990). It is also used as a complement to the insurer's capital-structure strategy.

At the empirical level, Garven and Lamm-Tenant (2003), as well as Powell and Sommer (2007), provided evidence that reinsurance purchases are positively related to insurer leverage. Shiu (2011) used data from the UK nonlife insurance industry to test the two-way relationship between reinsurance and insurers' capital structure. His results show that leverage exerts a positive influence on reinsurance purchases and that higher leverage is associated with more reinsurance purchases. He also found that the use of financial derivatives by insurers has a moderating impact on this two-way relationship.

More recently, Desjardins et al. (2022) studied the reciprocal causal relationship between reinsurance demand and liquidity creation by insurers. Financial institutions improve economic

[70] The same kind of argument was used by Doherty and Tinic (1981) to question the motivation of reinsurance demand by insurers.

[71] Rochet and Villeneuve (2011) showed that cash-poor firms should hedge using financial derivatives but not insure, whereas the opposite is true for cash-rich firms.

growth by creating liquidity in the economy or by investing in various assets or projects. Their generalized method of moments (GMM) estimation results show positive bicausal effects between the two activities for small insurers, but no significant effects for large insurers who do not seem to use reinsurance to protect their liquidity risk associated with long-term investments.

Taking a more general view, Hoyt and Liebenberg (2011) investigated whether an integrated risk management approach, as defined by Meulbroek (2002), has a positive influence on insurers' value. They used Tobin's Q as a measure of firm value and a maximum likelihood procedure to estimate joint equations for the determinants of an integrated risk management policy and its impact on Tobin's Q for a sample of 117 publicly traded US insurers. They found that insurers engaged in integrated risk management "are valued roughly 20% higher than other insurers after controlling for other value determinants and potential endogeneity bias" (p. 810). They also found that these insurers are larger, have less leverage, and rely less on reinsurance than other insurers.[72]

This result, added to those obtained above by Shiu (2011), can be related to earlier considerations by Doherty (1997) that insurers' management has been deeply influenced by developments in financial markets. The concept of asset-liability management, which has its roots in the portfolio approach mentioned above, means that insurers are less reliant on reinsurance as the natural instrument to hedge their risks and send signals to their partners. Indeed, developments in financial markets over the past 20 years have resulted in the emergence of derivative products intended to complement traditional reinsurance treaties from an integrated risk management view (Cummins and Weiss 2009).

6.4 Insurance and Financial Markets

In 1973, the insurance-banking interface was a sensitive subject. It was generally not well considered, in the insurance industry, to state that insurance is a financial claim and that insurers and bankers perform related functions in the economy. Some 50 years later, and after numerous experiences of mergers and agreements between banks and insurers, the question is not whether the two activities are closely related[73] but where they differ.[74]

It is easy for an economist of risk and insurance to provide a general answer to this question, founded on Borch's mutuality principle (see Sect. 2) and on subsequent work on risk sharing. Insurance and banking, like all financial activities, are concerned with the transfer of money across the two-dimensional space of time and states of nature. Insurance deals mainly—but not exclusively (see life insurance)—with transfers across states that do not necessarily involve a change in social wealth. In contrast, banking and financial markets perform transfers across states that often involve a change in social wealth. In other words, insurance is mainly concerned with diversifiable risk; banks and finance companies (such as mutual funds and hedge funds) are mainly concerned with undiversifiable (social) risk.

This kind of distinction has been used before to draw a line between private and public (social) insurance. According to this view, social insurance is called for when the limits of private insurability

[72] See Liebenberg and Hoyt (2025) for a recent review of this literature, and Cummins et al. (2009) for the effect of risk management on the efficiency of insurance firms.

[73] The convergence between reinsurance and investment banking was emphasized by Cummins (2005).

[74] The debate has regained importance after the 2008 financial market crisis and the collapse of AIG. Large insurance companies have been ranked with banks in the group of "Systemic Important Financial Institutions" (SIFI) and are threatened to become subject to the same regulations as banks. This is an occasion for the insurance industry to underscore the differences between banking and insurance (see Lehmann and Hofmann 2010; Geneva Association 2010).

are reached, in the sense that the insured events are not independent, so that diversifiability does not obtain: epidemics, losses from natural disasters, unemployment, etc.[75] But, social insurance is limited by national boundaries. In the absence of redistributive concerns and with market incompleteness due to asymmetric information, it has become increasingly obvious that financial markets are able to perform some social insurance functions in addition to their traditional function of sharing production risk.

A case in point is the evolution of the natural catastrophe branch of insurance. Since losses from natural disasters are correlated, they should be excluded from the private insurance area. Nonetheless, private insurance companies used to cover this risk because geographical dispersion seemed possible using the international reinsurance market. However, over the last two or three decades, the private insurability of this risk has been challenged by various developments such as the increased frequency and severity of hurricanes, huge losses, and a concentration of insured values in selected exposed areas of the globe: the US (mainly California and Florida), Japan, and Western Europe (mainly in the south). As a result, potential losses may have affected the financial capacity of the catastrophe reinsurance market (see Kielholz and Durrer 1997; Cummins et al. 2002). On the other hand, Dionne and Desjardins et al. (2022) pointed out that the capacity of the US insurance market has evolved over recent years. They showed that the capacity to pay for catastrophe losses was higher in 2020 than in 1997. This is explained by an increase in reinsurance demand and by higher premium rates for that risk. It is fair to say, however, that this increased capacity may not persist in the long run. Reinsurance is under pressure due to global warming and its observed impact on the severity and frequency of natural disasters around the world. Global warming reduces reinsurers' traditional role in international diversification. This calls for the development of new tools able to cope with climate risk. ILSs are considered an appropriate tool to complement reinsurance, but their international outstanding volume, at about $40B, is still very far from the needs expected for future years (Artemis website).[76]

A first possible solution to the insurability problem is traditional recourse to government insurance using increased taxation, i.e., social insurance. This is the solution adopted in France (Magnan 1995) and in some other countries,[77] where a reserve fund financed by specific taxes on property-liability insurance contracts indemnifies victims from natural disasters. The viability of this solution is however endangered in the long run by increasing risk due to global warming and by wrong incentives: development of activities and construction in areas exposed to catastrophe risk and pressure on the government to widen the scope of coverage while maintaining low rates.

A second solution is transferring the risk using special-purpose derivative markets. This was the solution proposed by the Chicago Board of Trade (CBOT) with the catastrophe options and futures contracts launched in December 1992: see D'Arcy and France (1992), Cummins and Geman (1995), and Aase (1999) for an analysis of these contracts.[78] However, the CBOT contracts were withdrawn after some years due to lack of success.[79] Following Hurricane Katrina in 2005, new contracts of

[75] Public insurance may also be justified by equity considerations, e.g., in medical insurance.

[76] Adding collateralized reinsurance, ILWs, sidecars, and other financial coverage products, the outstanding international volume is at about $60B (Artemis website).

[77] In the US, where a National Flood Insurance program has already existed for long, and where California has established a government earthquake insurance program (the California Earthquake Authority), the possible creation of state or regional catastrophe funds is being hotly debated, given the unconvincing examples of the two abovementioned programs (see Klein and Wang 2009).

[78] The early options and futures on four narrow-based indices of natural catastrophes were replaced in October 1995 by call spreads on nine broad-based indices. Lewis and Murdock (1996) proposed to have the same kind of contract supplied by federal authorities in order to complete the reinsurance market.

[79] Harrington and Niehaus (1999) had reached the conclusion that basis risk would not be a significant problem for Property Claims Services (PCS) derivative contracts, but later on, Cummins et al. (2004) reached a different conclusion, attributing the lack of success to basis risk. One might add that, possibly, the failure was due to the absence of arbitrage

the same type were nevertheless launched in 2007 by the Chicago Mercantile Exchange (CME) and the Insurance Futures Exchange. A main difference from the earlier CBOT contracts was their focus on US hurricanes and US tropical winds. Given the experience with the CBOT contracts, experts had doubts about the ultimate success of this new venture (see Cummins 2012). But the New York Mercantile Exchange of the CME still provides such options and futures derivative contracts under the group of "weather contracts." The geographical coverage was extended to Europe and Japan, even if the volumes are close to zero, meaning that these contracts are unlikely to substantially complement reinsurance capacity in the future.[80]

A third solution is the securitization of the risk, using more familiar securities, such as coupon bonds, issued by a special-purpose vehicle (on behalf of an insurer, reinsurer, or nonfinancial company) or by a public agency (on behalf of the State): see Litzenberger et al. (1996) and Loubergé et al. (1999) for early presentations and analyses of insurance-linked bonds (widely known as CAT bonds). In a CAT bond arrangement, a special-purpose reinsurer (SPR) issues a coupon bond on behalf of the sponsoring entity and improves the return on the bonds with a premium paid by the sponsor. The principal is then invested in first-class securities, such as government bonds. However, on the investor side, the principal and the coupons are at risk, in the sense that they may be lost partially or even totally if a catastrophe occurs and the CAT bond is triggered. In such a case, the proceeds of the investment are used by the SPR to pay indemnities to the sponsor. The catastrophe risk has been transferred to the financial market using familiar securities as a transfer vehicle. In contrast to CBOT derivatives, these ILSs were well received by the market. Their success was based on the huge pool of financial capacity provided by worldwide capital markets and the prospect of risk diversification made available to investors: catastrophic insurance losses are, in principle, uncorrelated with financial market returns. In addition, CAT bonds that have been based on an index of losses due to a specific catastrophic event, or triggered by such a specific event (parametric trigger), have made it possible to avoid the moral hazard arising from products based on the record of losses experienced by the sponsor.[81] On the supply side, CAT bonds provide sponsors with coverage that extends over several years, at fixed terms (unlike reinsurance contracts), and free from default risk since the proceeds from the bond issue are fully collateralized using highly rated securities.[82] CAT bonds have attracted broad interest among insurance practitioners (see Swiss Re 2009)[83] and academic researchers (see Barrieu and El Karoui 2002; Lee and Yu 2002; Nell and Richter 2004; Cummins 2008; Michel-Kerjan and Morlaye 2008; Barrieu and Loubergé 2009; Finken and Laux 2009).[84] The development of the market for these securities indicates that they filled a gap in the reinsurance market although the success has not been as huge as initially anticipated: CAT bond and ILS issues started with a few issues prior to 2000, then the market peaked with 31 issues in 2007, and regained momentum in 2010 (26 issues)

trading. Arbitrage trading between a derivative market and the market for the underlying instrument is essential to the provision of liquidity in derivatives trading for hedging and speculation purposes. However, in the case of PCS option contracts, such trading was impossible. The only market for the trading of insurance portfolios is the reinsurance market, which is not liquid enough for use as a vehicle in arbitrage trading.

[80] The pricing of these contracts was addressed by Perrakis and Boloorrforoosh (2018).

[81] Exposure to moral hazard for the investor is traded against basis risk for the sponsor.

[82] The risk of default by the reinsurance provider is a concern in the high-layer segment of the reinsurance market.

[83] Still, it remains that the use of insurance-linked securities raises sensitive issues in terms of regulation. Not because these instruments would represent a danger for the stability of the financial system, but because regulators, more particularly in the US, are reluctant to consider them as genuine alternative mechanisms for risk transfer: see Klein and Wang (2009).

[84] The success with CAT bonds aroused interest for other insurance-linked securities, particularly in the life insurance sector (mortality bonds, longevity bonds): see Cowley and Cummins (2005), Lin and Cox (2005), Albertini and Barrieu (2009), Cummins and Weiss (2009), and Chen and Cox (2009).

after a drop to 14 issues in 2008 due to the financial crisis. In 2022, 73 deals were observed, for a value of $10.5B (Artemis). The market is nevertheless developing over time, with positive and negative shocks being triggered by natural catastrophes and financial crises.[85]

A fourth solution available to an insurer to hedge catastrophe risk outside the reinsurance market is provided by Catastrophic Equity Puts (CatEPuts). Under this arrangement, the insurer or reinsurer purchases from the option writer the right to issue preferred stocks at a specific price following the occurrence of a catastrophe. This allows the insurers to take advantage of fresh funding at a predetermined cost in a situation where recourse to the capital market would be prohibitive for them. This illustrates the increased integration of insurance and investment banking, with both activities performing a fundamental economic function, i.e., the transfer of risk (Arnone et al. 2021).[86]

7 Conclusion

In the early 1970s, it was not clear how risk and insurance economics would develop over the years to come. Some 50 years later, it is comforting to realize that considerable developments have taken place: the length of the reference list below, unconventionally divided into pre-1973 and post-1973 references, gives an account of the qualitative and quantitative aspects of these developments.

This chapter focuses on developments that have mainly taken place along three avenues of research:

1. The theory of risk-taking behavior in the presence of multiple risks, which encompasses the theory of optimal insurance coverage, the theory of optimal portfolio investment, and the theory of optimal risk prevention. The dynamic behavior of economic agents over multiple periods is also considered.
2. The issues raised by asymmetric information for contract design and market equilibrium, a theme which extends beyond insurance economics and concerns all contractual relations in the economy, e.g., labor markets, product markets, and financial markets. Insurance economists have been very active in developing original empirical tests on the significance of asymmetric information, tests that can be applied in different markets.
3. The application of new financial paradigms, such as contingent claims analysis, to the analysis of insurance firms, insurance markets, and corporate risk management, a development that links insurance economics more closely to financial economics and insurance to finance.

Currently, risk and insurance economics represent a major theme in general economic theory. Risk and insurance education, per se, has also become a predominant theme in many countries. But risk and insurance issues are still pervasive in economic education, more particularly in microeconomics. To verify this statement, one needs only look at the post-1973 section of the following reference list to see that many important papers for the advancement of risk and insurance theory were published in general economic and financial journals and not only in the leading specialized reviews. Indeed, given that this goal of the 1970s was reached, one might wonder whether a different objective, that is, the development of more specialized risk and insurance education and research, which was given less importance back then, should not be reevaluated today.

[85] See Cummins (2012) for a comprehensive report on the state of the market at year-end 2011, and the recent survey by Barrieu et al. (2025).

[86] Other innovations, such as sidecars and ILWs, are different in nature from those presented in this section. They are innovations that improve the capacity of the reinsurance market, without introducing an alternative or complement to reinsurance contracts.

Given the tremendous amount of research activity we have witnessed in the study of financial markets over the past years, we could infer that specialized research in insurance economics would receive a major impulse from the creation of complete, reliable, and easily accessible insurance databases. Compared with the situation at the end of the 1990s, the last 30 years have been characterized by a breakthrough in empirical research in insurance economics, where the implications of models have been subject to empirical tests. These tests represent fundamental progress in the economics of risk and insurance. They provide results that enhance our understanding of insurance markets, and the authors must be congratulated for their efforts. But these contributions are still too often based on proprietary data, made available on a case-by-case basis, and not on widely available insurance databases, which reduces the power of the conclusions. Insurers are still reluctant to share their data. The availability of more databases would certainly trigger more interest in dissertations on risk and insurance from beginning doctoral students in economics and finance. The development of strong causality tests must be based on high-quality data (Dionne 2025).

Over the last 10 years, many new developments in the insurance literature have been observed. The solvency regulation of insurance firms became an important activity after the 2007–2009 financial crisis. Insurers must now be more transparent about their risk management activities, transmitting their yearly risk appetite statement to the regulator and having more board discussions about the computation of economic and solvency-regulated capital. This computation must now be based on market data instead of only actuarial data. Better dynamic stress tests of capital must also be implemented.

Natural catastrophe risk has been omnipresent in insurer losses in recent years. Some insurers have left the market, and important questions about the international coverage of these losses via reinsurance have been raised because global warming limits international diversification. The development of alternative financial coverage is positive but not very significant, at about 10% of the total reinsurance capital. Economic inflation is not well integrated in the management of insurance firms, along with social inflation costs. Cyber-risk costs and pandemic costs are becoming new sources of sizable losses. It is not clear that the insurers' traditional risk management tools are effective in limiting these large losses and expenses, which are often correlated.

The insurance market faces many new innovations that affect the management of insurance and reinsurance companies. The Internet and digital insurance are becoming significant management tools for insurance policy management. Machine learning and telematic information are used to improve safety monitoring and fraud auditing. In life insurance, unit-linked and equity-indexed contracts have been important innovative products in recent years. Private equity in life insurance markets is also growing, and the effect of interest rates on variable annuities must be better understood.

References

Publications until 1973

Akerlof GA (1970) The market for "lemons": quality uncertainty and the market mechanism. Q J Econ 84:488–500

Arrow KJ (1953) Le rôle des valeurs boursières pour la répartition la meilleure des risques. In: Econométrie, Fondements et applications de la théorie du risque en économétrie. CNRS, Paris, pp 41–47. English version (1964): The role of securities in the optimal allocation of risk-bearing. Rev Econ Stud 31:91–96

Arrow KJ (1963) Uncertainty and the welfare economics of medical care. Am Econ Rev 53:941–973

Arrow KJ (1970) Insurance, risk and resource allocation. In: Arrow KJ (ed) Essays in the theory of risk bearing. North Holland, pp 134–143

Borch K (1960) The safety loading of reinsurance premiums. Scand Actuar J 43:163–184

Borch K (1962) Equilibrium in a reinsurance market. Econometrica 30:424–444

Carter RL (1972) Economics and insurance. PH Press

Debreu G (1959) Theory of value. Wiley

Ehrlich J, Becker G (1972) Market insurance, self-insurance and self-protection. J Polit Econ 80:623–648
Friedman M, Savage LJ (1948) The utility analysis of choices involving risk. J Polit Econ 56:279–304
Gould JP (1969) The expected utility hypothesis and the selection of optimal deductibles for a given insurance policy. J Bus 42:143–151
Greene M (1971) Risk aversion, insurance and the future. Indiana University Press
Greene M (1973) Risk and insurance. South Western
Gurley JG, Shaw ES (1960) Money in theory of finance. Brookings, Washington, DC
Hammond JD (1968) Essays in the theory of risk and insurance. Scott Foresman
Joskow P (1973) Cartels, competition, and regulation in the property-liability insurance industry. Bell J Econ 4(2):375–427
Kihlstrom RE, Pauly M (1971) The role of insurance in the allocation of risk. Am Econ Rev 61:371–379
Lintner J (1965) Security prices, risk and maximal gain from diversification. J Financ 20:587–615
Mahr W (1964) Einführung in die Versicherungswirtschaft. Duncker & Humblot
Markowitz HM (1959) Portfolio selection – efficient diversification of investments. Wiley
Mehr R, Hedges B (1963) Risk management in the business enterprise. Irwin
Michaelsen JB, Goshay RC (1967) Portfolio selection in financial intermediaries: a new approach. J Financ Quant Anal 2:166–199
Mossin J (1966) Equilibrium in a capital asset market. Econometrica 34:768–783
Mossin J (1968) Aspects of rational insurance purchasing. J Polit Econ 79:553–568
Pashigian BP, Schkade L, Menefee GH (1966) The selection of an optimal deductible for a given insurance policy. J Bus 39:35–44
Pauly M (1968) The economics of moral hazard: comment. Am Econ Rev 58:531–536
Pfeffer I (1956) Insurance and economic theory. Irwin
Pratt J (1964) Risk aversion in the small and in the large. Econometrica 32:122–136
Ross S (1973) The economic theory of agency: the principal's problem. Am Econ Rev 63:134–139
Rothschild M, Stigliz J (1970) Increasing risk: I. A definition. J Econ Theory 2:225–243
Savage LJ (1954) Foundation of statistics. Wiley
Sharpe W (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19:425–442
Smith V (1968) Optimal insurance coverage. J Polit Econ 79:68–77
Spence M (1973) Job market signaling. Q J Econ 87:355–374
Spence M, Zeckhauser R (1971) Insurance, information and individual action. Am Econ Rev 61:380–387
Von Neumann J, Morgenstern O (1947) Theory of games and economic behavior. Princeton University Press
Wilson R (1968) The theory of syndicates. Econometrica 36:113–132

Publications after 1973

Aase K (1999) An equilibrium model of catastrophe insurance futures and spreads. Geneva Risk Insur Rev 24:69–96
Abbring J, Chiappori PA, Pinquet J (2003) Moral hazard and dynamic insurance data. J Eur Econ Assoc 1:767–820
Albertini L, Barrieu P (2009) The handbook of insurance-linked securities. Wiley, New York
Allard M, Cresta JP, Rochet JC (1997) Pooling and separating equilibria in insurance markets with adverse selection and distribution costs. Geneva Risk Insur Rev 22:103–120
Allen F (1985) Repeated principal-agent relationships with lending and borrowing. Econ Lett 17:27–31
Ambrose JM, Carroll AM, Regan L (2013) The economics of liability insurance. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, pp 315–347
Ames DA, Hines CS, Sankara J (2018) Board risk committees: insurer financial strength rating and performance. J Account Public Policy 37(2):130–145
Anderloni L, Moro O, Tanda A (2020) Governance and performance in insurance companies: a bibliometric analysis and a meta-analysis. Int J Econ Financ 12(11):1–20
AON (2023) Weather, climate and catastrophe insight. https://www.aon.com/weather-climate-catastrophe/index.aspx
Arnone M, Bianchi ML, Quaranta AG, Tassinari GL (2021) Catastrophic risks and the pricing of catastrophe equity put options. Comput Manag Sci 18:213–237
Arnott R (1992) Moral hazard and competitive insurance markets. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 325–358
Arnott R, Stiglitz JE (1990) The welfare economics of moral hazard. In: Loubergé H (ed) Risk, information and insurance: essays in the memory of Karl Borch. Kluwer Academic Publishers, pp 91–121
Arrow KJ (1974) Optimal insurance and generalized deductibles. Scand Actuar J 1:1–42
Arrow KJ (1978) Risk allocation and information: some recent theoretical developments. Geneva Pap Risk Insur Issues Pract 8:5–19

Auñon-Nerin D, Ehling P (2008) Why firms purchase property insurance. J Financ Econ 90:298–312
Baker T, Siegelman P (2013) The law and economics of liability insurance: a theoretical and empirical review. In: Arlen JH, Paige NZ (eds) Research handbook on the economics of torts. Elgaronline
Barrese J, Doerpinghaus H, Nelson J (1995) Do independent agent insurers provide superior service? The insurance marketing puzzle. J Risk Insur 62:297–308
Barrieu P, El Karoui N (2002) Reinsuring climatic risk using optimally designed weather bonds. Geneva Risk Insur Rev 27:87–113
Barrieu P, Loubergé H (2009) Hybrid cat bonds. J Risk Insur 76:547–578
Barrieu P, Braun A, Makariou D (2025) Catastrophe bonds. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Bauer D, Powell L, Su B, Zanjani GH (2023) Life insurance and annuity pricing during the financial crisis, revisited. Available on SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3799325
Bauer D, Phillips RD, Zanjani GH (2025) Pricing insurance risk: reconciling theory and practice. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Berger A, Cummins D, Weiss M (1997) The coexistence of multiple distribution systems for financial services: the case of property-liability insurance. J Bus 70:515–546
Berry-Stölzle T, Born P (2012) The effect of regulation on insurance pricing: the case of Germany. J Risk Insur 79:129–164
Biener C, Eling M (2012) Insurability in microinsurance markets: an analysis of problems and potential solutions. Geneva Pap Risk Insur Issues Pract 37:77–107
Biger N, Kahane Y (1978) Risk considerations in insurance ratemaking. J Risk Insur 45:121–132
Bleichrodt H (2022) The prevention puzzle. Geneva Risk Insur Rev 47(2):277–297
Bonato D, Zweifel P (2002) Information about multiple risks: the case of building and content insurance. J Risk Insur 69:469–487
Bond EW, Crocker KJ (1991) Smoking, skydiving and knitting: the endogenous categorization of risks in insurance markets with asymmetric information. J Polit Econ 99:177–200
Bond EW, Crocker KJ (1997) Hardball and the soft touch: the economics of optimal insurance contracts with costly state verification and endogenous monitoring costs. J Public Econ 63:239–264
Borch K (1974) Capital markets and the supervision of insurance companies. J Risk Insur 41:397–405
Born P, Klimaszewski-Blettner B (2013) Should I stay or should I go? The impact of natural disasters and regulation on U.S. property insurers' supply decisions. J Risk Insur 80(1):1–36
Born P, Viscusi WK (2006) The catastrophic effects of natural disasters on insurance markets. J Risk Uncertain 33(1/2):55–72
Boubakri N (2011) Corporate governance and issues from the insurance industry. J Risk Insur 78:501–518
Boubakri N, Valéry P (2025) Corporate governance issues from the insurance industry: updated review. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Boubakri N, Dionne G, Triki T (2008) Consolidation and value creation in the insurance industry: the role of governance. J Bank Financ 32:56–68
Boyer M, Dionne G (1989a) More on insurance, protection and risk. Can J Econ 22:202–205
Boyer M, Dionne G (1989b) An empirical analysis of moral hazard and experience rating. Rev Econ Stat 71:128–134
Boyer M, Jacquier E, Van Norden S (2012) Are underwriting cycles real and forecastable. J Risk Insur 79(4):995–1015
Brennan MJ, Schwartz E (1976) The pricing of equity-linked life insurance policies with an asset value guarantee. J Financ Econ 3:195–213
Briys E (1988) On the theory of rational insurance purchasing in a continuous time model. Geneva Pap Risk Insur Issues Pract 13:165–177
Briys E, de Varenne F (1994) Life insurance in a contingent claims framework: pricing and regulatory implications. Geneva Risk Insur Rev 19:53–72
Briys E, Loubergé H (1983) Le contrat d'assurance comme option de vente. Finance 4:139–153
Briys E, Loubergé H (1985) On the theory of rational insurance purchasing. J Financ 40:577–581
Briys E, Schlesinger H (1990) Risk aversion and the propensities for self-insurance and self-protection. South Econ J 57:458–467
Briys E, Kahane Y, Kroll Y (1988) Voluntary insurance coverage, compulsory insurance, and risky-riskless portfolio opportunities. J Risk Insur 55:713–722
Briys E, Dionne G, Eeckhoudt E (1989) More on insurance as a Giffen good. J Risk Uncertain 2:420–425
Briys E, Schlesinger H, von Schulenburg JM (1991) Reliability of risk management: market insurance, self-insurance and self-protection reconsidered. Geneva Risk Insur Rev 16:45–58
Caballé J, Pomansky A (1996) Mixed risk aversion. J Econ Theory 71:485–513
Cagle J, Harrington SE (1995) Insurance supply with capacity constraints and endogenous insolvency risk. J Risk Uncertain 11:219–232
Cardon J, Hendel I (2001) Asymmetric information in health insurance: evidence from the National Health Expenditure Survey. RAND J Econ 32:408–427

Cawley J, Philipson T (1999) An empirical examination of information barriers to trade in insurance. Am Econ Rev 89:827–846
Chamberlain S, Tennyson S (1998) Capital shocks and merger activity in the property-liability insurance industry. J Risk Insur 65:563–595
Chang YM, Ehrlich I (1985) Insurance, protection from risk and risk bearing. Can J Econ 18:574–587
Chang PM, Peng JL, Fan CK (2011) A comparison of bancassurance and traditional insurer sales channels. Geneva Pap Risk Insur Issues Pract 36:76–93
Chen H, Cox S (2009) Modeling mortality with jumps: applications to mortality securitization. J Risk Insur 76:727–751
Chen R, Wong KA, Lee HC (1999) Underwriting cycles in Asia. J Risk Insur 66:29–47
Cheng J, Elyasani E, Jia J (2011) Institutional ownership stability and risk-taking: evidence from the life-health insurance industry. J Risk Insur 78:609–641
Chesney M, Loubergé H (1986) Risk aversion and the composition of wealth in the demand for full insurance coverage. Schweiz Z Volkswirtsch Stat 122:359–370
Chiappori PA, Salanié B (2000) Testing for asymmetric information in insurance markets. J Polit Econ 108:56–78
Chiappori PA, Salanié B (2013) Asymmetric information in insurance markets: predictions and tests. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, pp 397–422
Chiappori PA, Macho I, Rey P, Salanié B (1994) Repeated moral hazard: the role of memory, commitment and the access to credit markets. Eur Econ Rev 38:1527–1553
Chiappori PA, Durand F, Geoffard PY (1998) Moral hazard and the demand for physician services: first lessons from a French natural experiment. Eur Econ Rev 42:499–511
Choi S, Hardigree D, Thistle P (2002) The property-liability insurance cycle: a comparison of alternative models. South Econ J 68:530–548
Cohen A (2005) Asymmetric information and learning in the automobile insurance market. Rev Econ Stat 87:197–207
Cohen A, Dehejia R (2004) The effects of automobile insurance and accident liabilities laws on traffic fatalities. J Law Econ 47:357–393
Cohen A, Einav L (2007) Estimating risk preferences from deductible choice. Am Econ Rev 97(3):745–788
Cohen A, Siegelman P (2010) Testing for adverse selection in insurance markets. J Risk Insur 77:39–84
Cook PJ, Graham DA (1977) The demand for insurance production: the case of irreplaceable commodities. Q J Econ 91:143–156
Cooper R, Hayes B (1987) Multi-period insurance contracts. Int J Ind Organ 5:211–231
Courbage C, Rey B (2006) Prudence and optimal prevention for health risks. Health Econ 15:1323–1327
Courbage C, Rey B (2012) Optimal prevention and other risks in a two-period model. Math Soc Sci 63:213–217
Courbage C, Stahel W (2012) Extreme events and insurance: 2011 annus horribilis. Geneva reports no 5. Geneva Association
Courbage C, Rey B, Treich N (2013) Prevention and precaution. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, pp 185–204
Courbage C, Loubergé H, Peter R (2017) Optimal prevention for multiple risks. J Risk Insur 84(3):899–922
Cowley A, Cummins JD (2005) Securitization of life insurance assets and liabilities. J Risk Insur 72:193–226
Crocker KJ, Morgan J (1998) Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts. J Polit Econ 106:355–375
Crocker KJ, Snow A (1985) The efficiency of competitive equilibria in insurance markets with adverse selection. J Public Econ 26:207–219
Crocker KJ, Snow A (1986) The efficiency effects of categorical discrimination in the insurance industry. J Polit Econ 94:321–344
Crocker KJ, Snow A (2008) Background risk and the performance of insurance markets under adverse selection. Geneva Risk Insur Rev 33:137–160
Crocker KJ, Snow A (2011) Multidimensional screening in insurance markets with adverse selection. J Risk Insur 78:287–307
Crocker KJ, Rothschild C, Snow A (2025) The theory of risk classification. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Cummins JD (1988) Risk-based premiums for insurance guaranty funds. J Financ 43:823–839
Cummins JD (2005) Convergence in wholesale financial services: reinsurance and investment banking. Geneva Pap Risk Insur Issues Pract 30:187–222
Cummins JD (2007) Reinsurance for natural and man-made catastrophes in the United States: current state of the market and regulatory reforms. Risk Manag Insur Rev 10:179–220
Cummins JD (2008) Cat bonds and other risk-linked securities: state of the market and recent developments. Risk Manag Insur Rev 11:23–47
Cummins JD (2012) Cat bonds and other risk-linked securities: product design and evolution of the market. Available at https://ssrn.com/abstract=1997467 or https://doi.org/10.2139/ssrn.1997467
Cummins JD, Danzon P (1997) Price, financial quality, and capital flows in insurance markets. J Financ Intermed 6:3–38

Cummins JD, Doherty N (2002) Capitalization of the property-liability insurance industry: overview. J Financ Serv Res 21:5–14
Cummins JD, Doherty N (2006) The economics of insurance intermediaries. J Risk Insur 73:359–396
Cummins JD, Geman H (1995) Pricing catastrophe futures and call spreads. J Fixed Income 4:46–57
Cummins JD, Harrington SE (1985) Property-liability insurance rate regulation: estimation of underwriting betas using quarterly profit data. J Risk Insur 52:16–43
Cummins JD, Harrington SE (1987) The impact of rate regulation on property-liability insurance loss ratios: a cross-sectional analysis with individual firm data. Geneva Pap Risk Insur Issues Pract 12:50–62
Cummins JD, Mahul O (2003) Optimal insurance with divergent beliefs about insurer total default risk. J Risk Uncertain 27:121–138
Cummins JD, Nini G (2002) Optimal capital utilization by financial firms: evidence from the property-liability insurance industry. J Financ Serv Res 21:15–54
Cummins JD, Outreville JF (1987) An international analysis of underwriting cycles in property-liability insurance. J Risk Insur 54:246–262
Cummins JD, Vanderhei JL (1979) A note on the relative efficiency of property-liability insurance distribution systems. Bell J Econ 10:709–720
Cummins JD, Weiss M (2009) Convergence of insurance and financial markets: hybrid and securitized risk-transfer solutions. J Risk Insur 76:493–545
Cummins JD, Xie X (2008) Mergers and acquisitions in the US property-liability insurance industry: productivity and efficiency effects. J Bank Financ 32(1):30–55
Cummins JD, Tennyson S, Weiss M (1999a) Consolidation and efficiency in the US life insurance industry. J Bank Financ 23:325–357
Cummins JD, Weiss M, Zi H (1999b) Organizational form and efficiency: the coexistence of stock and mutual property-liability insurers. Manag Sci 45:1254–1269
Cummins JD, Phillips RD, Tennyson S (2001) Regulation, political influence and the price of automobile insurance. J Insur Regul 20:9–50
Cummins JD, Doherty N, Lo A (2002) Can insurers pay for the "big one"? Measuring the capacity of the insurance market to respond to catastrophic losses. J Bank Financ 26:557–583
Cummins JD, Lalonde D, Phillips RD (2004) The basis risk of index-linked catastrophic loss securities. J Financ Econ 71:77–111
Cummins JD, Dionne G, Gagné R, Nouira A (2009) Efficiency of insurance firms with endogenous risk management and financial intermediation activities. J Prod Anal 32(2):145–159
Cummins JD, Klumpes P, Weiss MA (2015) Mergers and acquisitions in the global insurance industry: valuation effects. Geneva Pap Risk Insur Issues Pract 40(3):444–473. ISSN 1018-5895
Cutler D, Reber S (1998) Paying for health insurance: the trade-off between competition and adverse selection. Q J Econ 113:433–466
D'Arcy SP (1988) Application of economic theories of regulation to the property-liability insurance industry. J Insur Regul 7:19–52
D'Arcy SP, Doherty N (1990) Adverse selection, private information and lowballing in insurance markets. J Bus 63:145–163
D'Arcy SP, France VG (1992) Catastrophe futures: a better hedge for insurers. J Risk Insur 59:575–601
Dachraoui K, Dionne G, Eeckhoudt L, Godfroid P (2004) Comparative mixed risk aversion: definition and application to self-protection and willingness to pay. J Risk Uncertain 29:261–276
Dahlby B (1983) Adverse selection and statistical discrimination: an analysis of Canadian automobile insurance market. J Public Econ 20:121–131
Dahlby B (1992) Testing for asymmetric information in Canadian automobile insurance. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 423–443
Danzon PM (1983) Rating bureaus in US property-liability insurance markets: anti or pro-competitive? Geneva Pap Risk Insur Issues Pract 8:371–402
Danzon PM, Harrington SE (1992) The demand for and supply of liability insurance. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 25–60
Dardanoni V, Forcina A, Li Donni P (2018) Testing for asymmetric information in insurance markets: a multivariate ordered regression approach. J Risk Insur 85(1):107–125
De Garidel-Thoron T (2005) Welfare-improving asymmetric information in dynamic insurance markets. J Polit Econ 113:121–150
De Meza D, Webb D (2001) Advantageous selection in insurance markets. RAND J Econ 32:249–262
Deloitte (2020) Climate risk: regulators sharpen their forces, p 20. https://www2.deloitte.com/gr/en/blog/risk-advisory/2020/financial-risks-stemming-from-climate-change.html
Derrig R (2002) Insurance fraud. J Risk Insur 69:271–287

Desjardins D, Dionne G, N'Golo K (2022) Reinsurance demand and liquidity creation: a search for bicausality. J Empir Financ 66:137–154
Dionne G (1982) Moral hazard and state-dependent utility function. J Risk Insur 49:405–423
Dionne G (1983) Adverse selection and repeated insurance contracts. Geneva Pap Risk Insur Issues Pract 8:316–333
Dionne G (2025) Causality in empirical analyses with emphasis on asymmetric information and risk management. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Dionne G, Desjardins D (2022) A re-examination of the U.S. insurance market's capacity to pay catastrophe losses. Risk Manag Insur Rev 25(4):515–549
Dionne G, Doherty N (1994) Adverse selection, commitment and renegotiation: extension to and evidence from insurance markets. J Polit Econ 102:209–235
Dionne G, Eeckhoudt L (1984) Insurance and saving: some further results. Insur Math Econ 3:101–110
Dionne G, Eeckhoudt L (1985) Self-insurance, self-protection and increased risk aversion. Econ Lett 17:39–42
Dionne G, Eeckhoudt L (1988) Increasing risk and self-protection activities. Geneva Pap Risk Insur Issues Pract 13:132–136
Dionne G, Gollier C (1992) Comparative statics under multiple sources of risk with applications to insurance demand. Geneva Risk Insur Rev 17:21–33
Dionne G, Harrington SE (1992) An introduction to insurance economics. In: Dionne G, Harrington SE (eds) Foundations of insurance economics. Kluwer Academic Publishers, pp 1–48
Dionne G, Lasserre P (1985) Adverse selection, repeated insurance contracts and announcement strategy. Rev Econ Stud 52:719–723
Dionne G, Lasserre P (1987) Dealing with moral hazard and adverse selection simultaneously. Working paper, University of Pennsylvania
Dionne G, Li J (2011) The impact of prudence on optimal prevention revisited. Econ Lett 113:147–149
Dionne G, Li J (2014) When can expected utility handle first-order risk aversion? J Econ Theory 154:403–422
Dionne G, Liu Y (2021) Effects of insurance incentives on road safety: evidence from a natural experiment in China. Scand J Econ 123(2):453–477
Dionne G, Rothschild C (2014) Economic effects of risk classification bans. Geneva Risk Insur Rev 39:184–221
Dionne G, St-Michel P (1991) Workers' compensation and moral hazard. Rev Econ Stat 73(1):236–244
Dionne G, Gourieroux C, Vanasse C (2001) Testing for evidence of adverse selection in the automobile insurance market: a comment. J Polit Econ 109:444–453
Dionne G, Giuliano F, Picard P (2009) Optimal auditing with scoring: theory and application to insurance fraud. Manag Sci 55:58–70
Dionne G, Michaud PC, Dahchour M (2013) Separating moral hazard from adverse selection and learning in automobile insurance: longitudinal evidence from France. J Eur Econ Assoc 11(4):897–917
Dionne G, Fombaron N, Mimra W (2025) Adverse selection in insurance. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Doff R (2025) Risk and solvency regulation in Europe. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Doherty N (1984) Portfolio efficient insurance buying strategies. J Risk Insur 51:205–224
Doherty N (1997) Corporate insurance: competition from capital markets and financial institutions. Assurances 65:63–94
Doherty N, Dionne G (1993) Insurance with undiversifiable risk: contract structure and organizational form of insurance firms. J Risk Uncertain 6:187–203
Doherty N, Eeckhoudt L (1995) Optimal insurance without expected utility: the dual theory and the linearity of insurance contracts. J Risk Uncertain 10:157–179
Doherty N, Garven JR (1986) Price regulation in property-liability insurance: a contingent claims approach. J Financ 41:1031–1050
Doherty N, Garven JR (1995) Insurance cycles: interest rates and the capacity constraint model. J Bus 68:383–404
Doherty N, Jung HJ (1993) Adverse selection when loss severities differ: first-best and costly equilibria. Geneva Risk Insur Rev 18:173–182
Doherty N, Kang HB (1988) Price instability for a financial intermediary: interest rates and insurance price cycles. J Bank Financ 12:191–214
Doherty N, Phillips RD (2002) Keeping up with the Joneses: changing rating standards and the buildup of capital by U.S. property-liability insurers. J Financ Serv Res 21:55–78
Doherty N, Posey L (1998) On the value of a checkup: adverse selection, moral hazard and the value of information. J Risk Insur 65:189–211
Doherty N, Richter A (2002) Moral hazard, basis risk, and gap insurance. J Risk Insur 69(1):9–24
Doherty N, Schlesinger H (1983a) Optimal insurance in incomplete markets. J Polit Econ 91:1045–1054
Doherty N, Schlesinger H (1983b) The optimal deductible for an insurance policy when initial wealth is random. J Bus 56:555–565

Doherty N, Schlesinger H (1990) Rational insurance purchasing: considerations of contract non-performance. Q J Econ 105:243–253
Doherty N, Schlesinger H (1995) Severity risk and the adverse selection of frequency risk. J Risk Insur 62:649–665
Doherty N, Schlesinger H (2002) Insurance contracts and securitization. J Risk Insur 69:45–62
Doherty N, Thistle P (1996) Adverse selection with endogenous information in insurance markets. J Public Econ 63:83–102
Doherty N, Tinic S (1981) Reinsurance under conditions of capital market equilibrium. J Financ 36:949–953
Doherty N, Loubergé H, Schlesinger H (1987) Risk premiums with multiple sources of risk. Scand Actuar J 1-2:41–49
Drèze J (1979) Human capital and risk-bearing. Geneva Pap Risk Insur Issues Pract 12:5–22
Drèze J (1990) The role of securities and labor contracts in the optimal allocation of risk-bearing. In: Loubergé H (ed) Risk, information and insurance. Kluwer Academic Publishers, pp 41–65
Dror DM, Eling M (2025) Microinsurance: promoting resilience and welfare in the informational sector. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Eckart M, Räthke-Döppner S (2010) The quality of insurance intermediaries services – empirical evidence for Germany. J Risk Insur 77:667–701
Eeckhoudt L, Gollier C (2005) The impact of prudence on optimal prevention. Economic Theory 26:989–994
Eeckhoudt L, Kimball M (1992) Background risk, prudence, and the demand for insurance. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 239–254
Eeckhoudt L, Gollier C, Schlesinger H (1991) Increases in risk and deductible insurance. J Econ Theory 55:435–440
Eeckhoudt L, Gollier C, Schlesinger H (1996) Changes in background risk and risk-taking behavior. Econometrica 64:683–689
Eeckhoudt L, Meyer J, Ormiston MB (1997) The interactions between the demand for insurance and insurable assets. J Risk Uncertain 14:25–39
Eeckhoudt L, Rey B, Schlesinger H (2007) A good sign for multivariate risk taking. Manag Sci 53:117–124
Eeckhoudt L, Huang R, Tzeng L (2012) Precautionary effort: a new look. J Risk Insur 79(2):585–590
Einav L, Finkelstein A (2011) Selection in insurance markets: theory and empirics in pictures. J Econ Perspect 25:115–138
Einav L, Finkelstein A, Cullen M (2010) Estimating welfare in insurance markets using variations in prices. Q J Econ 125:877–921
Eisen R (1990) Problems of equilibria in insurance markets with asymmetric information. In: Loubergé H (ed) Risk, information and insurance. Kluwer Academic Publishers, pp 123–141
Ekern S, Persson SA (1996) Exotic unit-linked life insurance contracts. In: Loubergé H, Subrahmanyam M (eds) Financial risk and derivatives. Kluwer Academic Publishers, pp 35–63
Eling M, Gemmo I, Guxha D, Schmeiser H (2024) Big data, risk classification, and privacy in insurance markets. Geneva Risk Insur Rev 49:75–126
Eling M, Schmeiser H (2010) Insurance and the credit crisis: impact and ten consequences for risk management and supervision. Geneva Pap Risk Insur Theory Pract 35:9–34
Eling M, Schmeiser H, Schmit J (2007) The Solvency II process: overview and critical analysis. Risk Manag Insur Rev 10:69–85
Ernst and Young (2008) Strategic business risk: insurance. https://erm.ncsu.edu/library/article/business-risk-2008/
Fagart MC, Picard P (1999) Optimal insurance under random auditing. Geneva Risk Insur Rev 24:29–54
Fairley W (1979) Investment income and profit margins in property-liability insurance: theory and empirical results. Bell J Econ 10:192–210
Fang H, Keane M, Silverman D (2008) Sources of advantageous selection: evidence from the medigap insurance market. J Polit Econ 116:303–350
Faure MG (2014) The complementary roles of liability regulation and insurance in safety management: theory and practice. J Risk Res 17:689–707
Finkelstein A, McGarry K (2006) Multiple dimensions of private information: evidence from the long-term care insurance market. Am Econ Rev 96:938–958
Finkelstein A, Poterba J (2002) Selection effects in the United Kingdom individual annuities market. Econ J 112:28–50
Finkelstein A, Poterba J (2004) Adverse selection in insurance markets: policyholder evidence from the U.K. annuity market. J Polit Econ 112:183–2008
Finken S, Laux C (2009) Catastrophe bonds and reinsurance: the competitive effect of information-sensitive triggers. J Risk Insur 76:579–605
Finsinger J, Pauly M (1984) Reserve levels and reserve requirements for profi-maximizing insurance firms. In: Bamberg G, Spremann K (eds) Risk and capital. Springer Verlag, pp 160–180
Fluet C, Pannequin F (1997) Complete versus incomplete insurance contracts under adverse selection with multiple risks. Geneva Risk Insur Rev 22:81–101
Froot K (1999a) The evolving market for catastrophic event risk. Risk Manag Insur Rev 2:1–28
Froot K (1999b) The financing of catastrophe risk. University of Chicago Press

Froot K (2001) The market for catastrophic risk: a clinical examination. J Financ Econ 60:529–571
Froot K, O'Connell P (2008) On the pricing of intermediated risks: theory and application to catastrophe reinsurance. J Bank Financ 32:69–85
Froot K, Sharfstein D, Stein J (1993) Risk management: coordinating corporate investment and financial policies. J Financ 48:1629–1658
Fudenberg D, Tirole J (1990) Moral hazard and renegotiation in agency contracts. Econometrica 58:1279–1319
Garven JR (1987) On the application of finance theory to the insurance firm. J Financ Serv Res 1:57–76
Garven JR, Lamm-Tenant J (2003) The demand for reinsurance: theory and tests. Insur Risk Manag 71:217–238
Garven JR, Loubergé H (1996) Reinsurance, taxes and efficiency: a contingent claims model of insurance market equilibrium. J Financ Intermed 5:74–93
Gatzert N, Wesker H (2012) A comparative assessment of Basel II/III and Solvency II. Geneva Pap Risk Insur Issues Pract 37(3):539–570
Geneva Association (2010) Systemic risk in insurance – an analysis of insurance and financial stability. The Geneva Association
Geyer A, Kremslehner D, Mürmann A (2020) Asymmetric information in automobile insurance: evidence from driving behavior. J Risk Insur 87(4):969–995
Gollier C (1992) Economic theory of risk exchanges: a review. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 3–23
Gollier C (1995) The comparative statics of changes in risk revisited. J Econ Theory 66:522–536
Gollier C (2003) To insure or not to insure? An insurance puzzle. Geneva Risk Insur Rev 28:5–24
Gollier C (2005) Some aspects of the economics of catastrophe risk insurance. CESifo Working Paper Series, No. 1409. Also available on SSRN https://ssrn.com/abstract=668384
Gollier C (2024) Evaluating sustainability actions under uncertainty: the role of improbable extreme scenarios. Geneva Risk Insur Rev 49:59–74
Gollier C, Pratt JW (1996) Risk vulnerability and the tempering effect of background risk. Econometrica 64:1109–1123
Gollier C, Scarmure P (1994) The spillover effect of compulsory insurance. Geneva Risk Insur Rev 19:23–34
Gollier C, Schlee E (2006) Increased risk-bearing with backroung risk. Top Theor Econ 6:1–31. Published by De Gruyter. https://doi.org/10.2202/1534-598X.1249
Gollier C, Schlesinger H (1995) Second best insurance contract design in an incomplete market. Scand J Econ 97:123–135
Gollier C, Schlesinger H (1996) Arrow's theorem on the optimality of deductibles: a stochastic dominance approach. Economic Theory 7:359–363
Grace MF (2023) Loss ratio dynamics. Risk Manag Insur Rev 26(3):1–24
Grace MF (2025) Regulation of insurance markets in the US. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Grace MF, Hotchkiss JL (1995) External impacts on the property-liability insurance cycle. J Risk Insur 62:738–754
Grenier R (2019) The role of catastrophe models in the evolution of the flood insurance market. Air Worldwide Corporation
Gron A (1994) Evidence of capacity constraints in insurance markets. J Law Econ 37:349–377
Grossman S, Hart OD (1983) An analysis of the principal-agent problem. Econometrica 51:7–45
Guiso L, Jappelli T (1998) Background uncertainty and the demand for insurance against insurable risks. Geneva Risk Insur Rev 23:7–27
Haley J (1993) A cointegration analysis of the relationship between underwriting margins and interest rates: 1930-1989. J Risk Insur 60:480–493
Hansmann H (1985) The organization of insurance companies: mutual versus stock. J Law Econ Org 1:125–153
Harrington SE (1984) The impact of rate regulation on prices and underwriting results in the property-liability insurance industry: a survey. J Risk Insur 51:577–617
Harrington SE (1987) A note on the impact of auto insurance rate regulation. Rev Econ Stat 69:166–170
Harrington SE (1988) Prices and profits in the liability insurance market. In: Litan R, Winston C (eds) Liability: perspectives and policy. The Brooking Institution, pp 42–100
Harrington SE (2009) The financial crisis, systemic risk, and the future of insurance regulation. J Risk Insur 76:785–819
Harrington SE, Danzon P (1994) Price-cutting in liability insurance markets. J Bus 67:511–538
Harrington SE, Niehaus G (1999) Basis risk with PCS catastrophe insurance derivative contracts. J Risk Insur 66:49–82
Harrington SE, Niehaus G (2002) Capital structure decisions in the insurance industry: stocks versus mutual. J Financ Serv Res 21:145–163
Harrington SE, Niehaus G, Yu T (2013) Insurance price volatility and underwriting cycles. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, pp 647–667
Harrison GW (2024) Risk preferences and risk perceptions in insurance experiments: some methodological challenges. Geneva Risk Insur Rev 49:127–161
Harrison H, Karolyi GA, Scheinkman JA (2020) Climate finance. Rev Financ Stud 33(3):1011–1023

He E, Sommer D (2011) CEO turnover and ownership structure: evidence from the US property-liability insurance industry. J Risk Insur 78:673–701
Hellwig MF (1987) Some recent developments in the theory of competition in markets with adverse selection. Eur Econ Rev 31:319–325
Hellwig MF (1988) A note on the specification of interfirm communication in insurance markets with adverse selection. J Econ Theory 46:154–163
Helpman E, Laffont JJ (1975) On moral hazard in general equilibrium. J Econ Theory 10:8–23
Hemenway D (1990) Propitious selection. Q J Econ 105:1063–1069
Hendel I, Lizzeri A (2003) The role of commitment in dynamic contracts: evidence from life insurance. Q J Econ 118:299–327
Henriet D, Klimenko N, Rochet JC (2016) The dynamics of insurance prices. Geneva Risk Insur Rev 41(1):2–18
Hill RD (1979) Profit regulation in property-liability insurance. Bell J Econ 10:172–191
Hill RD, Modigliani F (1986) The Massachusetts model of profit regulation in nonlife insurance: theory and empirical results. In: Cummins JD, Harrington SE (eds) Fair rate of return in property-liability insurance. Kluwer Academic Publishers
Hilliard JI, Tennyson S (2025) Insurance distribution systems. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Ho CL, Lai GC, Lee JP (2013) Organizational structure, board composition, and risk taking in the US property casualty insurance industry. J Risk Insur 80(1):169–209
Hofmann A, Peter R (2015) Multivariate prevention decisions: safe today or sorry tomorrow? Econ Lett 128:51–53
Hofmann A, Peter R (2016) Self-insurance, self-protection, and saving: on consumption smoothing and risk management. J Risk Insur 83(3):719–734
Holmstrom B (1979) Moral hazard and observability. Bell J Econ 10:74–91
Hong SK, Lew KO, MacMinn R, Brockett P (2011) Mossin's theorem given random initial wealth. J Risk Insur 78:309–324
Hosios AJ, Peters M (1989) Repeated insurance contracts with adverse selection and limited commitment. Q J Econ 104:229–253
Hoy M (1982) Categorizing risks in the insurance industry. Q J Econ 97:321–336
Hoy M, Robson RJ (1981) Insurance as a Giffen good. Econ Lett 8:47–51
Hoyt R, Liebenberg A (2011) The value of enterprise risk management. J Risk Insur 78:795–822
Hoyt R, Mustard D, Powell L (2006) The effectiveness of state legislation mitigating moral hazard: evidence from automobile insurance. J Law Econ 49:427–450
Huang LY, Lai G, McNamara M, Wang J (2011) Corporate governance and efficiency: Evidence from US property-liability insurance industry. J Risk Insur 78:519–550
Huberman G, Mayers D, Smith CW (1983) Optimal insurance policy indemnity schedules. Bell J Econ 14:415–426
Ippolito R (1979) The effects of price regulation in the automobile insurance industry. J Law Econ 22:55–89
Jametti M, von Ungern-Sternberg T (2010) Risk selection in natural-disaster insurance. J Inst Theor Econ 166(2):344–364
Jang Y, Hadar J (1995) A note on increased probability of loss and the demand for insurance. Geneva Risk Insur Rev 20:213–216
Jean-Baptiste E, Santomero A (2000) The design of private reinsurance contracts. J Financ Intermed 9:274–297
Jewitt I (1989) Choosing between risky prospects: the characterization of comparative statics results, and location independent risk. Manag Sci 35(1):60–70
Jullien B, Salanié B, Salanié F (1999) Should more risk averse agents exert more effort? Geneva Risk Insur Rev 24:19–28
Kahane Y (1977) Capital adequacy and the regulation of financial intermediaries. J Bank Financ 1:207–218
Kahane Y, Kroll Y (1985) Optimal insurance coverage in situations of pure and speculative risk and the risk-free asset. Insur Math Econ 4:191–199
Kahane Y, Nye DJ (1975) A portfolio approach to the property-liability insurance industry. J Risk Insur 42:579–598
Karni E (1992) Optimal insurance: a nonexpected utility analysis. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 217–238
Karni E (1995) Non-expected utility and the robustness of the classical insurance paradigm – discussion. In: Gollier C, Machina M (eds) Non-expected utility and risk management. Kluwer Academic Publishers, pp 51–56
Kessler D (2014) (Re)insurance is not systemic. J Risk Insur 81:477–488
Kielholz W, Durrer A (1997) Insurance derivatives and securitization: new hedging perspectives for the US cat insurance market. Geneva Pap Risk Insur Issues Pract 82:3–16
Kihlstrom RE, Romer D, Williams S (1981) Risk aversion with random initial wealth. Econometrica 49:911–920
Kim H, Kim D, Im S, Hardin JW (2009) Evidence of asymmetric information in the automobile insurance market: dichotomous versus multinomial measurement of insurance coverage. J Risk Insur 76(2):343–366
Kimball MS (1990) Precautionary saving in the small and the large. Econometrica 58(1):53–73

Klein R (2012) Principles for insurance regulation: an evaluation of current practices and potential reforms. Geneva Pap Risk Insur Issues Pract 37:175–199
Klein R, Wang S (2009) Catastrophe risk financing in the United States and the European Union: a comparative analysis of alternative regulatory approaches. J Risk Insur 76:607–637
Klein R, Phillips R, Shiu W (2002) The capital structure of firms subject to price regulation: evidence from the insurance industry. J Financ Serv Res 21:79–100
Klick J, Stratmann T (2007) Diabetes treatment and moral hazard. J Law Econ 50:519–538
Klumpes P (2022) Consolidation and efficiency in the major European insurance markets. Working paper, Accounting Research Group, Aalborg University Business School
Konrad K, Skaperdas S (1993) Self-insurance and self-protection: a non-expected utility analysis. Geneva Risk Insur Rev 18:131–146
Kraus A, Ross SA (1982) The determinants of fair profits for the property-liability insurance firm. J Financ 37:1015–1030
Kunreuther H (1996) Mitigating disaster losses through insurance. J Risk Uncertain 12:171–187
Kunreuther H (2018) All-hazards homeowners insurance: challenges and opportunities. Risk Manag Insur Rev 21(1):141–155
Kunreuther H, Michel-Kerjan E (2009) At war with the weather. MIT Press, Cambridge, MA
Kunreuther H, Pauly M (1985) Market equilibrium with private knowledge: an insurance example. J Public Econ 26:269–288
Kunreuther H, Pauly M (2004) Neglecting disaster: why don't people insure against large losses? J Risk Uncertain 28:5–21
Kunreuther H, Pauly M (2006) Rules rather than discretion: lessons from Hurricane Katrina. J Risk Uncertain 33:101–116
Lambert R (1983) Long-term contracts and moral hazard. Bell J Econ 14:441–452
Lamm-Tennant J, Weiss M (1997) International insurance cycles: rational expectations/institutional intervention. J Risk Insur 64:415–439
Landsberger M, Meilijson I (1996) Extraction of surplus under adverse selection: the case of insurance markets. J Econ Theory 69:234–239
Lee K (1998) Risk aversion and self-insurance-cum-protection. J Risk Uncertain 17(2):139–151
Lee K (2012) Background risk and self-protection. Econ Lett 114:262–264
Lee JP, Yu MT (2002) Pricing default-risky cat bonds with moral hazard and basis risk. J Risk Insur 69:25–44
Lehmann A, Hofmann D (2010) Lessons learned from the financial crisis for risk management: contrasting developments in insurance and banking. Geneva Pap Risk Insur Issues Pract 35:63–78
Lemaire J (1990) Borch's theorem: a historical survey of applications. In: Loubergé H (ed) Risk, information and insurance. Kluwer Academic Publishers, pp 15–37
Leng CC, Meier U (2006) Analysis of multinational underwriting cycles in property-liability insurance. J Risk Financ 7:146–159
Lewis CM, Murdock KC (1996) The role of government contracts in discretionary reinsurance markets for natural disasters. J Risk Insur 63:567–597
Li J (2011) The demand for a risky asset in the presence of a background risk. J Econ Theory 146:372–391
Liebenberg A, Hoyt R (2025) Enterprise risk management. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Lin J, Cox S (2005) Securitization of mortality risks in life annuities. J Risk Insur 72:227–252
Litzenberger R, Beaglehole D, Reynold C (1996) Assessing catastrophe reinsurance-linked securities as a new asset class. J Portf Manag 23:76–86
Loubergé H (1983) A portfolio model of international reinsurance operations. J Risk Insur 50:44–60
Loubergé H, Watt R (2008) Insuring a risky investment project. Insur Math Econ 42:301–310
Loubergé H, Kellezi E, Gilli M (1999) Using catastrophe-linked securities to diversify insurance risk: a financial analysis of cat bonds. J Insur Issues 22:125–146
Machina M (1982) Expected utility analysis without the independence axiom. Econometrica 50:277–323
Machina M (1995) Non-expected utility and the robustness of the classical insurance paradigm. In: Gollier C, Machina M (eds) Non-expected utility and risk management. Kluwer Academic Publishers, pp 9–50
MacMinn RD, Ren Y (2011) Mutual versus stock insurers: a synthesis of the theoretical and empirical research. J Insur 34(2):101–111
MacMinn RD, Witt RC (1987) A financial theory of the insurance firm under uncertainty and regulatory constraints. Geneva Pap Risk Insur Issues Pract 12:3–20
Magee S, Schilling C, Sheedy E (2019) Risk governance in the insurance sector – determinants and consequences in an international sample. J Risk Insur 86(2):381–413
Magnan S (1995) Catastrophe insurance system in France. Geneva Pap Risk Insur Issues Pract 77:474–480
Mahul O (2000) Optimal insurance design with random initial wealth. Econ Lett 69:353–358
Main B (1982) Business insurance and large, widely-held corporations. Geneva Pap Risk Insur Issues Pract 7:237–247

Makki S, Somwaru A (2001) Evidence of adverse selection in crop insurance markets. J Risk Insur 68:685–708
Marshall JM (1974) Insurance theory: reserves versus mutuality. Econ Inq 12:476–492
Marshall JM (1976) Moral hazard. Am Econ Rev 66:880–890
Mayers D, Smith CW (1981) Contractual provisions, organizational structure, and conflict control in insurance markets. J Bus 54:407–434
Mayers D, Smith CW (1982) On the corporate demand for insurance. J Bus 55:281–296
Mayers D, Smith CW (1983) The interdependence of individual portfolio decisions and the demand for insurance. J Polit Econ 91:304–311
Mayers D, Smith CW (1986) Ownership structure and control: the mutualization of stock life insurance companies. J Financ Econ 16:73–98
Mayers D, Smith CW (1988) Ownership structure across lines of property-casualty insurance. J Law Econ 31:351–378
Mayers D, Smith CW (1990) On the corporate demand for insurance: evidence from the reinsurance market. J Bus 63:19–40
Mayers D, Smith CW (2002) Ownership structure and control: property-casualty insurer conversion to stock charter. J Financ Serv Res 21:117–144
Meier U, Outreville JF (2006) Business cycles in insurance and reinsurance: the case of France, Germany and Switzerland. J Risk Financ 7:160–176
Menegatti M (2009) Optimal prevention and prudence in a two-period model. Math Soc Sci 58(3):393–397
Meulbroek L (2002) Integrated risk management for the firm: a senior manager's guide. J Appl Corp Financ 14:56–70
Meyer J (1992) Beneficial changes in random variables under multiple sources of risk and their comparative statics. Geneva Risk Insur Rev 17:7–19
Meyer D, Meyer J (1998) Changes in background risk and the demand for insurance. Geneva Risk Insur Rev 23:29–40
Meyer D, Meyer J (2004) A more reasonable model of insurance demand. In: Aliprantis CD et al (eds) Assets, beliefs and equilibria in economic dynamics – essays in honor of Mordecai Kurz. Springer, pp 733–742
Meyer J, Ormiston MB (1995) Demand for insurance in a portfolio setting. Geneva Risk Insur Rev 20:203–212
Michel-Kerjan E (2012) How resilient is your country? Nature 491:497. https://www.nature.com/news/how-resilient-is-your-country-1.11861
Michel-Kerjan E, Morlaye F (2008) Extreme events, global warming, and insurance-linked securities: how to trigger the "tipping point". Geneva Pap Risk Insur Issues Pract 33:153–176
Mills E (2009) A global review of insurance industry responses to climate change. Geneva Pap 34:323–359
Mimra W, Wambach A (2019) Contract withdrawals and equilibrium in competitive markets with adverse selection. Economic Theory 67:875–907
Miyazaki H (1977) The rat race and internal labor markets. Bell J Econ 8:394–418
Moffet D (1977) Optimal deductible and consumption theory. J Risk Insur 44:669–683
Moffet D (1979) The risk-sharing problem. Geneva Pap Risk Insur Issues Pract 4:5–13
Monti A (2011) Public-private initiative to cover extreme events. In: Courbage C, Stahel WR (eds) Extreme events and insurance: 2011 annus horribilis. The Geneva reports – risk and insurance research. Geneva Association, pp 27–38
Mookherjee D, Png I (1989) Optimal auditing, insurance and redistribution. Q J Econ 104:205–228
Mormino CA (1979) Insurance cycles: an Italian experience. Études et Dossiers de l'Association de Genève 33:37–85
Munch P, Smallwood DE (1980) Solvency regulation in the property-liability insurance industry: empirical evidence. Bell J Econ 11:261–282
Munich Re (2019) Risk report. https://www.munichre.com/content/dam/munichre/mrwebsiteslaunches/2019-annual-report/MunichRe-Group-AR-2019-en_Risk-Report.pdf/_jcr_content/renditions/original./MunichRe-Group-AR-2019-en_Risk-Report.pdf
Myers SC, Cohn RA (1987) Insurance rate regulation and the capital asset pricing model. In: Cummins JD, Harrington SE (eds) Fair rate of return in property-liability insurance. Kluwer Academic Publishers, Norwell
Nell M, Richter A (2004) Improving risk allocation through indexed cat bonds. Geneva Pap Risk Insur Issues Pract 29:183–201
Nielsen JA, Sandmann K (1996) Uniqueness of the fair premium for equity-linked life insurance contracts. In: Loubergé H, Subrahmanyam M (eds) Financial risk and derivatives. Kluwer Academic Publishers, pp 65–102
OECD (2005) Catastrophic risk and insurance. OECD, Paris
Olivella P, Vera-Hernández M (2013) Testing for asymmetric information in private health insurance. Econ J 123(567):96–130
Parra A, Winter RA (2025) Optimal insurance contracts under moral hazard. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Pauly M (1974) Overinsurance and public provision of insurance: the role of moral hazard and adverse selection. Q J Econ 88:44–62
Pauly M, Kleindorfer PR, Kunreuther H (1986) Regulation and quality competition in the US insurance industry. In: Finsinger J, Pauly M (eds) The economics of insurance regulation. Palgrave Macmillan UK, London, pp 65–107
Perrakis S, Boloorrforoosh A (2018) Catastrophe futures and reinsurance contracts. J Futur Mark 38(1):104–128

Peter R (2017) Optimal self-protection in two periods: on the role of endogenous saving. J Econ Behav Organ 137:19–36
Peter R (2021a) Who should exert more effort? Risk aversion, downside risk aversion and optimal prevention. Economic Theory 71(4):1259–1281
Peter R (2021b) Prevention as a Giffen good. Econ Lett 208:110052
Peter R (2022) When is safety a normal good? Working paper, University of Iowa. Available at SSRN 44147815
Peter R (2024) The economics of self-protection: a tribute to EGRIE. Geneva Risk Insur Rev 49:6–35
Peter R (2025) The self-selection problem. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Picard P (1996) Auditing claims in insurance markets with fraud: the credibility issue. J Public Econ 63:27–56
Picard P (2025) Economic analysis of insurance fraud. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer
Plantin G (2006) Does reinsurance need reinsurers? J Risk Insur 73:153–168
Polemarchakis H (1990) Competitive allocation when the asset market is incomplete. Geneva Risk Insur Rev 15:5–16
Powell LS, Sommer DW (2007) Internal versus external markets in the insurance industry: the role of reinsurance. J Financ Serv Res 31:173–188
Priest GL (1996) The government, the market and the problem of catastrophic losses. J Risk Uncertain 12:219–237
Puelz R, Snow A (1994) Evidence on adverse selection: equilibrium signalling and cross-subsidization in the insurance market. J Polit Econ 102:236–257
Quiggin JC (1982) A theory of anticipated utility. J Econ Behav Organ 3:323–343
Radner R (1981) Monitoring cooperative agreements in a repeated principal-agent relationship. Econometrica 49:1127–1148
Raviv A (1979) The design of an optimal insurance policy. Am Econ Rev 69:84–86
Razin A (1976) Rational insurance purchasing. J Financ 31:133–137
Rea SA (1992) Insurance classifications and social welfare. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers, pp 377–396
Rees R, Gravelle H, Wambach A (1999) Regulation of insurance markets. Geneva Risk Insur Rev 24:55–68
Rey B (2003) A note on optimal insurance in the presence of a nonpecuniary background risk. Theor Decis 54:73–83
Richaudeau D (1999) Automobile insurance contracts and risk of accident: an empirical test using French individual data. Geneva Risk Insur Rev 24:97–114
Riley JG (1979) Informational equilibrium. Econometrica 47:331–359
Robinson C, Zheng B (2010) Moral hazard, insurance claims and repeated insurance contracts. Can J Econ 43:967–993
Rochet JC, Villeneuve S (2011) Liquidity management and corporate demand for hedging and insurance. J Financ Intermed 20:303–323
Ross S (1981) Some stronger measures of risk aversion in the small and in the large with applications. Econometrica 49:621–638
Rothschild C (2011) The efficiency of categorical discrimination in insurance markets. J Risk Insur 78:267–285
Rothschild C (2024) Welfare analysis in insurance markets with adverse selection. Geneva Risk Insur Rev 49:36–58
Rothschild M, Stiglitz JE (1970) Increasing risk: I – A definition. J Econ Theory 2:225–243
Rothschild M, Stiglitz JE (1976) Equilibrium in competitive insurance markets: the economics of markets with imperfect information. Q J Econ 90:629–650
Rothschild M, Stiglitz JE (1997) Competition and insurance twenty years later. Geneva Risk Insur Rev 22:73–79
Rowell D, Connelly L (2012) A history of the term "moral hazard". J Risk Insur 79:1051–1075
Rubinstein A, Yaari ME (1983) Repeated insurance contracts and moral hazard. J Econ Theory 30:74–97
Saito K (2006) Testing for asymmetric information in the automobile insurance market under rate regulation. J Risk Insur 73:335–356
Sandroni A, Squintani F (2007) Overconfidence, insurance and paternalism. Am Econ Rev 97:1994–2004
Schlee E (1995) The comparative statics of deductible insurance in expected- and non-expected utility theories. In: Gollier C, Machina M (eds) Non-expected utility and risk management. Kluwer Academic Publishers, pp 57–72
Schlesinger H (1984) Optimal insurance for irreplaceable commodities. J Risk Insur 51:131–137
Schlesinger H (1997) Insurance demand without the expected utility paradigm. J Risk Insur 64:19–39
Schlesinger H (1999) Decomposing catastrophic risk. Insur Math Econ 24:95–101
Schlesinger H (2013) The theory of insurance demand. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, pp 167–184
Schlesinger H, Doherty N (1985) Incomplete markets for insurance: an overview. J Risk Insur 52:402–423
Segal U, Spivak A (1990) First order versus second order risk aversion. J Econ Theory 51:111–125
Shavell S (1979) On moral hazard and insurance. Q J Econ 93:541–562
Shavell S (1982) On liability and insurance. Bell J Econ 13:120–132
Shavell S (1986) The judgment proof problem. Int Rev Law Econ 6:45–58
Shavell S (2000) On the social function and the regulation of liability insurance. Geneva Pap Risk Insur Issues Pract 25:166–179
Shavell S (2018) Liability for accidents in the New Palgrave Dictionary of Economics. Palgrave MacMillan, Springer, London, pp 7797–7803

Shim J (2011) Mergers & acquisitions, diversification and performance in the U.S. property-liability insurance industry. J Financ Serv Res 39:119–144
Shiu YM (2011) Reinsurance and capital structure: evidence from the United Kingdom non-life insurance industry. J Risk Insur 78:475–494
Sinn HW (1982) Kinked utility and the demand for human wealth and liability insurance. Eur Econ Rev 17:149–162
Smart M (2000) Competitive insurance markets with two unobservables. Int Econ Rev 41:153–169
Smith C, Stulz R (1985) The determinants of firms' hedging policies. J Financ Quant Anal 20:391–405
Smith C, Smithson W, Wilford DS (1990) Financial engineering: why hedge? In: Handbook of financial engineering. Harper & Row, pp 126–137
Spence M (1978) Product differentiation and performance in insurance markets. J Public Econ 10:427–447
Stiglitz JE (1977) Monopoly, non-linear pricing and imperfect information: the insurance market. Rev Econ Stud 44:407–430
Stiglitz JE (1983) Risk, incentives and insurance: the pure theory of moral hazard. Geneva Pap Risk Insur Issues Pract 8:4–33
Stiglitz JE, Weiss A (1981) Credit rationing in markets with imperfect information. Am Econ Rev 71:393–410
Stulz RM (1984) Optimal hedging policies. J Financ Quant Anal 19:127–140
Stulz RM (1996) Rethinking risk management. J Appl Corp Financ 9(3):8–25
Sweeney GH, Beard R (1992) The comparative statics of self-protection. J Risk Insur 59(2):301–309
Swiss Re (2009) The role of indices in transferring insurance risks to capital markets. Sigma 4
Tibiletti L (1995) Beneficial changes in random variables via copulas: an application to insurance. Geneva Risk Insur Rev 20:191–202
Townsend R (1979) Optimal contracts and competitive contracts with costly state verification. J Econ Theory 22:265–293
Tsetlin I, Winkler RL (2005) Risky choices and correlated background risk. Manag Sci 51:1336–1345
Tufano P (1996) Who manages risk? An empirical examination of risk management practices in the gold mining industry. J Financ 51:1097–1137
Turnbull S (1983) Additional aspects of rational insurance purchasing. J Bus 56:217–229
Venezian E (1985) Ratemaking methods and profit cycles in property and liability insurance. J Risk Insur 52:477–500
Villeneuve B (2003) Concurrence et antisélection multidimensionnelle en assurance. Ann Econ Stat 69:119–142
Viscusi WK (1995) Insurance and catastrophes: the changing role of the liability system. Geneva Risk Insur Rev 20:177–184
Von Schulenburg M (1986) Optimal insurance purchasing in the presence of compulsory insurance and insurable risks. Geneva Pap Risk Insur Issues Pract 38:5–16
Wang J, Li J (2015) Precautionary effort: another trait for prudence. J Risk Insur 82(4):977–983
Weiss M (2007) Underwriting cycles: a synthesis and further directions. J Insur Issues 30(1):31–45
Wharton Risk Management Center (2007) Managing large-scale risks in a new era of catastrophe. Wharton School, Philadelphia
Wilson C (1977) A model of insurance markets with incomplete information. J Econ Theory 12:167–207
Winter RA (1994) The dynamics of competitive insurance markets. J Financ Intermed 3:379–415
Yaari M (1987) The dual theory of choice under risk. Econometrica 55:95–115
Young VR, Browne MJ (1997) Explaining insurance policy provisions via adverse selection. Geneva Risk Insur Rev 22:121–134
Zanjani G (2002) Pricing and capital allocation in catastrophe insurance. J Financ Econ 65:283–305
Zeckhauser R (1995) Insurance and catastrophes. Geneva Risk Insur Rev 20:157–175
Zou H, Adams M (2008) Debt capacity, cost of debt, and corporate insurance. J Financ Quant Anal 43:433–466
Zweifel P, Ghermi P (1990) Exclusive vs. independent agencies: a comparison of performance. Geneva Risk Insur Rev 15:171–192

Part II

Climate Risk Protection

Fifty Years of U.S. Natural Disaster Insurance Policy

Kendra Marcoux and Katherine R. H. Wagner

1 Introduction

In 2022, the United States experienced over $140 billion of natural disaster damages. Globally, damages exceeded $270 billion (Straub et al. 2023). This year is not an anomaly, but rather reflects a trend in increasingly severe weather events. Natural disaster insurers are struggling to respond to the rate of increase of costs, with over 200 bills related to natural disaster insurance reform currently discussed at various levels of government in the US, while homeowners are experiencing increasing difficulty obtaining affordable insurance coverage (Wagner 2022b; Wing et al. 2018).

Large natural disasters are not a new phenomenon in the United States, and have caused insurance market upheaval in the past. The Great Mississippi Flood of 1927 is one such example, after which private insurers who previously sold flood coverage walked away from the flood insurance market completely. Their justification for doing so was that the massive losses they had sustained and would inevitably experience in the future required raising insurance premiums to unaffordable levels. These private insurance companies were replaced in 1968 by the National Flood Insurance Program (NFIP), a public program that sells government-backed flood insurance at subsidized rates to homeowners throughout the country (Gaul 2019). Fast forward fifty years and flood insurance in the US is provided almost exclusively through the NFIP today.

However, the scale and scope of current and predicted future damages have fostered concern about whether markets designed fifty years ago are appropriate for adapting to climate change risks. Over its fifty-year history, the NFIP has amassed a $40 billion revenue shortfall, relying on its ability to borrow from the US. Treasury and its funding by federal taxpayers to survive where private insurers could

Contact: Marcoux: University of California at Berkeley, Giannini Hall, Berkeley CA 94720, USA. Wagner: University of British Columbia, 6000 Iona Drive, Vancouver BC V6T2E8, Canada. We thank Patrick Baylis, Benjamin Collier, Brian Copeland, Carolyn Kousky, and Philip Mulder for helpful discussions.

K. Marcoux
UC Berkeley, Berkeley, CA, USA
e-mail: kendra_marcoux@berkeley.edu

K. R. H. Wagner (✉)
University of British Columbia, Vancouver, BC, Canada
e-mail: katherine.wagner@ubc.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_2

not (Gaul 2019). And the costly events that sank private flood insurance markets decades ago are now arriving in markets for natural disasters other than floods. In recent years, for example, insurers have started to pull out of areas at high risk of wildfire in the western US, leading policymakers to question whether natural disaster insurance risk is insurable by private markets, or at all (Cignarale et al. 2019; Rothschild and Stiglitz 1976). The general divergence between risk and premiums in natural disaster insurance markets has created uncertainty about the future solvency of insurers in a world affected by climate change and the ability of homeowners to insure themselves against its effects.

The goal of this chapter is to provide an overview of how insurers and homeowners are responding to challenges in natural disaster insurance markets in the US, and to highlight areas of fruitful future study. Our intent is to survey both public and private markets, many of which focus on different, specific disasters; while literature on the NFIP has burgeoned in recent years, our analysis places the NFIP both in a historical context and relative to other markets that have received less attention. Though we limit our discussion here to U.S. natural disaster insurance markets on which we have both conducted research, many of the behaviors of insurers and homeowners are common to international markets that are also experiencing the effects of climate change.

Why review natural disaster insurance separately from other insurance markets? A first reason is that these markets have received limited attention in comparison with other insurance markets, such as health, unemployment, and disability insurance. And yet, the rate at which costs are increasing in natural disaster insurance markets suggests an urgency in understanding insurer and homeowner behavior to undertake efficient reforms. A second reason to focus on analysis of natural disaster risk is that it is not clear that conclusions from other, more commonly studied insurance markets generalize. Natural disaster risk is different from other types of insurable risk because losses are highly spatially correlated, less frequent, and larger than most other types of insurable risk, such as health (Wagner 2022b). The resulting difficulty in predicting risk and the potential downside of getting the predictions wrong changes both insurer and homeowner behavior, as we will discuss.

We begin our study with discussion of the supply side of natural disaster insurance markets in Sect. 2. The public NFIP provides a starting point for our analysis of the challenges faced by insurers even in the presence of large geographic diversification and seemingly unlimited borrowing potential. We then discuss how these challenges differ in private markets for other natural disaster insurance types that are also complicated by greater insolvency risk, more stringent rate regulations, and more limited ability to diversify risk across space. In Sect. 3, we review how homeowner behavior on the demand side of the market complicates these challenges faced by insurers. Section 4 comprises analysis of how changes in natural disaster insurance markets affect other related markets, such as housing. We conclude in Sect. 5 with a discussion of important areas for future work.

2 Supply of Natural Disaster Insurance

This section reviews the supply side of natural disaster insurance markets in the US. We distinguish between public and private markets, which face different challenges. We discuss potential market failures that arise in both of these settings and policy solutions that could address them.

2.1 Public Insurance Markets

We begin with an analysis of public natural disaster insurance markets in the US. The largest, and perhaps best known, such market is the NFIP, which services almost all demand for flood insurance in the country. In addition, several other public natural disaster insurance markets exist in individual

states, for different disasters. All of these markets share common benefits and challenges that are characteristic of publicly provided natural disaster insurance, which we discuss in detail here.

The motivations for public natural disaster insurance markets have a theoretical basis. Natural disasters are infrequent and catastrophic, and the ongoing effects of climate change on the distribution of risk are hard to predict. Weitzman (2009) discusses how this uncertainty surrounding the distribution of climate change risk creates challenges for calculating the probabilities of natural disasters and, by extension, actuarially fair insurance premiums. Rothschild and Stiglitz (1976) claim more strongly that their standard model of insurer behavior does not apply to natural disaster risk and that the potential uninsurability of such risks by private markets justifies public provision. Their insight is that it is difficult to credibly model natural disaster insurers as risk neutral, as is typical in insurance markets, because premiums and claims are rarely in balance: In some years, no natural disaster occurs, while losses in other years can be catastrophic. By contrast, other insurance lines, such as health and automobile insurance, typically break even in any given year (Michel-Kerjan 2010). Private natural disaster insurers would have to build up a large amount of liquid assets in case a loss occurs—which is characteristic of risk-averse behavior. Institutional features of the tax code and legal framework in the US discourage the accumulation of large quantities of cash for rainy days; the lack of financial reserves was a contributing factor to the exit of private flood insurers in the early twentieth century, for example (Gaul 2019). Since then, the expansion of reinsurance markets has allowed insurers to reduce their exposure to natural disaster risk, a topic to which we return in Sect. 4.2 below (OECD 2018). While purchasing such reinsurance and reforming capital markets to encourage the accumulation of such reserves are possible solutions, an alternative is for the government to supply insurance directly (Jaffee and Russell 1997).[1]

The greater geographic diversification and solvency potential of governments suggest that public provision of natural disaster insurance may have advantages relative to private markets.[2] A public insurer can more easily smooth risk over a larger geography, therefore reducing the risk that any individual event results in bankruptcy (Oh et al. 2021). They also potentially have access to greater financing capacity than any individual private insurer does, allowing them to better weather large loss events, such as Hurricane Katrina, by borrowing from taxpayers.

The NFIP exemplifies these advantages of public natural disaster insurance markets. As a federal program administered by the Federal Emergency Management Agency (FEMA), they write over 5 million annual flood insurance contracts for homeowners nationwide, insuring both buildings and their contents. Figure 1 shows that the number of contracts sold by the NFIP has increased substantially over the program's lifetime. Their geographically dispersed market allows them to cross-subsidize claims in one part of the country with contemporaneous premium revenues from unflooded places. A distinguishing feature of natural disaster insurance markets is that claims are highly correlated in space, and so greater geographic diversification increases the amount of premium revenue available to offset claims in disaster areas. Today, the NFIP collects over $4.6 billion of annual premium revenue from customers across every state and insures over $1.3 trillion of assets (Horn 2022; Michel-Kerjan 2010).

The NFIP also benefits from seemingly unlimited borrowing potential. Flood insurance premium revenue has remained relatively constant in real terms since 1990, but flood insurance claims, by

[1] Jaffee and Russell (1997) propose several other capital markets solutions, including reforming accounting standards to permit tax-deductible catastrophe reserves and relaxing regulatory constraints that currently prohibit large premium increases. We return to the topic of financial market reform in more detail below.

[2] The benefits of public provision of insurance here are in addition to possible welfare gains from addressing textbook market failures, such as adverse selection, that have been studied extensively in other settings. See Einav et al. (2021) for a review of selection in insurance markets.

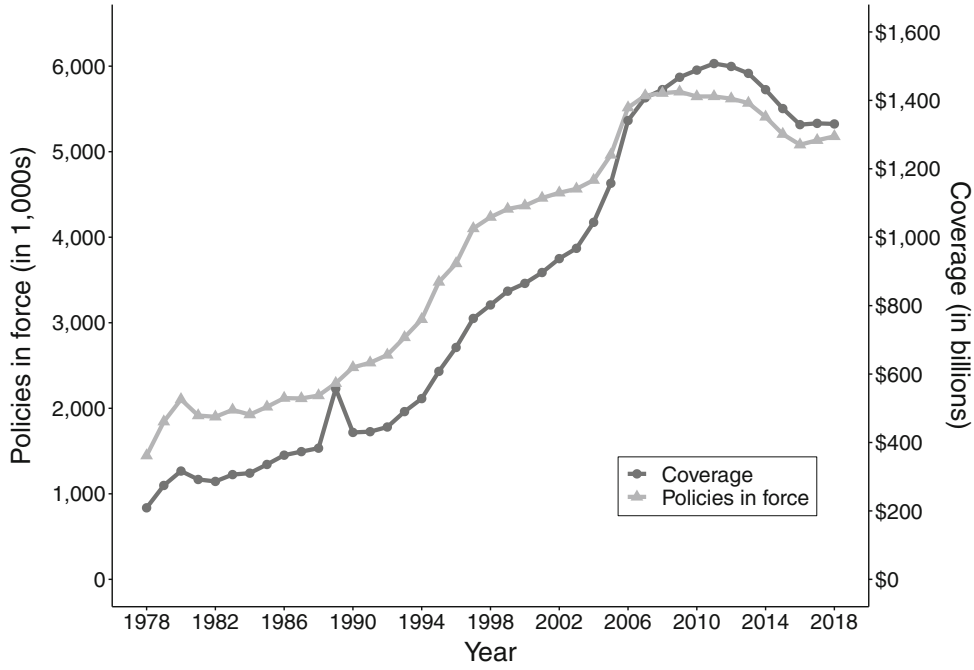


Fig. 1 History of NFIP Policy Counts and Coverage: This figure shows the total number of policies sold and total dollar value of coverage written by the National Flood Insurance Program, for each year from 1978–2018. All dollar values are in 2018 USD. Data are from the Federal Emergency Management Agency

contrast, are highly variable (Kelly 2017). Unlike health risks, natural disaster insurance payouts are infrequent but are more likely to be catastrophic if they happen (Jaffee and Russell 1997). Since seventeen of the most costly flood events in the US have occurred since 2000, claims have eclipsed premiums even with the NFIP's geographic diversification (Gaul 2019); this increasing annual premium shortfall is shown in Fig. 2. The program is currently over $20 billion in debt, but it continues to stay afloat by borrowing from the Department of the Treasury to finance their payouts; this number excludes the $16 billion of debt that Congress forgave in 2017. Under these circumstances, any private insurance market would conceivably have been declared insolvent years ago. By contrast, Congress continues to increase the NFIP's borrowing limit, which currently exceeds $30 billion and is 1500% higher in real terms than its initial 1968 level (Horn 2022).

In addition to the NFIP, several other U.S. public natural disaster insurance markets exist. Many of these other markets were also created in response to a large loss event that triggered the exit of private insurers. For example, the California Earthquake Authority (CEA) formed in 1996 after the 1994 Northridge earthquake precipitated the exit of private homeowners insurers from the earthquake line; this state-run earthquake insurer fills the void left in the private market by selling earthquake-specific policies throughout California at actuarially fair rates (Marshall 2017). The public Florida Citizens Property Insurance Corporation serves as an insurer of last resort for Florida homeowners who cannot obtain property and casualty coverage in private markets due to high perceived risk from flooding, wind, and other natural perils; the original market was established after high claims in private markets for damage caused by Hurricane Andrew in 1992 (Citizens 2020). The Louisiana Citizens Property Insurance Corporation and the California Fair Access to Insurance Requirements (FAIR) Plan Association similarly publicly supply homeowners' insurance to residual markets in their respective states; the latter has become increasing important in the last five years as private insurers

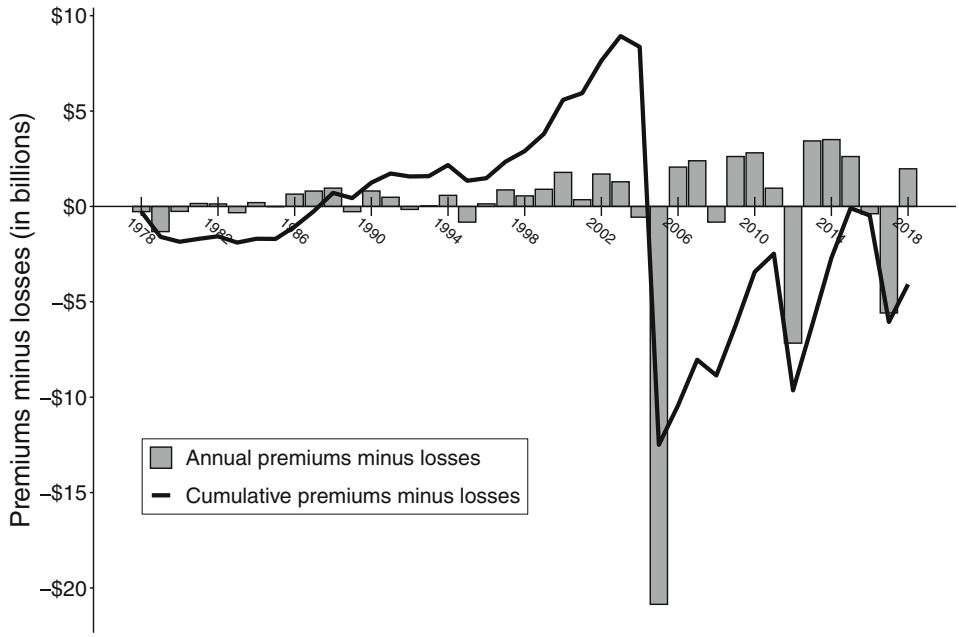


Fig. 2 History of NFIP Premiums and Losses: This figure shows the National Flood Insurance Program's annual premium shortfall or surplus. All dollar values are in 2018 USD. Data are from the Federal Emergency Management Agency

in California struggle to bear rising fire costs, as we will discuss in Sect. 2.2 (Cignarale et al. 2019). While many of these markets are smaller in scope, they share the characteristics and potential benefits common to public natural disaster insurance markets.

Given these potential benefits of public markets, why is widescale reform of the NFIP currently discussed? Regulatory concerns are centered on the high debt burden that the NFIP has incurred despite their ability to smooth risk across localities throughout the country. Many of today's challenges stem from the original design of the NFIP. Initial flood insurance premiums were set to cover losses incurred in an average loss year, and it was understood that the program would not break even during catastrophes as a result and would borrow from Congress. During the interim period between the exit of private flood insurers in the 1920s and the inception of the NFIP in 1968, the federal government had already been providing public disaster relief to victims of floods; the creation of the NFIP formalized this process. The program's original dual objectives were to increase the availability of flood insurance coverage and to manage development in floodplains. To accomplish the first objective, the program included explicit premium subsidies, below actuarially fair levels, on existing construction in high-risk flood zones to maintain affordability of flood insurance contracts for homeowners for whom actuarially fair rates would be very high (Michel-Kerjan 2010).[3] To accomplish the second objective, the NFIP also worked with the U.S. Army Corps of Engineers to produce Flood Insurance Rate Maps (FIRMs) that delineate areas at high and low risk of flooding; new construction in high-risk flood zones is required to meet minimum standards to reduce their flood risk or otherwise pay very high actuarially fair premiums. These minimum construction standards came into effect in 1974 and mandate that the elevation of houses' foundations exceeds the height of the

[3] The NFIP further increased these subsidies in 1972 and again in 1974 to encourage uptake. These additional subsidies were partially repealed in 1981 (AIR 2005).

flood that has a 1% probability of occurring (Hovekamp and Wagner 2023).[4] The gradual mapping of the US expanded the NFIP's scope to include almost all of the country today.

Since 1968, it has proven difficult to make changes to this original set-up. Together, political opposition to increasing premiums to reflect changes in flood risk and modeling constraints on determining what those risks actually are help explain why flood insurance premiums have remained largely constant in real terms for over half of the NFIP's history (Kelly 2017). In response to rising costs for the older structures grandfathered into the program at lower rates, Congress attempted to phase out statutory subsidies in 2012 through the Biggert-Waters Flood Insurance Reform Act. Specifically, this bill proposed to remove the subsidies for houses that were built before flood maps came into effect at a rate of 25% per year if the house was sold, grandfathered into a riskier flood zone, or classified as "severe repetive loss" by the NFIP.[5] However, the spatial nature of these premium increases leads to widespread concern from homeowners that coastal properties would experience substantial devaluation and from local governments that property tax revenues would fall. These premium changes were ultimately so unpopular that they were rescinded after only one year and replaced with more modest increases in a second bill, the Homeowners Flood Insurance Affordability Act, in 2013. This alternative legislation implemented more modest premium increases of between 5% and 18% for all properties receiving explicit subsidies, until premiums reached levels that the government's flood risk models suggested were actuarially fair. The new bill excluded severe repetitive loss properties, which were still subject to the original rate increases of 25% (NRC 2015).

Updating flood maps is challenging for the same political reasons, but also for scientific ones. Political opposition to updating natural disaster risk maps arises because real estate markets and natural disaster insurance markets are inextricably linked, and so rezoning houses into high-risk flood zones reduces their value (Hino and Burke 2021).[6] However, a more binding constraint on using NFIP flood maps to price flood insurance is that many of the maps do not in fact reflect current risk levels. Approximately two-thirds of official flood maps are out-of-date and do not incorporate the latest flood risk information available to the government; approximately 20% of maps were last updated more than 15 years ago (GAO 2021a). The scale at which these maps are produced also means that many of the historical government flood risk models use coarse-resolution terrain data and simplifying assumptions about the physics of local hydrology, including the effects of extreme precipitation events that are becoming more common.[7] Recent research using independent flood risk models shows that there are more than three times as many houses in high-risk flood zones as the NFIP's maps suggest (Wing et al. 2018). One change that the NFIP did make was the implementation of a mandatory flood insurance purchase requirement for homeowners with federally backed mortgages (Michel-Kerjan 2010). However, even this mandate has been difficult to enforce in practice and evidence supporting its effect on take-up is mixed (Petrolia et al. 2013); we return to demand-side challenges in Sect. 3.

How might these supply-side rate setting and information challenges be addressed? There is increasing acknowledgment that continually increasing the NFIP's borrowing from the Treasury will be untenable if flood insurance costs continue to increase; the NFIP has actually begun to purchase reinsurance products to reduce the Treasury's exposure to the contracts they have already written. In

[4] These construction standards apply in areas that have a "1-in-100 year" or higher chance of flooding, which the NFIP refers to as "Special Flood Hazard Areas" (Horn and Brown 2018).

[5] Severe repetitive loss properties are those that have made at least four claims in excess of $5,000 or at least two claims that, in total, exceed the property's value (Horn and Brown 2018). Grandfathered properties are structures that were remapped from a low-risk flood zone to a high-risk one and kept their low-risk rates (NRC 2015).

[6] Outside of the flood context, the state of Oregon withdrew new wildfire risk maps last year in response to public backlash from homeowners who opposed the zoning changes (Profita 2022).

[7] The NFIP's coarse rate schedule reflects their lack of granular data on flood risk. Premiums are set based on a handful of structural characteristics, including construction year, flood zone, foundation elevation, and basement depth.

addition, the federal government has committed over four billion dollars to update NFIP flood maps to provide accurate information about current risk levels through the 2009 FEMA Risk Mapping, Assessment and Planning Program. Since issuing a new flood map can take five to seven years, modernizing all flood maps is a long-term objective (GAO 2021a; Weill 2022). Using the information from these new maps and from independent catastrophe models, the NFIP is currently implementing a new series of premium changes, called Risk Rating 2.0, with the goal of bringing premiums into greater alignment with actuarially fair levels. This new rate setting methodology is the first major change to NFIP rate schedules since the 1970s. The new methodology will incorporate structure-specific characteristics and risk from heavy rainfall, which historically have been excluded from the calculation of premiums. Using more granular flood risk models will result in the removal of implicit subsidies to risky homes, while explicit subsidies to older construction will continue to be phased out following the guidelines of the Homeowners Flood Insurance Affordability Act; about 75% of policy holders are expected to pay higher premiums as a result (Horn 2022; Kousky and Mulder 2023).

In addition, there is early evidence that the government may be increasing enforcement of the mandatory flood insurance purchase requirement for homeowners, which will increase the salience of the social costs of living in a high-risk flood zone (GAO 2021b). As these costs rise, homeowners may find lower-risk properties increasingly attractive. "Managed retreat" policies explicitly compensate homeowners in exchange for relocating further out of harm's way. FEMA also runs a voluntary residential buyouts program that purchases flood-prone properties, allowing their residents to move to areas with lower flood risk and the property to be converted into open space. This federal buyouts initiative is the largest managed retreat-type program in the country and yet has purchased fewer than 50,000 properties since its creation (Shi et al. 2022). Overall, the implementation of such policies to encourage migration away from risky areas in the US has been limited.

Addressing these problems is worthwhile because publicly provided natural disaster insurance may also distort signals about risk in other markets. The government's historical subsidization of flood insurance, below expected payouts, and coarse flood maps signal to homeowners that their risk is lower than reality. Homeowners may then underestimate their risk levels and, by extension, overvalue property located in risky areas (McGuire 2018; Mulder 2022). We discuss how natural disaster insurance markets interact with complementary markets, such as real estate, in greater detail in Sect. 4.

Despite these challenges, there is some discussion now of whether the benefits of the public model outweigh the costs, and whether private natural disaster insurance markets should in fact become public as a result. We turn now to discuss private natural disaster insurance markets.

2.2 Private Insurance Markets

While the public provision of flood insurance through the NFIP is the primary channel through which homeowners can insure themselves against flood risk, insurance markets for other types of natural disasters remain predominantly private. Insurance contracts that provide coverage against other natural disaster types are sold by many of the same private insurers who sell standard homeowners' multiperil policies. These private markets are characterized by additional challenges to adapting to changing distributions of natural disaster risk. We discuss the differences between public and private markets, and the challenges that these differences create, in this section.

The structure of private natural disaster insurance markets differs along several dimensions from that of public markets. First, the contract structure itself is different: Unlike the contracts sold by the NFIP, coverage for other natural disasters is not typically provided through a separate policy. Instead, basic coverage is either included in standard homeowners' policies, as is the case for wildfire

coverage in California or windstorm coverage in Florida, or is sold as a supplementary policy to extend basic coverage limits, for extreme events such as earthquakes.[8] Second, many private natural disaster insurance markets service homeowners in individual states, and so their markets are less geographically diverse than the NFIP (USGAO 2007). Third, while public markets may borrow from taxpayers to elastically supply insurance contracts, private insurers do not necessarily have the same borrowing potential. Indeed, private insurers have historically had the option to refuse to renew contracts or to pull out of certain geographies altogether, in contrast with the national market of the public NFIP (Cignarale et al. 2018). Finally, the private natural disaster insurance market structure typically comprises several competitive insurers regulated by the state, rather than a single public provider (Oh et al. 2021). While public insurers have historically offered statutory subsidies to encourage take-up, private insurers have a profit-maximizing objective that aims to set premiums high enough to finance payouts while remaining competitive with other insurers and complying with state regulations. As a result, premiums may be closer to actuarially fair levels in these markets.

The market for wildfire insurance in California is one salient example of a private natural disaster insurance market in the US with these characteristics. Damages resulting from fire have traditionally been covered by standard home insurance policies across the country, and the same is true for wildfire-related property damage in the California. Unlike the federal market for flood insurance, there are over 220 distinct insurers selling wildfire coverage in California. The premiums charged by these insurers are regulated by the state, with increases in excess of 7% per year subject to extensive regulatory review (APCIA 2020; Chan 2022). Historically, private insurers had fewer difficulties aggregating wildfire damages with damages from other insured perils, but the rate of wildfire occurrence and severity has been steadily increasing since 1980. Average annual wildfire losses in California were $30 million between 1979 to 1988, and forty years later, this number has increased to almost $1 billion annually (Buechi et al. 2021). In September 2015 alone, the Valley and Butte wildfires caused over $1 billion in insured damages and destroyed over 1,700 homes, and then again in October 2017, wildfires damaged or destroyed over 14,700 insured homes (Cignarale et al. 2018). These extreme claim levels precipitated the withdrawal of a number of private insurers from high-risk areas and the increase of some premiums to unaffordable levels. Difficulty in finding insurance plans forced many homeowners in high-risk areas to purchase more expensive policies through the California FAIR plan (Liao et al. 2022).

Wildfires are not the only natural disaster that continues to be covered by standard homeowners' insurance contracts. Damage caused by wind is also covered by most multiperil policies, for example. In coastal areas that are prone to hurricanes, the distinction between flood and wind damage can be important. Public flood insurance only covers damage that is the direct result of rising water levels, which excludes damage caused by rain or roof leaks (FEMA 2023). Hurricanes typically cause both wind and water damage, so NFIP coverage tends to falls short of the full cost of home repairs in these instances, with the balance borne by private home insurers. In states such as Florida, where a large share of the market is exposed to hurricanes, damages from wind can be very costly, placing strain on private insurers and even creating difficulty for homeowners to find affordable policies. As a result, Florida and other coastal states have developed insurance markets of last resort that have similar objectives to California's FAIR market and that sell wind insurance policies to homeowners who cannot find coverage in private markets (Zhang et al. 2022).

[8] These supplementary policies also exist in markets that are dominated by a public insurer. For example, high-value homes that require coverage exceeding the NFIP maximums, which are currently set at $250,000 for building property and $100,000 for personal property, can supplement this coverage with additional insurance from a private provider (Horn and Webel 2018). Such supplementary policies are the norm for earthquake coverage in most states as well, though California exceptionally has a public earthquake insurance market (Marshall 2017).

The differences between public and private natural disaster insurance markets make it even more challenging for these private insurers to adapt to increasing climate change risk, as compared with public insurers. Private insurers have had similar difficulties raising premiums at the same rate as natural disaster costs, but it is unclear whether private markets will be able to continue to offer affordable premiums in the future. The additional difficulties stem primarily from the more limited geographic diversification, greater insolvency potential, and more stringent regulation of premiums that characterize the California wildfire insurance market, for example, and distinguish it from the public NFIP (Wagner 2022b).

First, private insurance markets are more often characterized by high market concentration and low geographic diversification. These two factors together imply that each individual firm may lack sufficient information on claims history to accurately assess risk for new homes, which in turn results in difficulty setting actuarially fair premiums (Cignarale et al. 2018; Horn and Webel 2018). Competition disincentivizes insurers from sharing historical claims data, which when combined with the changing distribution of natural disaster risk leads to high levels of uncertainty about how to determine actuarially fair premiums. While insurers may concentrate their operations in space to obtain high-quality information about local risk, high levels of market concentration and spatial coverage reduce the ability of individual insurers to cross-subsidize payouts in high-risk areas with premiums in other geographies, as is more feasible for the federal NFIP; by contrast, private insurers have been state-regulated since 1945, at which time state regulators also began to require prior review and approval of proposed homeowners' insurance rate increases (Oh et al. 2021).[9] As a result, the catastrophic and spatially correlated nature of natural disaster risk makes it more difficult for private insurers operating in limited geographies to break even when a natural disaster occurs, so that geographic concentration and exit probability are positively correlated (Jaffee and Russell 1997).

Second, private insurers typically face much larger solvency concerns following natural disasters than public insurers. This greater risk of insolvency arises because, unlike public insurers, private insurers are not backed by state or federal funding and so cannot rely on borrowing from taxpayers to finance payouts. While the NFIP can reliably borrow from the Treasury if claims exceed premium revenue following a costly flood, private insurers must either independently finance their payouts through premium and investment revenue or mitigate their payout risk through reinsurance. The capital demands required for private markets to self-finance all payouts are likely prohibitive; while realized natural disaster damages are highly variable, the risk of a natural disaster is theoretically constant, and so private insurers would have to keep enough capital on hand each year to cover the payouts of an extreme event.[10] More concretely, for a natural disaster that has a 1% chance of occurring, private insurers would need enough capital to cover 100 times the annual expected loss (Jaffee and Russell 1997). Practically, insurers' leverage ratios (i.e., the ratio of debt to assets) are regulated, which requires them to maintain a minimum amount of capital to pay out an intermediate realization of risk (NAIC 2023). Faced with high capital requirements, insurers may exhibit "safety-first" behavior by acting to minimize insolvency potential, rather than maximizing profits (Kousky and Kunreuther 2018; Roy 1952; Stone 1973). Such a model of behavior would support recent insurer decisions to pull out of areas at high risk of wildfire in California, rather than raise premiums above affordable levels (Liao et al. 2022). The increasing development of financial markets, which we discuss in more detail below, helps to mitigate some of these solvency concerns.

[9] The National Association of Insurance Commissioners (NAIC) oversees these state regulatory bodies (Webel and Cobb 2005).

[10] This capital requirement would increase in the more realistic scenario of currently increasing, rather than constant, risk.

Third, there is greater tension in private markets between setting premiums that allow insurers to remain solvent and charging rates that homeowners can afford. Unlike the federally backed NFIP, private insurers cannot operate at the loss they would incur if they subsidized premiums; indeed, historically low flood insurance premiums kept private flood insurers out of this market for much of the NFIP's history (Gaul 2019). However, homeowners' insurance is a requirement to obtain federally back mortgages in the US, and so insurance regulators are increasingly intervening in private markets to keep premiums affordable and homeowners insured (Kousky and Kunreuther 2018). These policies can take several forms. For example, one policy common in many states is imposing effective limits on annual homeowners insurance premium increases. Such a constraint has existed since 1988 in California, where wildfire insurance premium increases in excess of 7% per year trigger administrative hearings that can delay the implementation of new premiums by months or years (CLI 1993). In response, there is significant bunching of rate increase requests at 6.9% per year (APCIA 2020). Other types of regulation include limits on the variables or models that insurers can use to price risk, and therefore the extent to which premiums can reach actuarially fair levels. For example, some states prohibit insurers from pricing risk that is correlated with race, religion, income, or other demographic characteristics that could result in regressive premium schedules (Powell 2020). California also limits which wildfire risk model insurers are permitted to use to set premiums; regulators' concerns here are that select projections of future risk could result in unaffordable rates (APCIA 2020).

The effects of these additional complexities are already evident in private natural disaster insurance markets in the US. In California's wildfire insurance market, limited geographic diversification, insolvency issues, and government regulation are all making wildfire risk increasingly difficult for for-profit companies to insure. California's strict regulatory environment and constraints on premium increases limit insurers' ability to cross-subsidize wildfire risk with premiums from other states where some insurers might operate. Within California, insurers with significant market share in high-risk areas are increasingly reluctant to continue to sell these policies; between 2016 and 2019, when California insurers experienced unprecedented wildfire claims, the number of high-risk policies dropped by insurers rose by more than 100% (Liao et al. 2022; Oh et al. 2021). At the start of this time period, complaints about premium increases and nonrenewals were already 200% higher than the previous decade. This market upheaval prompted the state government to introduce a moratorium on insurer nonrenewals in 2020 (Cignarale et al. 2018). The difficulty finding insurance in the private market is evident from the rising market share of insurers of last resort. The FAIR plan in California, for example, almost doubled its market share in high-risk zip codes between 2018 and 2020, while Florida Citizens Property Insurance Corporation, the analogous insurer of last resort in Florida, experienced a similarly sized increase in market share over the same time period (Liao et al. 2022).

There are a number of potential solutions to the series of challenges outlined above, many of which would also benefit public natural disaster insurance markets. Indeed, the extreme nature of natural disaster risk suggests that some type of public-private partnership may be helpful for natural disaster insurance markets to function well. The public sector may have advantages over the private sector for implementing adaptation policies that reduce costs to insurers, transferring risk away from insurers, and communicating risk to homeowners. For example, expanded state funding for adaptation, such as sea walls or rubber coating for power lines in forested areas, could reduce overall natural disaster risk and mitigate property damage in areas covered by both public and private markets (Kousky and Kunreuther 2018). Encouraging the adoption of reinsurance policies and other financial products would also allow private insurers with geographically concentrated markets to better diversify their natural risk. Allowing private insurers to pass reinsurance expenses at least

partially through to consumers in jurisdictions that currently prohibit it (e.g., California) may be one feasible way to encourage uptake of these policies. Utilizing these financial markets to reduce local exposure to natural disasters would reduce the need to explicitly cross-subsidize risky areas with lower-risk policies in other, less stringently regulated states (Oh et al. 2021). In addition, the public sector may have advantages communicating information about risk to homeowners. Public markets such as the NFIP and the CEA make flood and earthquake risk maps publicly available to homeowners, and publicizing analogous risk information for natural disasters more commonly covered by private markets would provide informative signals to help homeowners and developers make informed decisions about mitigation and adaptation to natural disasters (Kousky and Kunreuther 2018). Information also plays an important role in stimulating demand for natural disaster insurance, which is the focus of the following section.

3 Demand for Natural Disaster Insurance

This section reviews the demand side of natural disaster insurance markets. We begin with a discussion of take-up in public and private natural disaster insurance markets and then turn to determinants of these take-up rates and possible policy solutions to increase them.

3.1 Demand Stylized Facts in Public and Private Markets

A key difference between demand for natural disaster insurance in public and private markets is the extent to which natural disaster risk is bundled with other types of insurable risk. Private natural disaster insurance markets typically bundle some natural disaster risks with other perils, with separate policies available for coverage amounts in excess of the basic limits included in standard homeowners policies. This structure is characteristic of the Californian wildfire insurance market, for example: While most wildfire damages are insured by standard homeowners' policies, extended coverage policies can be purchased separately to provide full insurance in the case of complete destruction of the house by fire (Klein 2018). Other natural disasters are explicitly excluded from homeowners' multiperil policies, including the risks of earthquakes and floods discussed above. Insurance contracts for these natural disasters must be purchased separately from the CEA or NFIP, respectively, for example, and these contracts will include coverage only for the natural disaster insured by that particular market (Kousky and Kunreuther 2018).

Demand for natural disaster insurance seems to reflect these differences in contract structures. When natural disaster risk is bundled with other perils, homeowners are more likely to be insured against those risks. Over 95% of homeowners in the US carry a general insurance policy against standard perils such as theft, vandalism, and other common risks to property (Klein 2018). By contrast, take-up of policies for specific natural disasters that are excluded from standard home insurance policies is much lower. For example, only approximately 10% of homeowners in California purchase earthquake insurance (Marshall 2017). Take-up of flood insurance is also so low that the literature routinely refers to the "insurance gap" in this market, even though Fig. 1 suggests that flood insurance demand is much higher than at most times in the fifty-year history of the NFIP (Kousky and Kunreuther 2018). Overall, only approximately 4% of homeowners in the US purchase a flood insurance contract, with much of this take-up driven by demand in high-risk flood zones (Bradt et al. 2021). However, even in these high-risk flood zones, over 40% of homeowners are uninsured (Wagner 2022a). The supplementary coverage policies available for bundled natural disaster risks, such as wildfires, are also less likely to be purchased than the basic multiperil policies (Klein 2018).

This puzzlingly low uptake for natural disaster-specific policies suggests that homeowners' behavior in these markets differs from other perils, which underscores the importance of research on these markets.

These low extensive-margin take-up rates for natural disaster insurance are especially remarkable given the large expected benefits of insurance for homeowners. Standard theory predicts that risk-averse homeowners should be willing to pay their expected insurance benefit plus a risk premium that reflects their benefits from the reduction in the variance of consumption provided by insurance (Hendren 2020). Though there are few estimates of risk aversion in natural disaster insurance markets, homeowners are generally thought to be risk averse, as in most other domains (Sydnor 2010).[11] Willingness to pay for insurance rises as the variance of realized risk rises, and the infrequent, catastrophic nature of natural disasters creates a large difference between the financial consequences of a natural disaster if a homeowner is insured relative to when they are uninsured. It is therefore surprising that many homeowners are uninsured against natural disaster risk when contracts are actuarially fair, as in the case of the CEA, or even better than actuarially fair, as for subsidized policies sold through the NFIP; risk-averse homeowners should be willing to pay more than an actuarially fair premium, and yet we find that a majority of homeowners throughout the country are unwilling to do so in markets for natural disaster insurance specifically. The gradual phase-out of flood insurance subsidies after 2012, for example, led to reductions in the share of insured homeowners well before premiums reached actuarially fair levels, and the extent to which current premium changes under Risk Rating 2.0 will affect coverage levels remains to be seen. Overall, extensive-margin price elasticities of demand for flood insurance are estimated to be around -0.30 (i.e., relatively inelastic) (Bradt et al. 2021; Wagner 2022a).[12]

In addition to the decision of whether to purchase a natural disaster insurance contract, unbundled contracts for natural disaster risk also typically allow an intensive-margin purchase choice: Conditional on purchase, homeowners can choose different levels of coverage for buildings and their contents. Though there has been more limited analysis of the decision of how much natural disaster coverage to purchase, existing work suggests that homeowners are not price elastic on this margin; intensive-margin demand elasticities for flood insurance coverage are in the range of -0.01, with the small demand response arising through adjustments in contents coverage purchased rather than coverage for the structure itself (Wagner 2022a). Conditional on purchasing a flood insurance contract, for example, homeowners appear to fully insure their homes, rather than making a piecemeal decision to insure a fraction of the house's value. Indeed, total purchased coverage exceeds damages for almost all claims submitted to the NFIP and in some instances also exceeds the value of the house (Collier and Ragin 2020). This stylized fact justifies the treatment of natural disaster insurance contracts as full insurance contracts in empirical work even though the homeowner could in theory decide sequentially whether to purchase insurance and then how much to buy. The relatively few purchasors of natural disaster insurance therefore seem to fully insure themselves against these risks. Next, we discuss candidate explanations for why demand for natural disaster insurance is so puzzlingly low.

[11] In fact, the limited work estimating risk aversion parameters for homeowners' insurance suggests that risk aversion is higher than in more commonly studied insurance settings such as health or unemployment, and almost implausibly so, with coefficients of absolute risk aversion in the range of 1.7×10^{-3} instead of around 5×10^{-4} (Hendren 2020; Sydnor 2010).

[12] A more extended set of papers based on panel regressions without quasi-experimental price variation or on simulation models estimate demand elasticities in the range of -0.49 to -0.06 (Browne and Hoyt 2000; NRC 2015; Kriesel and Landry 2004).

3.2 Challenges and Policy Solutions

Homeowners' low levels of demand for natural disaster insurance are surprising because premiums underestimate the expected benefits of insurance in many of these markets. Natural disaster insurance premiums have increased much more slowly than natural disaster risk in the last fifty years, so that many policies are implicitly subsidized relatively to current actuarially fair levels; the current flood insurance subsidy is about 30% of payouts on average, for example (Wagner 2022a). Many of these uninsured homeowners are in fact at risk of natural disasters. After Hurricane Harvey in 2017, for example, only approximately 20% of homeowners who experienced flooding were insured (Healy 2018). This same pattern is evident in other markets after natural disasters: Two-third of homeowners affected by the Tubb Fire in California in that same year did not have the supplementary wildfire coverage needed to rebuild their homes (Frazee 2017).

There are many candidate explanations for low take-up of natural disaster insurance, some of which are economic and some of which are behavioral. Economic explanations for low observed demand allow that homeowners are rational, but that unobserved additional costs or benefits influence homeowners' decisions, leading to lower uptake than we might expect based on average payouts alone. Some examples in this setting that could decrease willingness to pay include private information about natural disaster risk (i.e., adverse selection), implicit insurance through government bailouts, limited liability due to low home equity, and credit constraints. Adverse selection is the textbook economic explanation for inefficiently low demand for insurance; it arises when the lowest cost homeowners also have the lowest willingness to pay, so that average insurance costs increase when prices rise above the willingness to pay of the lowest cost individuals. The distinction between marginal and average costs in selection markets means that some homeowners choose not to purchase insurance because their private, marginal insurance costs are lower than the premiums set at average cost that they are charged. Prior work shows that adverse selection is an important driver of insurance purchase decisions in insurance markets with less infrequent, spatially correlated, and extreme risk realizations, such as health, unemployment, and life insurance; Einav et al. (2021) review the implications of selection in these other markets in detail. Finkelstein et al. (2019) also provide evidence that public bailouts—a second economic explanation—contribute to low health insurance demand in some markets, where uninsured individuals do not in fact pay for their uninsured healthcare expenditures. In addition, there are several economic costs that could be excluded from measured average payouts, including search costs of acquiring information about true risk, hassle costs of obtaining risk assessments necessary to calculate natural disaster premiums, and information processing costs of understanding contract characteristics. Many of these costs exist in other insurance markets. For example, high search costs needed to learn about health insurance plan characteristics seem to discourage individuals from optimally choosing a contract (Abaluck and Gruber 2011; Handel 2013; Handel and Kolstad 2015).

Behavioral explanations for low take-up include misperception of natural disaster risk, mistaken expectations about the likelihood of public bailouts, mistaken beliefs about included perils in homeowners insurance, cognitive dissonance in relation to the probability of climate change occurring, and inertia in purchase decisions, among others. Existing work shows that many of these frictions distort demand in other insurance markets. For example, consumers of health insurance are in some instances misinformed about both the benefits of health insurance and the treatments that it covers. Pauly and Blavin (2008) propose that some individuals may not understand the value provided by different plan characteristics and make suboptimal insurance purchase decisions as a result of these mistakes, while Zhou-Richter et al. (2010) show a similar pattern for long-term care insurance purchases. Bhargava et al. (2017) show that individuals often choose health insurance plans that are financially dominated, plausibly because they lack the competence required to translate premiums into relevant out-of-pocket

costs. Papers that estimate search costs in health insurance markets also point to similar behavioral explanations, such as inertia and mistaken beliefs about the ways in which health insurance plans differ, as contributing factors to the very high switching costs that these models find (e.g., Abaluck and Gruber 2011; Handel 2013; Handel and Kolstad 2015). Handel and Schwartzstein (2018) review additional examples of behavioral mistakes in health insurance and other markets.

The distinction between these competing explanations is important because the existence of additional unmeasured economic costs that are excluded from average insurance payouts would mean that homeowners are making rational decisions when forgoing the purchase of natural disaster insurance; accounting for these additional costs borne by homeowners would explain why demand is puzzlingly low in these markets. By contrast, if individuals are making behavioral mistakes, then there is a role for policy to remove these frictions to allow homeowners to make more informed choices (Handel and Schwartzstein 2018).

Existing work supports that both economic costs and behavioral frictions may contribute to low take-up for natural disaster insurance, though in ways that differ from other insurance types such as health and unemployment. Economic explanations for low take-up in other insurance settings appear to play a smaller role in natural disaster insurance markets and alone seem unable to explain the extent of uninsurance in these markets. For example, while adverse selection is evident in many other insurance markets, private information in natural disaster insurance markets appears to be limited (Wagner 2022a). Homeowners do not seem to have more information about their private natural disaster risk than insurers do, plausibly because most determinants of natural disaster risk are based on features of local geography and therefore in theory could be perfectly observed by both sides of the market.[13] In addition, the public bailouts that depress demand in some health insurance markets are much less important in natural disaster insurance markets. As mentioned above, Finkelstein et al. (2019) find that public payment of uninsured low-income individuals' medical bills can fully explain low uptake of insurance in low-income health insurance exchanges, but the analogous federal grants post-disaster are only about 0.03% of average flood insurance payouts during hurricanes; state grants are larger, but homeowners may obtain these funds regardless of whether or not they have purchased insurance, so we would not expect these grants to distort willingness to pay for insurance (Horn 2018). Search costs or hassle costs also seem unable to fully rationalize low willingness to pay because many homeowners purchase insurance for only one year (i.e., in their first year of owning their home) and then fail to renew it. It then seems implausible that such unmeasured costs would explain low demand because these costs are highest in that first year of policy tenure, when regulators undertake an assessment of the structure to determine the premium that applies (NRC 2015).

Credit constraints and limited liability may both play larger roles in natural disaster insurance markets than in other insurance markets, though both seem unlikely to be able to explain the extent of uninsurance. Low-income homeowners may indeed forgo flood insurance if they cannot afford it; however, the average flood insurance premium is about 1% of median income in high-risk, high-amenity flood zones, which suggests the cost of flood insurance is not prohibitive for most homeowners (CBO 2017). Limited liability may be comparatively more important in this setting because the insurable risk is to houses rather than people, though bankruptcy occasionally acts as implicit insurance for uninsured individuals in health insurance markets (Liao and Mulder 2021; Mahoney 2015). The intuition here is that homeowners with mortgages may in fact have relatively little capital at risk, and uninsured homeowners with little home equity could choose to default on their mortgages and avoid paying reconstruction costs in the event of a natural disaster. However, given that most homeowners have equity in their home at least equal to the average flood insurance

[13] Indeed, the development of more sophisticated risk models suggests if anything that we might expect insurers to have an information advantage relative to consumers.

payout due to their down payment, it seems somewhat unlikely that limited liability is the primary explanation for low take-up (Li and Goodman 2016).

Behavioral explanations, such as misperception of risk and incomplete understanding of policy structure, also appear to play some role in natural disaster insurance markets. For example, a growing body of literature points to underestimation of natural disaster risk as a plausibly important explanation for low willingness to pay. This evidence is based both on direct surveys of homeowners' beliefs and on indirect hedonics analysis of house price responses to new information about natural disaster risk. Surveys of flood risk beliefs typically elicit risk probabilities that are lower than both the NFIP's conservative inundation model and the more sophisticated models of private companies (Bakkensen and Barrage 2021; Royal and Walls 2019). Hedonic regressions of house prices on natural disaster risk measures also suggest that realized floods or fires or updated map information about areas at high risk of these disasters provide novel information that homeowners use to update their priors on their risk (Gibson and Mullins 2020; Hino and Burke 2021; McCoy and Walsh 2018). Gallagher (2014) also shows that homeowners respond to the direct experience of a flood by purchasing insurance the day after one occurs. Similar to the spike in insurance demand after natural disasters, there is also typically a spike in litigation by homeowners who feel duped by flood or fire insurers about the terms of their contract and its benefits (Klein 2018). This response suggests that homeowners' incomplete understanding of natural disaster coverage and exclusions from standard homeowners' insurance policies may bias demand downward. In addition, many of the other behavioral explanations for low take-up that exist in other insurance markets (e.g., inertia) have yet to be tested in these markets, and we view this line of inquiry as a fruitful direction for future work.

The nature of risk in natural disaster insurance markets suggests that behavioral frictions may potentially be more important in these markets than in more commonly studied insurance markets, such as health insurance. The large, infrequent, and spatially correlated natural disasters that complicate the calculation of actuarially fair premiums on the supply side also create difficulties for homeowners to assess their risk. The high costs and low probabilities involved may be more difficult for individuals to conceptualize than the comparatively larger probabilities of making a health insurance or automobile insurance claim, especially if a homeowner has never experienced a natural disaster directly. There is some evidence that individuals may round very small probabilities to zero, even if the highly unlikely event would result in a large negative income shock (Sunstein 2002). Hence, if the probability of experiencing a natural disaster is perceived to be below some critical threshold, homeowners' willingness to pay could be quite low. The difficulty of collecting information on the probability of a natural disaster that is infrequently observed may also make it easier for homeowners to underestimate the probability of it ever occurring. Consistent with this interpretation, Gallagher (2014) shows that a Bayesian learning model, in which experiencing a flood provides new information about true flood risk and encourages homeowners to update their priors on flood risk probabilities, can explain observed insurance demand increases immediately after an event occurs.

The presence of behavioral mistakes is problematic because it means that homeowners' full benefit from insurance is not reflected in their willingness to pay. Many of these frictions seem likely to bias willingness to pay downward: Homeowners will undervalue natural disaster insurance if they underestimate their probability of experiencing a flood, a fire, or an earthquake, for example. Hence, at any given level of premiums, many homeowners may be uninsured who would actually benefit from insurance in expectation, and raising premiums can exacerbate the extent of inefficient underinsurance. Using this revealed preference willingness to pay to evaluate the effects of proposed policies may lead to biased estimates of changes in social welfare. If these frictions are important in natural disaster insurance markets, then the ongoing phase-out of NFIP subsidies and wildfire

insurers' regulatory filings requesting increases in premiums for wildfire coverage seems likely to lead to even lower levels of demand for insurance—even if premiums remain actuarially fair or better.

To encourage homeowners to purchase insurance, regulators could employ either allocative or mechanism policies (Handel and Schwartzstein 2018). An allocation policy assigns consumers to specific options; mandating the purchase of natural disaster insurance would be one example. By constrast, mechanism policies typically target specific distortions, by encouraging uptake through the provision of accurate information to correctly align homeowners' beliefs with reality, for instance. The objective of both types of policies is to attain the efficient level of insurance coverage, where all homeowners who would benefit from insurance are covered. Mechanism-type policies typically require more information about the source of any distortion since these policies would target the removal of one or another specifically; allocation policies more bluntly require (or strongly suggest) that homeowners purchase a certain level of coverage, and so identifying the underlying source of the distortion is less important.[14]

Both allocation and mechanism policies are beginning to be more broadly employed to address the "coverage gap" in natural disaster insurance markets. The existing mandate for homeowners with federally backed mortgages to purchase flood insurance is one example of an allocation policy. Mortgage lenders appear to be beginning to enforce this mandate more stringently than at any time since its inception in 1973 (Michel-Kerjan 2010). New information-based mechanism policies are also being rolled out in many states. For example, the NFIP flood map updates that more accurately represent homeowners' risk and the housing disclosure laws mandating the publishing of flood zone status for new transactions are two current policies that emphasize providing accurate information about flood risk to homeowners. The implementation of these policies changes house prices, suggesting that homeowners do in fact update their beliefs about their flood risk based on this new information (Hino and Burke 2021; Lee 2022). Another solution that is currently discussed is the auto-renewal of flood insurance policies, to avoid the decline in demand after the year that a house is purchased (Kousky et al. 2019). In wildfire insurance markets, basic coverage is already mandatory under the current market structure because most homeowners are required to carry standard multiperil policies to obtain a mortgage. In addition, ongoing improvements in climate risk models will also allow insurers to obtain more accurate information about homeowners' risk of natural disasters, and efforts to digitize many of these maps could improve homeowners' access to the information they need to make informed insurance purchase decisions (Weill 2022).

4 Complementary Markets

Reforms to natural disaster insurance markets also have spillover effects in other complementary markets. In this section, we discuss how natural disaster insurance markets interact with housing markets, financial markets, construction decisions, and other public programs.

4.1 Real Estate Markets

The geographic nature of natural disaster risk creates a tight link between housing markets and natural disaster insurance markets. Natural disaster insurance contracts are house-specific, rather than individual-specific in the manner of health insurance, life insurance, automobile insurance, or most other insurance contract types. The nature of this risk means that natural disaster insurance

[14] Handel and Schwartzstein (2018) review a wide range of policies employed in health insurance and other markets.

premiums are one feature of houses that homeowners could conceivably consider when choosing between houses to purchase. If these insurance premiums rise, potential homeowners could find risky homes less appealing, and so increases in location-specific premiums could drive decreases in property values. As a result, local governments have an unusually high level of interest in natural disaster insurance market regulation because of the potential implications for their property tax base. Current estimates of homes exposed to serious flooding suggest that the number of Americans living in high-risk flood zones is almost 41 million—approximately 3 times as high as estimates based on NFIP flood maps—and so the potential house price effect of making flood insurance actuarially fair is a significant concern for many local governments throughout the country (Wing et al. 2018).

There is an extended literature showing that risk signals provided either through insurance premiums or through information policies do in fact affect house prices. The gradual phase-out of flood insurance premium subsidies, for example, seems to have contributed to house price declines in risky areas that began to face higher, more accurate risk-based premiums; ongoing premium increases seem likely to contribute to additional devaluation (Gibson and Mullins 2020). Updating natural disaster risk maps also reduces prices for houses newly mapped into high-risk zones, for floods (Hino and Burke 2021), fires (Garnache 2020), and earthquakes (Singh 2019); these effects also seem likely to persist as maps reflecting new climate change risk distributions are disseminated. Providing risk information through other means, such as a separate (dis)amenity listing on house sales profiles, has a similar effect as map updates (Lee 2022). Other work finds that natural disaster events themselves reduce house prices for undamaged properties by up to 20% (Hallstrom and Smith 2005; Kousky 2010).

The price response to new information about risk provides further evidence that homeowners have an incomplete understanding of the extent to which they are exposed to natural disasters. If homeowners' priors accurately reflected their risk, then natural disaster map updates conceivably would not contain any new information, and we would not expect a house price response. By contrast, Baldauf et al. (2020) show that increases in future inundation risk are differentially capitalized into house prices depending on whether homeowners report that they believe climate change is occurring. In addition, Bakkensen and Barrage (2021) provide direct survey evidence that homeowners who live in risky flood zones are more optimistic about flood risk than homeowners who live in low-risk flood zones—a pattern that is opposite to their true risk. One implication of these biased beliefs in their model is that natural disaster risks, such as coastal sea-level rise, are incompletely capitalized into property markets, so that prices of at-risk houses are higher than they would be with accurate beliefs. Their estimates of the wedge created by unpriced climate risk suggest that houses in high-risk flood zones are over-valued by at least 13%. Mispricing of natural disaster risk in property markets is not a new phenomenon; these markets have historically undercapitalized natural disaster risk since informative signals from changes in natural disaster insurance premiums and updates to maps were limited throughout much of the history of these markets; recent estimates suggest that homes in high-risk flood zones are over-valued by approximately $200 billion in aggregate (Gourevitch et al. 2023). However, the wedge between true value and market value of at-risk homes could plausibly increase as exposure to climate change risk increases.

Providing accurate signals about risk through natural disaster insurance premiums is one way to help correct the over-valuation of at-risk property. Charging homeowners actuarially fair insurance premiums forces them to internalize true costs of living in risky areas; if these costs are high, then it may be efficient to incentivize homeowners to move out of harm's way. Panjwani (2022) estimates that both excluding flood risk premiums from mortgages and providing homeowners with funds to rebuild lead to inefficiently high levels of construction and density in high-risk flood zones; this chapter estimates that a tax on homeowners equal to the median flood insurance premium would have reduced the number of New Jersey homeowners exposed to Hurricane Sandy in 2012 by 17%. Baylis

and Boomhower (2023) show a similar phenomenon for public funding intended to prevent wildfires from destroying homes (i.e., before insurance claims are made); such funding provides an implicit subsidy exceeding 20% of the value of the homes in the areas at highest risk of wildfires in the Western US. However, despite the efficiency gains in property markets by providing accurate insurance price signals about true risk, increasing natural disaster insurance prices to be actuarially fair continues to be strongly opposed by both homeowners and local governments. In addition, in the presence of behavioral mistakes, implementing actuarially fair insurance prices can have unintended welfare costs, such as discouraging homeowners who underestimate their risk from purchasing insurance at higher prices (Wagner 2022a). Correcting distortions in both insurance and real estate markets will require separate policies designed to address the distinct challenges in each of these related markets.

4.2 Financial Markets

The interactions between natural disaster insurance markets and financial markets change the incidence of natural disaster risk. The linkages between natural disaster insurance markets and financial markets complicate the functioning of these other markets, but also create possible solutions for some of the supply-side challenges we identify in Sect. 2.

The substantial uncertainty surrounding the costs of financing homeowners who might be exposed to natural disaster risk creates challenges in financial markets. The difficulties accurately pricing climate risk that complicate setting actuarially fair insurance premiums and allowing homeowners to learn about their risk also create pricing challenges for mortgages, for example. It is difficult to predict the effects of climate change over the lifespan of a conventional thirty-year mortgage in the US, and this uncertainty is evident in banks' pricing decisions. Residential mortgage lenders do not price sea-level rise exposure in a systematic way: They charge higher interest rates for thirty-year mortgages in areas that recently experienced a hurricane, but this premium declines over time and is attenuated in areas where fewer residents believe that climate change is occurring (Hurst et al. 2016; Nguyen et al. 2022). Consistent with other markets, financial institutions also have relatively limited ability to accurately predict how climate change will affect their products.

A reaction of financial markets to this uncertainty surrounding costs has been to transfer risk to other agents in the market. This pattern is especially evident for uninsured homeowners, who have a greater risk of default or delinquency than their insured counterparts. Mortgage lenders, for example, transfer risk both back to uninsured homeowners and to securitizers of mortgages. Sastry (2022) shows that homeowners without flood insurance are required to make higher down payments than their insured counterparts; this higher equity at risk (i.e., lower loan-to-value ratio) reduces exposure of banks at the expense of homeowners with more natural disaster exposure. There is also evidence supporting a similar pattern of risk transfer upstream: Lenders appear to offload mortgage debt written in high-risk flood zones to government-sponsored enterprises (e.g., Fannie Mae and Freddie Mac) who do not price flood risk in their guarantee fees, though the results are currently debated (LaCour-Little et al. 2022; Ouazad and Kahn 2022). Similar to the distribution of mortgage down payments, this risk transfer is smaller in areas where flood insurance is required. Higher flood insurance penetration rates would therefore alleviate some of the risk to homeowners and financial institutions by sharing risk with insurers.

Despite this pricing uncertainty, financial markets offer potential remedies to challenges in natural disaster insurance markets by reducing the incidence of natural disaster risk on insurers. Markets for reinsurance contracts and newer financial products, such as catastrophe bonds, have developed as possible solutions to insurers' insolvency risk resulting from highly spatially correlated, costly natural disasters. Reinsurance markets allow natural disaster insurance markets to make payments to

designated reinsurance companies, such as Munich Re or Swiss Re, in exchange for indemnifying specific natural disaster losses; global reinsurers can therefore pool risk for many disasters across larger geographies than any federal or state-level natural disaster insurer (Kousky and Kunreuther 2018). By spreading the risk of catastrophic losses amongst many insurers, these secondary financial markets increase the ability of natural disaster insurers to write policies without maintaining high capital reserves to avoid going bankrupt; insurers need only pay claims that are not covered by reinsurance out of their reserves. In response to the increases in natural disaster insurance costs, reinsurance premiums have been rising steadily over the last two decades. Despite these premium increases, global reinsurance capital grew by approximately 30% in the last decade and exceeds over $630 billion today, and it is widely believed that these markets have bolstered the ability of natural disaster insurers to continue to write policies in the face of escalating natural disaster costs in the US and abroad (Gallagher-Re 2023; OECD 2018).

Catastrophe bonds are a newer alternative financial instrument used to reduce exposure to natural disasters. These bonds pay investors (e.g., financial institutions) interest up to the point where a prespecified disaster occurs, at which point at least some of the principle is transferred to the victims of that disaster. Such alternative financial instruments for transferring risk first became prevalent in the 1990s when the traditional reinsurance markets faced challenges in smoothing risk the same way some private natural disaster insurance markets did; the first catastrophe bonds were issued after Hurricane Andrew in Florida in 1992 bankrupted a dozen insurers due to insufficient supply of reinsurance. Catastrophe bond markets have steadily increased in importance and market capitalization since then, with markets for these types of bonds and other insurance-linked securities exceeding $100 billion today. These types of investments are increasingly popular with investors due to their relatively stable, high-single-digits returns and low correlation with shocks to other types of assets (Braun and Kousky 2021; Ruoff 2023). Other innovative financial products, such as parametric insurance contracts, which pay out differentially depending on the magnitude of individual disasters, are also becoming increasingly available around the world. These markets seem likely to become even more important as climate change is predicted to drive demand for creative ways to reduce exposure to rising natural disaster risk (Michel-Kerjan 2010).

4.3 Construction

Markets for in-place adaptation, such as resilient construction materials, interact with natural disaster insurance markets in two main ways. First, adaptation reduces natural disaster damages and therefore lowers actuarially fair insurance premiums for houses that are better protected. Second, natural disaster insurance coverage discourages investment in adaptation capital.

Adaptation reduces pressure on natural disaster insurance markets by lowering expected payouts for more resilient houses. Compliance with stringent construction codes in high-risk flood or wildfire zones reduces insurance payouts and increases the probability of house survival. For example, elevating houses at high risk of flooding reduces insurance payouts by 30%, relative to those that are not elevated, with the highest benefits in the most catastrophic floods (Hovekamp and Wagner 2023). Similarly, mandated minimum construction standards in wilderness areas in California increase the probability that houses will survive wildfires by 40% and also have positive spillovers on neighboring houses by improving the likelihood that these houses also survive by 6% (Baylis and Boomhower 2021). In both of these markets, this expected construction benefit is partially priced into insurance rates, so that homeowners who undertake defensive investments obtain a discount on insurance. Flood insurance sold by the NFIP is cheaper for houses that have elevated foundations, for example, though this discount does not fully offset the expected benefits of adaptation (Wagner 2022a).

Despite this insurance discount, adaptation can also substitute for natural disaster insurance. Wagner (2022a) finds that houses that are required to be elevated are 25% less likely to be insured against flooding than houses that are not required to be elevated. More generally, Fried (2022) finds that the provision of post-disaster payouts reduces adaptation capital by 5% in a model that includes multiple types of natural disasters. High natural disaster insurance subsidies may be at least partially responsible for this substitution: At-risk homeowners have limited incentive to incur private costs to reduce damages that are borne by insurers. For example, several surveys of homeowners in high-risk flood zones find that only 5 to 17% of respondents reported that they had taken any measures to fortify their homes (Michel-Kerjan 2010). Risk-reducing investments are usually cost-effective when implemented during the course of construction, and so mandating more stringent adaptation at the onset may be an effective way to reduce the cost burden on insurers (Baylis and Boomhower 2021).[15] However, the benefits of more resilient construction are not capitalized into house prices, and so adaptation is likely to be underprovided by private homeowners (Hovekamp and Wagner 2023; Ostriker and Russo 2022). Many of the reasons for low-insurance take-up discussed above also help explain why homeowners may underinvest in defensive investments against natural disasters, including misperceptions of risk, cognitive dissonance, and myopia. In addition to encouraging insurance uptake, reducing this undercapitaliztion of the costs of natural disasters and the benefits of private risk-reducing investments into house prices is another reason to improve access to accurate natural disaster risk information.

4.4 Other Public Programs

Natural disaster insurance markets are one of a suite of programs that support homeowners after natural disasters. Insured homeowners receive insurance payouts after natural disasters occur, but other federal, state, and local programs also provide post-disaster bailouts, and these other public payouts may be tied to the purchase of natural disaster insurance. Collier and Kousky (2023) provide a detailed description of financing available to homeowners in the aftermath of natural disasters. The two main grants available to homeowners from the federal government are the Individuals and Households Program and the Small Business Administration Disaster Loan Program, both of which require flood insurance to be purchased as a condition of accepting public assistance after a flood. Receipts from these programs are typically in the range of a few thousand dollars. Other post-disaster grants are administered by state and local governments and would typically provide higher payouts to insured homeowners than their uninsured counterparts to encourage uptake (Horn 2018). Deryugina (2018) shows that payments from social programs that are not directly related to natural disasters, such as unemployment insurance and public medical payouts, also increase in their aftermath. While this literature does not distinguish between insured and uninsured homeowners, it seems plausible that uninsured homeowners may rely on some form of government transfer from these other programs to recover from a natural disaster.

These other social programs may substitute for formal natural disaster insurance to some extent. As discussed in Sect. 3, the receipt of public bailouts from these other programs after natural disasters may depress natural disaster insurance demand. There is some evidence that receipts of Individual Assistance grants and Small Business Administration loans lower the intensive margin of flood insurance coverage after hurricanes, with no effect on the extensive margin of take-up (Kousky et al.

[15] Ostriker and Russo (2022) and Baylis and Boomhower (2021) show respectively that lower natural disaster insurance premiums and higher natural disaster insurance bailouts encourage construction in risky areas. Accurate signals about risk may therefore also help reduce the density of property in harm's way and, by extension, insurance payouts.

2018). While the magnitudes of the transfers from these programs seem too small individually to explain the extent of uninsurance, the cumulative role of these public payouts in explaining puzzlingly low take-up is an open empirical question.

5 Conclusion

This chapter surveys evidence of current challenges in natural disaster insurance markets, why these challenges arise, and what can be done to address them. We discuss how the infrequent, spatially correlated, and extreme events that characterize natural disaster insurance markets complicate both the supply of and demand for natural disaster insurance, with spillovers to interrelated markets such as real estate. These distinguishing market features motivate a growing body of literature studying natural disaster insurance separately from other insurance types, such as health. A key theme throughout this chapter is that the behavior of insurers and homeowners in these markets may necessitate different policy solutions than other insurance markets.

Many of these current challenges discussed in this chapter stem from the historical design of natural disaster insurance markets and the policy choices made over the course of their half-century history. Natural disaster insurance markets, such as the NFIP, were not typically designed to weather the increasingly high number of costly disasters that these markets are experiencing, and were certainly not designed to account for the non-stationarity of the distribution of climate change risk and the uncertainty of how this risk translates into realized floods or fires. Initially, subsidized and heavily regulated insurance premiums were introduced to accomplish take-up objectives under very different distributions of climate risk than those prevailing today, but difficulty changing the initial market structure coupled with rising costs of natural disasters have made it more challenging for insurers to operate in today's changing climate (Gaul 2019). Indeed, this chapter suggests that natural disaster insurance policy changes have not kept pace with the changing distribution of risk over the last fifty years. While this institutional lock-in has largely prevented previous attempts at natural disaster insurance reform, such as the Biggert-Waters Flood Insurance Reform Act of 2012, from accomplishing their objectives, the current rate of increase of costs in these markets suggests that ongoing reforms will likely continue to ensure that public or private insurers can carry on offering natural disaster insurance in the US. We highlight here a few areas where we believe important work to advance this agenda remains to be done.

First, we note that the effects of policies in natural disaster insurance markets could offer insight and predictions for other insurance markets. Natural disasters damage capital beyond homes; the market for crop insurance, for example, shares some similarities with flood and fire insurance markets, as the same disasters that destroy homes can wipe out crop fields. Similar to the NFIP, the Federal Crop Insurance Program (FCIP) subsidizes crop insurance to encourage take-up, and total subsidies have grown substantially since the program's inception, reaching almost $12 billion in 2022 (Coble and Barnett 2013; Tsiboe and Turner 2023; GAO 2023). Given the rising costs of the FCIP, the federal government is exploring new policies to improve the efficiency and equitability of the program, and future research could examine how lessons learned from natural disaster insurance can inform new crop insurance policies (GAO 2023).

The current frontier in the insurance literature more generally is incorporating behavioral mistakes into insurance markets. Theoretical models of insurance markets are beginning to emphasize the potential role of frictions such as biased beliefs, myopia, and inertia on demand, and to highlight the additional empirical estimates needed to quantify welfare in insurance markets in the presence of such frictions (Handel and Schwartzstein 2018; Handel et al. 2019). Willingness to pay for natural disaster insurance may be more biased by these types of mistakes than in other insurance markets due

to the additional challenges of learning about risk in a market with infrequently observed and extreme events. The resulting distortions in demand have important implications for the supply side of natural disaster insurance markets as well: Private insurers cannot break even if homeowners are unwilling to pay even their own expected damage costs for insurance. These distortions also support implementing information-based policies that provide accurate signals to homeowners about their risk (e.g., modern risk maps), in addition to the current actuarially fair pricing reforms (Wagner 2022a). Future research quantifying the contributions of different frictions to remarkably low willingness to pay for natural disaster insurance would help inform the implementation of policies designed to address them.

An additional open question is the extent to which these proposed policies may have inequitable effects across different populations in the US. The high and increasing costs of natural disasters mean that many of the solutions that are debated focus on the first-order problem of ensuring continued access to insurance, and the incidence of premium changes on consumers and insurers. However, the effects of these policies, and indeed the effects of climate change themselves, are distributed heterogeneously across the country. Low-income and minority homeowners tend to sort into higher-risk neighborhoods because these houses are often cheaper even if their prices do not fully reflect underlying risk. Insurance premiums are also highest in these high-risk areas, and so low-income households are more exposed to higher natural disaster risk and higher insurance costs. The NFIP, for example, is typically considered a regressive program (Bin et al. 2012). Further, some evidence supports that although currently proposed premium reforms would increase overall welfare, these reforms would disproportionately burden low-income homeowners; by contrast, actuarially fair premiums in combination with federal transfers to low-income households could improve both equity and economic efficiency (Bakkensen and Ma 2020; Kousky and Kunreuther 2018). The effects of climate change seem likely to continue to increase annual natural disaster costs for the foreseeable future, and so the design of policies to minimize the burden on both homeowners and insurers will be an important area for future research.

References

Abaluck J, Gruber J (2011) Choice inconsistencies among the elderly: evidence from plan choice in the Medicare Part D Program. Am Econ Rev 101(4):1180–1210

AIR (2005) A chronology of major events affecting the national flood insurance program. Technical report, American Institute for Research

APCIA (2020) Insurance industry response to CDI investigatory hearing on homeowners' insurance availability and affordability. Technical report, American Property Casualty Insurance Association

Bakkensen L, Barrage L (2021) Flood risk belief heterogeneity and coastal home price dynamics: going under water? Rev Financ Stud 35(8):3666–3709

Bakkensen L, Ma L (2020) Sorting over flood risk and implications for policy reform. J Environ Econ Manag 104:102362

Baldauf M, Garlappi L, Yannelis C (2020) Does climate change affect real estate prices? Only if you believe in it. Rev Financ Stud 33(3):1256–1295

Baylis P, Boomhower J (2021) Mandated vs. voluntary adaptation to natural disasters: The case of U.S. wildfires. Working Paper.

Baylis P, Boomhower J (2023) The economic incidence of wildfire suppression in the United States. Am Econ J Appl Econ 15(1):442–473

Bhargava S, Loewenstein G, Sydnor J (2017) Choose to lose: health plan choices from a menu with dominated option. Q J Econ 132(3):1319–1372

Bin O, Bishop JA, Kousky C (2012) Redistributional effects of the National Flood Insurance Program. Public Finan Rev 40(3):360–380

Bradt J, Kousky C, Wing O (2021) Voluntary purchases and adverse selection in the market for flood insurance. J Environ Econ Manag 110:102515

Braun A, Kousky C (2021) Catastrophe bonds. Technical report, Wharton Risk Management and Decision Processes Center

Browne MJ, Hoyt RE (2000) The demand for flood insurance: empirical evidence. J Risk Uncertainty 20:291–306
Buechi H, Weber P, Heard S, Cameron D, Plantinga AJ (2021) Long-term trends in wildfire damages in California. Int J Wildland Fire 30(10):757–762
CBO (2017) The national flood insurance program: financial soundness and affordability. Technical report, Congressional Budget Office
Chan R (2022) 2021 California property and casualty premium and loss summary. Technical report, California Department of Insurance
Cignarale T, Laucher J, Allen K, Landsman-Smith L (2018) The availability and affordability of coverage for wildfire loss in residential property insurance in the wildland-urban interface and other high-risk areas of California: CDI summary and proposed solutions. Technical report, California Department of Insurance
Cignarale T, Laucher J, Allen K, Landsman-Smith L (2019) Availability and affordability of coverage for wildfires. Technical report, California Department of Insurance
Citizens (2020) Strategic plan. Technical report, Florida Citizens Property Insurance Corporation
CLI (1993) Insurance code. Technical report, California Legislative Information
Coble KH, Barnett BJ (2013) Why do we subsidize crop insurance? Am J Agric Econ 95:498–504
Collier B, Kousky C (2023) Economic and financial resilience of populations after climate risk events: the role of insurance. Working Paper
Collier B, Ragin M (2020) The influence of sellers on contract choice: evidence from flood insurance. J Risk Insurance 87(2):523–557
Deryugina T (2018) The fiscal cost of hurricanes: disaster aid versus social insurance. Am Econ J Econ Policy 9(3):202–233
Einav L, Finkelstein A, Mahoney N (2021) The IO of selection markets. Handbook Ind Organ 5(1):389–426
FEMA (2023) Wind damage vs. water damage: what you need to know when filing a claim. Technical report, FEMA. GAO-08-7
Finkelstein A, Hendren N, Shepard M (2019) Subsidizing health insurance for low-income adults: evidence from Massachusetts. Am Econ Rev 109(4):1530–1567
Frazee G (2017) California's wildfire victims could be like most homeowners: underinsured
Fried S (2022) Seawalls and stilts: a quantitative macro study of climate adaptation. Rev Econ Stud 89(6):3303–3344
Gallagher J (2014) Learning about an infrequent event: evidence from flood insurance take-up in the United States. Am Econ J Appl Econ 6(3):206–233
Gallagher-Re (2023) Reinsurance market report - full year 2022. Technical report, Arthur J. Gallagher and Co
GAO (2021a) FEMA flood maps: better planning and analysis needed to address current and future flood hazards. Technical report, Government Accountability Office
GAO (2021b) National flood insurance program: congress should consider updating the mandatory purchase requirement. Technical report, Government Accountability Office
GAO (2023) Farm bill: reducing crop insurance costs could fund other priorities. Technical report, Government Accountability Office
Garnache C (2020) Does the salience of risk affect large, risky asset prices? Working Paper
Gaul GM (2019) The geography of risk. Sarah Crichton Books
Gibson M, Mullins J (2020) Climate risk and beliefs in New York floodplains. J Assoc Environ Resour Econ 7(6):1069–1111
Gourevitch JD, Kousky C, Liao P, Nolte C, Pollack AB, Porter JR, Weill JA (2023) Unpriced climate risk and the potential consequences of overvaluation in US housing markets. Nat Climate Change 13:250–257
Hallstrom D, Smith VK. (2005) Market responses to hurricanes. J Environ Econ Manag 50(3):541–561
Handel B (2013) Adverse selection and inertia in health insurance markets: when nudging hurts. Am Econ Rev 103(7):2643–2682
Handel B, Kolstad J (2015) Health insurance for humans: information frictions, plan choice, and consumer welfare. Am Econ Rev 105(8):2449–2500
Handel B, Kolstad J, Spinnewijn J (2019) Information frictions and adverse selection: policy interventions in health insurance markets. Rev Econ Stat 101(2):326–340
Handel B, Schwartzstein J (2018) Frictions or mental gaps: what's behind the information we (don't) use and when do we care? J Econ Perspect 32(1):155–178
Healy J (2018) By all appearances, they overcame hurricane Harvey. Appearances are deceiving
Hendren N (2020) Measuring ex-ante welfare in insurance markets. Rev Econ Stud 88(3):1193–1223
Hino M, Burke M (2021) The effect of information about climate risk on property values. Proc Natl Acad Sci 118(17):1–9
Horn D (2022) National flood insurance program borrowing authority. Technical report, Congressional Research Service
Horn DP. (2018) Federal disaster assistance: the national flood insurance program and other federal disaster assistance programs available to individuals and households after a flood. Technical report, Congressional Research Service

Horn DP, Brown J (2018) Introduction to the national flood insurance program (NFIP). Technical report, Congressional Research Service
Horn DP, Webel B (2018) Private flood insurance and the national flood insurance program. Technical report, Congressional Research Service
Hovekamp WP, Wagner KRH (2023) Efficient adaptation to flood risk. Working Paper
Hurst E, Keys BJ, Seru A, Vavra J (2016) Residential redistribution through the U.S. mortgage market. Am Econ Rev 106(10):2982–3028
Jaffee DM, Russell T (1997) Catastrophe insurance, capital markets, and uninsurable risk. J Risk Insurance 64(2):205–230
Kelly JV (2017) FEMA needs to improve management of its flood mapping programs. Technical report, Department of Homeland Security
Klein K (2018) Minding the protection gap: resolving unintended, pervasive, and profound homeowner insurance. Conn. Insur. Law J 25(24):36–114
Kousky C (2010) Learning from extreme events. Land Econ 86(3):395–422
Kousky C, Kunreuther H (2018) Risk management roles of the public and private sector. Risk Manag Insur Rev 21(1):181–204
Kousky C, Lingle B, Kunreuther H, Shabman L (2019) Moving the needle on closing the flood insurance gap. Technical report, Wharton Risk Management and Decision Center
Kousky C, Michel-Kerjan EO, Raschky PA (2018) Does federal disaster assistance crowd out flood insurance? J Environ Econ Manag 87:150–164
Kousky C, Mulder P (2023) Risk rating without information provision. Working Paper
Kriesel W, Landry C (2004) Participation in the national flood insurance program: an empirical analysis for coastal properties. J Risk Insur 71(3):405–420
LaCour-Little M, Pavlov AD, Wachter SM (2022) Adverse selection and climate risk: a response to Ouazad and Kahn (2021). Working Paper
Lee S (2022) Adapting to natural disasters through better information: evidence from the home seller disclosure requirement. Working Paper
Li W, Goodman L (2016) How much house do Americans really own? Measuring America's accessible housing wealth by geography and age. Technical report, Urban Institute Housing Finance Policy Center
Liao P, Mulder P (2021) What's at stake? Understanding the role of home equity in flood insurance demand. Working Paper
Liao Y, Walls MA, Wibbenmeyer M, Pesek S (2022) Insurance availability and affordability under increasing wildfire risk in California. Resour Future Issue Brief 22-09
Mahoney N (2015) Bankruptcy as implicit health insurance. Am Econ Rev 105(2):61–107
Marshall D (2017) An overview of the California earthquake authority. Technical report, Resources for the Future
McCoy SJ, Walsh RJ (2018) Wildfire risk, salience, and house price demand. J Environ Econ Manag 91:203–228
McGuire CJ (2018) Considering the role of government in communicating climate change: lessons from the us public flood insurance program. In: Filho WL (ed) Handbook of climate change communication, vol. 3. Springer, Berlin
Michel-Kerjan EO (2010) Catastrophe economics: the national flood insurance program. J Econ Perspect 24(4):165–186
Mulder P (2022) Mismeasuring risk: the welfare effects of flood risk information. Working Paper
NAIC (2023) Risk-based capital. Technical report, National Association of Insurance Commissioners
Nguyen DD, Ongena S, Qi S, Sila V (2022) Climate change risk and the cost of mortgage credit. Rev Financ 26(6):1509–1549
NRC (2015) Affordability of national flood insurance program premiums, report 1. Technical report, National Research Council
OECD (2018) The contribution of reinsurance markets to managing catastrophe risk. Technical report, OECD
Oh S, Sen I, Tenekedjieva A-M (2021) Pricing of climate risk insurance: regulatory frictions and cross-subsidies. Available at SSRN 3762235
Ostriker A, Russo A (2022) The effects of floodplain regulation on housing markets. Working Paper
Ouazad A, Kahn M (2022) Mortgage finance and climate change: securitization dynamics in the aftermath of natural disasters. Rev Financ Stud 35(8):3617–3665
Panjwani A (2022) The effects of federal policies on homeowners' exposure to climate change risk. Working Paper
Pauly MV, Blavin FE (2008) Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. J Health Econ 27(6):1407–1417
Petrolia DR, Landry CE, Coble KH (2013) Risk preferences, risk perceptions, and flood insurance. Land Econ 89(2):291–306
Powell L (2020) Risk-based pricing of property and liability insurance. J Insur Regul 1
Profita C (2022) Oregon postpones wildfire risk mapping and rulemaking plans after public backlash
Rothschild M, Stiglitz J (1976) Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Q J Econ 90(4):629–649

Roy AD (1952) Safety first and the holding of assets. Econom J Econ Soc 431–449
Royal A, Walls M (2019) Flood risk perceptions and insurance choice: do decisions in the floodplain reflect overoptimism? Risk Anal 39(5):1088–1104
Ruoff S (2023) Catastrophe bonds: a perfectly uncorrelated storm. Technical report, Schroder Investment Management Australia Limited
Sastry P (2022) Who bears flood risk? Evidence from mortgage markets in Florida. Working Paper
Shi L, Fisher A, Brenner RM, Greiner-Safi A, Shepard C, Vanucchi J (2022) Equitable buyouts? Learning from state, county, and local floodplain management programs. Climatic Change 174:29
Singh R (2019) Seismic risk and house prices: evidence from earthquake fault zoning. Region Sci Urban Econ 75:187–209
Stone JM (1973) A theory of capacity and the insurance of catastrophe risks (part I). J Risk Insur 231–243
Straub S, Rakette A, Apel U, Thoms F, McNamara M, Ng L (2023) Climate change and La Nina driving losses: the natural disaster figures for 2022. Technical report, Munich Re
Sunstein C (2002) Probability neglect: emotions, worst cases, and the law. Yale Law J 112:61–107
Sydnor J (2010) Overinsuring modest risks. Am Econ J Appl Econ 2(4):177–199
Tsiboe F, Turner D (2023) Crop insurance at a glance
USGAO (2007) Natural disasters: public policy options for changing the federal role in natural catastrophe insurance. Technical report, USGAO. GAO-08-7
Wagner KRH (2022a) Adaptation and adverse selection in markets for natural disaster insurance. Am Econ J Econ Policy 14(3):380–421
Wagner KRH (2022b) Designing insurance for climate change. Nat Climate Change 12(12):1070–1072
Webel B, Cobb C (2005) Insurance regulation: history, background, and recent congressional oversight. Technical report, Congressional Research Service
Weill JA. (2022) Perilous flood risk assessments. Working Paper
Weitzman ML (2009) On modeling and interpreting the economics of catastrophic climate change. Rev Econ Stat 91(1):1–19
Wing OEJ, Bates PD, Smith AM, Sampson CC, Johnson KA, Fargione J, Morefield P (2018) Estimates of present and future flood risk in the conterminous United States. Environ Res Lett (13):1–7
Zhang J, Czajkowski J, Botzen WW, Robinson PJ, Tesselaar M (2022) Assessing the drivers of intrinsically complex hurricane insurance purchases: lessons learned from survey data in Florida. In: Hurricane risk in a changing climate. Springer, Berlin, pp 283–321
Zhou-Richter T, Browne MJ, Gruendl H (2010) Don't they care? Or, are they just unaware? Risk perception and the demand for long-term care insurance. J Risk Insur 77(4):715–747



Climate Risk and Insurance

W. J. Wouter Botzen and Peter J. Robinson

1 Introduction

Natural disasters can cause extensive losses to human life, damages to physical assets, economic changes to production and employment, as well as social effects like distributional impacts (Markhvida et al. 2020). Between 1980 and 2014, there was a three-factor increase in the number of natural disasters, which was primarily attributable to weather hazards like storms and floods (Hoeppe 2016). It has been shown that the rise in the frequency of natural disasters is in part due to climate change (van Aalst 2006; Otto et al. 2018), which has also led to an increase in extreme damages from natural disasters over time (Coronese et al. 2019). Increasing natural disaster risks can pose challenges for supplying insurance coverage against damages from extreme weather at affordable premiums (Botzen 2021). Additionally, demand for insurance against natural disasters is typically low (Kunreuther and Pauly 2004), which can be partly caused by irrationalities in human decision-making processes (Meyer and Kunreuther 2017). In 2016, only 30% of natural disaster losses were insured worldwide (Swiss Re 2017).

While natural disaster risks associated with climate change relate to decades-long uncertainty about future climate conditions, disaster insurance is an annual contract. In this regard, it is important that insurers ensure that risks are priced accurately and financed sufficiently. However, as expected losses grow over time, premiums also rise, which means that policyholders carry a significant burden of the long-term risks of climate change. Managing climate risk requires avoidance (e.g., relocation) and effective risk reduction (e.g., improved building codes and the implementation of individual-level adaptation strategies). Higher insurance prices may also motivate households at risk to avoid or engage in risk reduction. Nevertheless, it is not clear to what extent insurance price signals will encourage individuals to adapt, given the oftentimes low demand for insurance against natural

W. J. W. Botzen (✉) · P. J. Robinson
Department of Environmental Economics, Institute for Environmental Studies (IVM), VU Amsterdam, Amsterdam, Netherlands
Utrecht University School of Economics (U.S.E.), Utrecht University, Utrecht, Netherlands
e-mail: wouter.botzen@vu.nl

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_3

disasters observed in practice. Other insurance market challenges remain with respect to climate change. This chapter will draw insights from empirical and modeling studies to discuss present supply- and demand-side challenges in detail regarding the insurance of climate risk, as well as issues that may arise over time with respect to climate change impacts.

As a starting point, it is useful to note a few effects of climate change on insurance markets that can already be observed. For instance, natural disaster insurance premiums have risen in response to higher insurance claims over time, and some insurers have had to withdraw coverage altogether in very high-risk areas (Loncy 2023). In addition, insurers have adapted their risk assessment practices and developed new insurance products to address climate risks (Golnaraghi 2018). One example is parametric insurance in which payouts are based on the magnitude of weather-related events as opposed to the magnitude of losses (NAIC 2023). To mitigate future risk, insurance companies are also integrating environmental, social, and governance (ESG) investment criteria and divesting from fossil fuel companies (Egli et al. 2022). Some insurers also offer incentives for individual-level risk reduction, including premium reductions for policyholders who retrofit their homes against weather-related damages (Kunreuther 2021).

There are also several impacts that insurance companies may face in the future, given the trend of rising natural disaster damages caused by climate change. For example, there are direct impacts on the way insurers operate, as is the case when they miscalculate risk-based policy premiums due to reliance on outdated models and/or historical damage data that insufficiently account for current climate change trends (Krauss 2021). Furthermore, insurer insolvency is more likely in a changing climate, especially since many losses occurring at once in a single event become more probable due to the high correlation of climate risks (Botzen and van den Bergh 2008).[1] For this reason, it is vital that insurers hold higher capital reserves or purchase additional reinsurance coverage to enhance their capacity to reimburse claims from natural hazards.

Other effects of climate change on the insurance sector will occur as a result of a change in the types of contracts offered by insurers. That is, insurance contracts are subject to uncertainties inherent to climate change impacts over time. This is because climate risks are difficult to predict because relatively limited past data are available (Kunreuther 1996),[2] and there is considerable uncertainty regarding climate scenarios and the actual impact of climate change on the frequency and severity of natural disaster risks. Overall, this means that the future claims faced by insurance companies are also uncertain. Consequently, insurance companies may raise deductibles (Tucker 1997) or refuse to offer long-term insurance products, such as contracts with fixed annual prices that cover policyholders for time periods in excess of one year, because of lack of insight into future risk (Aerts and Botzen 2011a). This can also affect consumer welfare, as it has been shown that potential policyholders prefer long-term fixed-price insurance against natural disasters over short-term policies with fluctuating annual prices (Kunreuther and Michel-Kerjan 2015). Insurers are likely to continue to raise premiums, in light of climate change uncertainty, to a level that is significantly higher than actuarially fair levels due to their uncertainty aversion (Kunreuther et al. 1995). Swiss Re (2020) predicts that, in the case of the National Flood Insurance Program (NFIP) in the United States, the average premium per policy will increase by 70% by 2100.

Although climate change poses a number of significant risks for insurance markets, the insurance sector can play a role in managing increasing climate risks through partnerships formed with the public sector and by implementing strategies that stimulate risk reduction by policyholders. This chapter makes proposals for the design of insurance and other policies to effectively adapt to climate

[1] Climate risks are correlated both spatially and over multiple hazards (climate change can impact the incidence of windstorms as well as river and coastal flooding, and all of these risks can occur at once).

[2] Predicting future climate risk is difficult even with sufficient historical data due to model parameter uncertainty.

change. In addition, the chapter summarizes the state of the art with respect to the assessment of climate risk for the design of insurance. Suggestions for future research are also presented in this chapter.

The remainder of this chapter is structured as follows: Section 2 provides an overview of the main natural disaster risks that are expected to change in the future as a result of climate change and discusses methods (statistical approaches and catastrophe models) for assessing climate risk for insurance. Section 3 discusses the impacts of climate change on the supply side of insurance. It outlines in detail the different types of (private, public, public–private) arrangements for the supply of insurance against natural disaster risks and potential reforms of these arrangements for coping with climate change. Section 4 reviews empirical evidence of demand for insurance against low-probability natural disasters and describes various theories and behavioral biases that may explain why this demand is often low. Moreover, it discusses the results of modeling studies that simulate how insurance demand may change under future climate conditions. Section 5 proposes how insurance arrangements, complemented with other policies, can stimulate policyholders to take risk reduction measures and adapt to climate change risks. Section 6 concludes with the main takeaways for insurers and policymakers and outlines an agenda for future research.

2 Assessing Climate Change Risks for Insurance

It is pivotal for insurers to accurately assess natural disaster risks and account for changes in the underlying risk drivers over time in these assessments. These natural disaster risk assessments are commonly used in the insurance industry to estimate the insured losses caused by specific types of natural hazards, such as floods or storms, in areas where the insurer has covered damages that can be caused by these events, such as property losses. Outputs of such risk assessments include expected insured losses per natural hazard event, loss distributions that indicate which loss amounts occur with specific probabilities, and annual expected losses. These risk assessments commonly form the basis for both setting insurance premiums and insurers' financial risk management, such as decisions about required reserves and reinsurance coverage (Grossi and Kunreuther 2005). Hence, changes that have already occurred in the frequency and/or severity of extreme weather events as a result of climate change should ideally be reflected in assessments of current natural disaster risks. Moreover, forward-looking assessments of future natural disaster risk may be useful for stress-testing how current arrangements for insuring natural disaster losses cope with climate change impacts (Charpentier 2008). Such climate risk assessments provide insights into future changes of risks, which can be input for decisions by insurance companies to adapt insurance products, such as premiums and coverage conditions, or financial risk management practices (Mills 2012). Moreover, they inform adaptation strategies that reduce natural disaster risks so that insurers can stimulate their policyholders to take or implement them together with the public sector, such as improved building codes (see Sect. 5).

Assessments of natural disaster risks can be based on historical claims payments of losses from extreme weather events or simulation methods that are often called catastrophe models (Zhou et al. 2023). Risk assessments of historical claims are commonly applied for extreme weather types that occur relatively frequently, such as rainfall events, meaning that sufficient historical observations are available for assessing loss distributions. A disadvantage of using historical loss records may be that the resulting loss distributions are no longer representative once climate change has altered the frequency or severity of extreme weather events (Dlugolecki 2008). Therefore, some studies have estimated the relationships between historical extreme weather indicators and insurance claims, which have been used to forecast future changes in losses (e.g., Botzen et al. 2010; Botzen and Bouwer 2016). This approach extrapolates losses under scenarios of changes in extreme weather

characteristics, as a result of climate change, derived from climate models. An example is Botzen and Bouwer (2016) who estimated the statistical relationships between historical claims of losses caused by hailstorms to vehicles and local weather indicators derived from meteorological weather stations in the Netherlands. Forecasting analyses based on these relationships under climate change scenarios indicate that these losses may increase up to 33% by the year 2050. This method is suitable for hailstorm risk projections, given the sufficient number of historical claims available for statistical analysis. However, this is not applicable to low-probability extreme weather events, such as extreme floods at a particular location. Another limitation is that additional adaptation measures that could limit future losses are not accounted for.

Natural disaster risk of low-probability extreme weather types, such as extreme floods, is commonly assessed with catastrophe models (Botzen et al. 2019). These models simulate hypothetical, but realistic, natural hazard characteristics and estimate potential damages based on exposed property values and their vulnerability (i.e., susceptibility to damage). Examples are flood damage models that simulate potential flood inundation extents and depths at a particular location for potential flood events with various severities and associated occurrence probabilities (see de Moel et al. 2015 for a review). Next, these spatially explicit flood maps are combined with information on exposed properties and their values and vulnerability in terms of depth–damage curves that estimate the building value lost at a particular inundation depth. The final model results include potential damages for various types of floods at a particular location and associated loss probability distributions. These models can be forward-looking and simulate future changes in flood risks based on the changes in flood hazard characteristics under climate change combined with expected changes in exposure, for instance, to assess potential changes in flood insurance premiums and their affordability (Hudson et al. 2016).

Some studies applying catastrophe models also assess how adaptation to changing natural disaster risks over time limits the impacts of climate change on losses. These models are often applied through simplified scenarios of adaptation measures (see Aerts 2020 for a review). An exemplary study is Kunreuther et al. (2013) who assessed the implications of changes in future hurricane risk under climate change scenarios for insurance pricing in Florida. Moreover, they simulated how building codes that enhance the resistance of properties to wind damage can significantly limit the future rise in hurricane risk. However, in practice, adaptation decisions depend on risk reduction behavior by households, governments, and private sector agents that are dependent on various factors, such as the occurrence of natural disasters, changing risks, and other decisions.

Recent studies have proposed agent-based models (ABMs) for integrating human decision-making into forward-looking natural disaster risk assessments. ABMs allow for representing decisions on risk reduction from heterogeneous household and business agents that respond to each other's decisions, their environment, changing risks, and government and insurance strategies (Aerts et al. 2018). For instance, Haer et al. (2019) integrate human adaptation decisions in an EU-scale flood damage model using an ABM to estimate future flood risk under climate change scenarios. The results show that government investments in flood protection and household investments in flood-proofing homes can significantly limit the increase in flood risk from climate change (Haer et al. 2019). Moreover, since flood risk awareness that influences adaptation decisions in the ABM is triggered by the occurrence of flood events, the adoption of risk reduction measures as well as the resulting risk outcome fluctuates over time depending on the timing of flood disasters (Haer et al. 2019). These forward-looking ABMs are able to capture key dynamics in adaptation and risk that are also observed in practice, but a remaining challenge is to empirically calibrate and validate the behavioral rules that guide the adaptation decisions in the model (Aerts 2020).

3 Supply of Natural Disaster Insurance in a Changing Climate

The effect of climate changes on natural disaster risk for insurance consumers depends on the way disaster coverage is supplied. Damage caused by relatively common extreme weather events, such as extreme rainfall and windstorms, is often insurable in private markets in high-income countries (Botzen 2021). However, rare extreme weather events that may cause relatively high damages in concentrated areas, such as extreme floods, are more challenging to privately insure at affordable premiums (Botzen and van den Bergh 2008). The reasons include that the correlated nature of these risks implies that high accumulated losses may occur following a disaster event, which requires holding large capital reserves and relatively expensive reinsurance coverage. Moreover, uncertainties associated with estimating the risks of rare and extreme events add to high premium loading factors (Kunreuther et al. 1995). Furthermore, adverse selection may happen when only consumers with a relatively high risk purchase the insurance and those with a relatively low risk do not. This selection creates a small pool of consumers for spreading risks that add to high insurance costs.

Despite these challenges, private insurance products for extreme natural disaster events are available for purchase in voluntary private markets in various countries, such as flood insurance in Germany (Schwarze et al. 2011). In other countries, natural disaster insurance is offered by a public insurer and combined with requirements for consumers to purchase coverage to create a large pool for spreading risks. An example is Spain (Paudel et al. 2012). One advantage of private market insurance is that it is likely to be more efficient because, in a competitive market, insurers have an incentive to offer coverage at a low cost and premium for consumers. This efficiency holds at least if the risks covered are not too extreme since private reinsurance coverage for such risks can be relatively expensive. Additional transaction costs are especially limited if coverage against extreme weather events is included in common property insurance products, such as for fires and thefts. An advantage of public insurance is that it does not require high premium loading factors for profit and reinsurance costs, while governments also have a strong capacity to spread risks intertemporally by borrowing from international financial markets (to be repaid by future taxes). Some countries, such as France, have combined these comparative advantages in public–private natural disaster insurance coverage (Poussin et al. 2013). In such a system, private insurers cover part of the natural disaster risks (e.g., up to a certain maximum loss amount) and handle the administration of policies, while a public reinsurer covers the extreme portion of the risks. This involvement of the public sector can guarantee the financial stability of the insurance system and prevent the need to purchase relatively expensive private reinsurance, which contributes to the affordability of premiums. Moreover, combining public–private coverage with purchase requirements can result in a high degree of spreading of risks across policyholders (Hudson et al. 2020).

Tesselaar et al. (2020a, 2022) estimate how changing flood risk under various climate change scenarios affects existing private, public, and public–private flood insurance arrangements in EU countries. They also examine which insurance system reforms are needed to cope with climate risk. For this purpose, they further develop the dynamic integrated flood and insurance (DIFI) model from Hudson et al. (2019), which is a partial equilibrium model of the demand and supply of insurance coverage against river flood risk. DIFI is integrated into a flood damage model that simulates spatially explicit flood risk under various climate change scenarios. These flood risk estimates determine insurance supply decisions, such as the premiums to be charged to consumers. Additionally, flood risk is input into a consumer demand module in which households that maximize their (subjective) expected utility decide whether or not to purchase insurance based on the flood risk they face, insurance costs, wealth, and disposable income.

Tesselaar et al. (2020a) show that climate change is expected to increase flood risk in various regions across the EU due to an increased frequency or severity of flooding. This increase in flood risk faced by insurers is passed on to consumers and results in rises in flood insurance premiums, especially in countries where flood insurance is supplied in private markets without purchase requirements. In regions with relatively low income, such as in Eastern Europe, many households may be unable to afford these higher premiums and may decide to drop their flood insurance coverage. These premium increases and drops in insurance demand are less problematic in countries with public or public–private flood insurance arrangements. In order to cope with climate risk, model simulations by Tesselaar et al. (2020a) show that it may be useful to introduce insurance reforms that include elements of public–private flood insurance arrangements. Such a reform helps keeping premiums affordable due to public reinsurance, especially if incentives to stimulate risk reduction are integrated into insurance (see Sect. 5). Tesselaar et al. (2022) extend this analysis to include existing arrangements for (ad hoc) government compensation for flood damage. Although such compensation can be useful for addressing uninsured losses, its existence crowds out demand for private flood insurance coverage in markets where it can be purchased voluntarily. This contributes to a widening of the insurance coverage gap under climate change.

4 Demand for Natural Disaster Insurance in a Changing Climate

Well-functioning natural disaster insurance markets, without strict purchase requirements, depend on sufficient consumer demand. This section provides an overview of empirical studies of demand for insurance against climate risk, with a focus on low-probability natural disasters. There are two general approaches to modeling risky choices (Pachur et al. 2017). The first approach focuses on predicting decisions based on theories that describe decision-making in terms of a mathematical expectation calculation, based on (transformed) probabilities and outcomes, e.g., expected utility theory (EUT) and prospect theory (PT) (see sections "Theory", "Charity Hazard", "Correlated Risk", "Probability Weighting"). The second approach characterizes decision-making according to simple mental shortcuts (heuristics), which means that relevant information, such as probabilities and outcomes, may be systematically ignored and that behavioral biases are present (see section "Heuristics and Behavioral Biases").[3] The main takeaway of this section is that there are various reasons why individuals might be underinsured in practice against low-probability natural disasters based on a description of the empirical evidence base of demand for insurance. These reasons are discussed in the context of decision models that guide patterns of behavior.

How demand for insurance against natural disaster risk may be influenced by climate change is also described based on insights derived from empirical and simulation studies (Sect. 4.2). That is, the possible evolution of insurance demand against climate risk over time is described according to studies that observe and simulate how insurance demand may change under future climate conditions.

[3] The included models of behavior are the most commonly referred to theories in the literature on insurance choices under low-probability/high-impact risks, such as climate risks, or frequently cited heuristic models used to reconcile why individuals underprepare for natural disasters (Robinson and Botzen 2019; Meyer and Kunreuther 2017).

4.1 Low Demand for Insurance Against Natural Disasters

Theory

The most commonly adopted numerical theory of decision-making applied in economic studies of low-probability/high-impact (LPHI) insurance choices, such as those that concern natural disasters associated with climate risk, is the EUT (Robinson and Botzen 2019). EUT, first axiomized by von von Neumann and Morgenstern (1944), is a rational model of risky choice that assumes that individuals make decisions involving probabilities and money by valuing final wealth amounts under different states of nature and then weighting these valuations by their probabilities of occurring. Utility determines how individuals value money, i.e., final wealth. According to the EUT, individuals choose options that maximize their expected utility. Applying EUT to insurance decisions for natural disaster events associated with climate risk implies that individuals process the, often very low, likelihood of risk occurrence linearly and that final wealth states with high-impact losses are transformed by an at-risk individual's utility function.

A major puzzle for insurance economics, and EUTs, is that individuals tend to prefer insuring high-probability/low-impact (HPLI) losses to LPHI losses (Browne et al. 2015). This pattern of behavior has been shown to occur in experimental studies that mimic disaster risk decisions (Slovic et al. 1977; Schoemaker and Kunreuther 1979),[4] as well as studies based on market data that compare flood insurance purchase and insurance purchase for bicycle theft (Browne et al. 2015). Assuming that individuals are risk averse, under EUT, there should be a higher preference for insuring LPHI risk to HPLI risk of the same expected value.

Charity Hazard

The tendency for individuals to underinsure LPHI risk, compared with HPLI risk, has led many to offer explanations of the observed behavior, some of which can be reconciled within the framework of the EUT. For instance, insufficient demand for insurance against climate risk may be due to individuals rationally anticipating that the government will compensate for their uninsured damages after a disaster has occurred, i.e., charity hazard (Raschky and Weck-Hannemann 2007).

Nevertheless, empirical evidence on the prevalence of charity hazards in practice is mixed, which may be due to inconsistencies in applied study methodologies. Observational studies adopting NFIP policy-in-force data (Browne and Hoyt 2000) as well as surveys of flood and wind insurance purchases (Petrolia et al. 2013, 2015) predict a positive or null effect of anticipated government relief on insurance demand. However, it has been suspected that there is endogeneity involved in studying the charity hazard problem due to a number of potential sources, such as differences in objective risk, which can impact both demand for insurance and government compensation payments; differences in individuals' education levels, which may cause variations in insurance demand and their ability to navigate the bureaucratic red tape required to successfully apply for government compensation; and reverse causality due to the desire of the government to assist uninsured households, which means that low demand for insurance may lead to high levels of government assistance, not vice versa.

More recent studies of insurance against natural disaster risk, which control for endogeneity by exploiting data that results in exogenous variation in government compensation, like timing of presidential elections, swing states, or other political factors, find evidence in favor of charity hazards (Kousky et al. 2018; Landry et al. 2021). Robinson et al. (2021b), in an experimental study with randomly varied schemes of government compensation, find that compensation has the potential to

[4] More recent studies by Laury et al. (2009) and Bajtelsmit et al. (2015) showed that this behavioral pattern may be an artifact of nonincentive compatible reward structures adopted by the earlier studies.

crowd out demand for insurance against natural disaster risk in cases where the probability and extent of compensation are known with certainty.

Correlated Risk

Friedl et al. (2014) offer an alternative account of the underinsurance of LPHI risk. The authors disentangle two types of insurable risk, i.e., risk where losses are correlated with those of one's peers and idiosyncratic risk where losses do not exhibit this correlation. The former closely resembles the case of climate risk since future losses from natural disasters will most likely affect not only one household or business but also neighboring households and businesses. The latter is typical of burglary risk and the risk of loss of one's mobile phone, for instance. It is proposed that social comparison, as well as the assumption that individuals are inequality averse and dislike being worse off than their peers to a larger extent than they like being better off than their peers (Fehr and Schmidt 1999),[5] may lead a substantial subgroup of the population to remain uninsured against correlated climate risk.

The extent to which individuals actually observe the insurance decisions of others in practice is perhaps unclear, although some observational studies suggest that individuals do form social expectations (e.g., Lo 2013). Nevertheless, the intuition is that there is less of a feeling of loss if peers also lose, which is the case for correlated risk. Whereas, for uncorrelated risk, the loss is for the individual only, and therefore felt more strongly (Richter et al. 2014).

Social comparison in insurance demand is formalized by Friedl et al. (2014) in a model that uses EUT as its foundation and extends preferences to account for inequality aversion. It is shown that one Nash equilibrium is for individuals not to purchase insurance against correlated risk when strategic interaction is present. An experiment was also conducted by Friedl et al. (2014) to compare willingness-to-pay for insurance against correlated versus idiosyncratic risk and revealed that, on average, willingness-to-pay is lower for covering the correlated risk, in line with the theoretical findings.

Probability Weighting

Violations of EUT's independence axiom (e.g., Allais 1953), which implies that preferences between options should not be affected by commonly added risky outcomes or lotteries to the options, have led to the development of more descriptively accurate theories of decision-making that include, among other things, the tendency of individuals to weigh probability information nonlinearly. The most well known of these theories is PT, which relaxes the independence axiom by replacing probabilities with decision weights to accommodate individual nonlinear weighting of probabilities (Tversky and Kahneman 1992). This means that climate risks, which have very low probabilities of occurrence, can be either under- or overweighted, depending on an individual's risk attitude (Hajimoladarvish 2017).

Robinson and Botzen (2020) find that flooding risks are significantly underweighted compared with the risks investigated by other papers that study probability weighting for more general lottery formats in the loss domain (e.g., Etchart-Vincent 2004; Abdellaoui 2000). Moreover, it has been shown explicitly that individuals tend to weigh probabilities based on the frequency of occurrence of the event despite providing objective risk information. Specifically, Kusev et al. (2009) found that probability weighting functions derived for low-frequency risks, like climate risks, display more underweighting than high-frequency risks, like theft and home burglary. The implication is that individuals weigh probabilities based on the accessibility of events in memory and that experiences leak into risky decisions despite a specified probability and loss of information.

[5] See also the concept of loss aversion, which implies that individuals are more sensitive to losses than to gains in decisions (Kahneman and Tversky 1979).

Some have also used probability weighting to explain the low demand for insurance against LPHI risks based on the delay dependency of risk preferences. For instance, Fehr-Duda and Fehr (2016) proposed that individuals assign less weight to low-probability events when uncertainty resolves further into the future. The result is that natural disaster events caused in part by climate change, which are typically perceived as potentially occurring many years from now, are significantly underweighted in decision processes. This is contrary to short-term insurance contracts that had very high demand in the past, like accidental death policies that were sold at airports in the United States to cover flight-related accidents that also ended once flights had completed (Eisner and Strotz 1961).

Heuristics and Behavioral Biases

Aside from theoretical justifications for the tendency to underinsure climate risk based on EUT and PT, another view is that underinsurance could be caused by mental shortcuts. Mental shortcuts (heuristics) are processes that make problems less complex by (un)consciously ignoring some information pertaining to the decision problem. In the presence of climate risk, it is perhaps reasonable to expect that due to cognitive limitations, which imply that individuals find it difficult to process very low probabilities (Kunreuther et al. 2001) and limited access to information (e.g., probabilities and losses),[6] individuals adopt mental shortcuts to deal with these limitations (Kahneman 2003). This subsection focuses on evidence of individuals utilizing heuristics or displaying behavioral biases in decisions involving climate risk, i.e., errors in judgment that deviate behavior from what would be considered "optimal" behavior in the sense of EUT (Gigerenzer 1991).

Availability Heuristic

Individuals may judge the probability of an event by the ease with which similar examples come to mind (Tversky and Kahneman 1974). For instance, one may assess the risk of flooding by recalling flood events that have occurred recently, among friends and family, or have been reported in the media and ignore other objective information that may signal flood risk, like flood maps, property elevation, and distance of one's home to the nearest river.

There are a number of papers that use market data on flood insurance in the United States showing that demand for insurance is positively related to previous disaster experiences (Kriesel and Landry 2004; Browne and Hoyt 2000; Michel-Kerjan and Kousky 2010; Zahran et al. 2009; Kousky 2011, 2017), in line with the availability heuristic. Nevertheless, the influence of experience on flood insurance demand diminishes after several years of a flood event (Atreya et al. 2015). Gallagher (2014) explores the relationship between flood experience and insurance demand in more detail and finds that, in addition to a rise in demand for insurance among individuals directly affected by a flood, unaffected individuals who share the same television content as those in affected areas have a higher demand for insurance following a flood. However, geographically close neighbors to flooded areas who do not share the same television content do not purchase more insurance following a flood event. It appears that flood insurance demand is impacted by a flood event among individuals for which information about flooding becomes available following the event.

Threshold of Concern Heuristic

Open-ended answers to questions posed by Slovic et al. (1977) in their experimental study of why individuals fail to insure low-probability risks demonstrated that they may follow a threshold model decision heuristic. The model entails that individuals are not concerned about losses whose probability

[6] See Kunreuther and Pauly (2004).

is below some threshold. Therefore, loss probabilities that fall below the threshold are treated as though they equal zero.

Other experiments of willingness-to-pay for insurance against low-probability risk find that there is a bimodal pattern of payments. That is, a large proportion of individuals are willing to pay nothing to cover the risk, and another proportion of individuals, who are very risk averse, are willing to pay values significantly larger than the expected value of risk (McClelland et al. 1993; Irwin et al. 1992; Robinson and Botzen 2018). The behavior of the first subgroup is consistent with a threshold model decision heuristic and perhaps reflects the judgment that the risk is too small to be concerned about. It has also been shown that individuals who indicate that they adopt the threshold of concern decision rule explicitly have lower risk perceptions and lack awareness of living in a hazard-prone area (Botzen et al. 2015).

Inertia

Meyer and Kunreuther (2017) suggest that the tendency to maintain the status quo may be an underlying cause of low natural disaster insurance demand, especially in cases where there is considerable uncertainty about the possible benefits of holding insurance. Implicitly, if one were to change a default of no insurance coverage against a disaster in homeowners' policies to include disaster coverage in the policy by default, significantly more homeowners would purchase such coverage.

Two recent empirical studies support this hypothesis. Kunreuther et al. (2021) find, in a study of disaster insurance demand among Canadian homeowners, that individuals have a probability greater than 1.6 of purchasing insurance coverage when they are assigned to a default where they must opt out to cancel their coverage (in exchange for a premium discount) rather than a default where they must opt in to purchase coverage by adding it to a homeowners' policy. Robinson et al. (2021a) show that defaults raise the likelihood of purchasing insurance against flood risk by between 17 and 18 percentage points among Dutch homeowners who have faced relatively minor flood losses in practice and for whom only a small percentage had purchased flood insurance at the time of the study. However, the same study found a null effect of the default on flood insurance demand in the United Kingdom, where the actual purchase of flood insurance and flooding experience is more common. It seems that learning about risk and experience with the decision at hand can impact the magnitude of the default effect, an insight that has been replicated in other decision-making contexts (Löfgren et al. 2012).

Herding

In addition to inertia, it was proposed by Meyer and Kunreuther (2017) that there is a tendency for individuals to base insurance decisions involving natural disaster risk on the observed actions of others. Consistent with this proposal, some survey studies find that there is a relationship between whether individuals perceive that there is a social norm for insurance against disaster risk and the purchase of such insurance (Lo 2013; Zhang et al. 2022). Therefore, it can be inferred that if strong social norms for insurance are developed, this may help raise coverage rates among individuals in disaster-prone areas.

Nevertheless, experimental studies by Mol et al. (2021) and Robinson and Botzen (2022), who employed a norm nudge that conveyed to respondents the flood preparedness behavior of the majority, on aggregate, found the nudge had no influence on respondents' protection choices. The latter study also examined moderators of the norm nudge and found that there are two subgroups for which the

nudge would increase flood insurance demand, namely, those whose prior beliefs are consistent with the norm, as well as those who trust insurance agents, who were used as messengers of the norm. This finding highlights the importance of considering heterogeneity with regard to consumer beliefs and preferences since a one-size-fits-all approach to nudging behavior may be ineffective or in some cases results in the opposite of the intended behavior.

4.2 Impact of Climate Change on Demand for Insurance Against Natural Disasters

A number of studies investigate the potential influence of climate change on demand for insurance against natural disasters, either by using empirical data on individual insurance choices under higher objective loss probabilities or by implementing computational models that simulate the microeconomic decisions of individuals and other societal actors facing disaster risk. Regarding the former, Botzen and van den Bergh (2012a) examine the effects of climate change on flood insurance demand in a choice experiment with flood probabilities that resemble current flood protection levels in the Netherlands and higher flood probabilities that resemble changes that may occur due to future climate variability. A preference for the purchase of flood insurance, ceteris paribus, is found to be positively associated with flood probability. However, this relationship is nonlinear and concave, which means that the relative change in the proportion of individuals predicted to purchase flood insurance is more pronounced for changes to very low flood probabilities (e.g., from a 1/1000 to a 2/1000 yearly chance of flooding) than for changes to higher flood probabilities (e.g., from a 9/1000 to a 10/1000 yearly chance of flooding). This suggests that the demand for insurance against disaster risk may eventually plateau for higher loss probabilities caused by climate change. A similar effect is noted by Robinson and Botzen (2020), who show that the relative risk preferences of individuals who face flood risk change more prominently for variations in low-risk levels compared to more moderate levels of risk.

The aforementioned studies on the impact of higher objective loss probabilities, however, do not account for changes in risk perceptions due to natural disasters becoming more prevalent in practice because of climate change. That is, individual disaster risk judgments and perceptions of climate risk may be expected to change in the future as natural disasters become more commonplace. Risk perceptions and beliefs about climate change are often strongly correlated with insurance demand against disaster risk (Hossain et al. 2022; Botzen and van den Bergh 2012b; Akter et al. 2017); therefore, there may be an elevated response of insurance demand to higher actual occurrences of natural disasters if (climate) risk perceptions are influenced by these higher occurrences, an effect that is predicted by the availability heuristic.

Additional insights into the possible effects of climate change on the purchase of insurance against natural disasters can be found in simulation studies that estimate future insurance premiums under various scenarios of climate change. For instance, the partial equilibrium model of Tesselaar et al. (2020a, b) finds that the uptake of flood insurance in Europe may decline as a result of climate change. This finding can be explained by higher premium loading caused by an increase in capital market volatility for private reinsurers since reinsurers often cover extreme losses, which ultimately leads to rising unaffordability of insurance for low-income households. In sum, even though demand for insurance against climate risk may rise over time, the actual uptake of insurance may fall unless measures are in place to support households with low levels of income (e.g., public reinsurance; see Sect. 3). It is worth noting that even though flood insurance uptake may fall over time due to higher premium loading, climate change information provision may increase as natural disaster events become more frequent. Therefore, by holding prices constant, the uptake of insurance could rise.

5 Natural Disaster Insurance and Climate Change Adaptation

Despite the challenges that climate change poses for insurance markets, the insurance sector can also play a role in managing increasing climate risks through implementing strategies that stimulate risk reduction by policyholders or by facilitating the development of minimum risk reduction requirements. Insurance companies can collaborate with governments to link natural disaster insurance arrangements to build code policies that make properties more resistant to the impacts of extreme weather and to develop zoning policies that prohibit new construction in areas with high natural disaster risk, such as floodplains. In general, a strong regulatory framework, with building codes, embedded in law, that enforces standards that make new and existing structures more resilient to climate change, e.g., by requiring wet- and dry-proofing as well as the elevation of certain buildings, can contribute to lowering future vulnerability to natural disaster events (Aerts and Botzen 2011b). Moreover, insurance coverage can be connected to conditions in policy regulations that require policyholders to have certain risk reduction measures in place to be eligible for insurance coverage. Building codes, zoning regulations, and insurance policy conditions can be useful for setting minimum risk reduction requirements in areas with high natural disaster risk. Moreover, insurers can stimulate risk reduction beyond these minimum requirements by providing information and financial incentives, as discussed in the remainder of this section.

Raising individuals' general awareness of risk can encourage risk reduction among the general public. Warner et al. (2009) proposed that insurers and the government work together to raise awareness among the general public by providing information about climate risks. One example is the open-access HORA risk zoning platform in Austria, which represents potential hazards due to floods, earthquakes, and storms, to the general public with the aim of increasing their awareness of risk.[7] The insurance industry contributes modeling expertise, in order to aid such risk mapping. In addition to offering risk information, insurers can advise their policyholders on effective risk reduction strategies and how to implement these measures.

Insurers may also provide financial incentives to their policyholders for reducing risks associated with climate change. For example, insurers can offer discounted premiums for individuals who are able to demonstrate that their property has installed fixtures that would reduce possible loss amounts in case of a disaster (Ward et al. 2008). Experimental evidence by Mol et al. (2020a, b) and DIFI model simulations by Tesselaar et al. (2020b) illustrate that offering policyholders discounts on flood insurance premiums is effective in stimulating the flood-proofing of homes, thereby limiting increases in flood risk under climate change. Offering premium discounts to individuals who take risk reduction measures encourages the charging of risk-based policy premiums. In cases where premiums do not adequately signal risk and/or zoning policies are insufficient, adverse effects may occur, such as new developments in disaster-prone areas (de Ruig et al. 2022). Another financial incentive for risk reduction that can be provided by insurance is the deductible. The deductible refers to the part of the damage that the policyholder pays out of pocket before reimbursement from insurance starts. This gives policyholders "skin in the game" and an incentive to minimize natural disaster losses. However, the limited existing empirical evidence shows that deductibles play only a small role in the implementation of natural disaster risk reduction measures (Hudson et al. 2017; Mol et al. 2020b). In summary, this evidence suggests that premium discounts may be more effective in stimulating adaptation to climate change risks.

[7] As an aside, Chaudhry et al. (2020) and Keller et al. (2006) found that climate risk awareness and insurance demand is higher in situations whereby probabilities are framed over a long time horizon, such as framing a 1/100 yearly probability of natural disaster occurrence as a 1/3 probability of experiencing a disaster at least once over a 40 year time horizon.

If insurers offer premium discounts for implementing risk reduction measures, long-term loans from the government, banks, or other financial institutions to homeowners for mitigation could encourage individuals to invest in cost-effective measures that promote risk reduction (Kleindorfer and Kunreuther 1999). In Connecticut in the United States, a state-run program has been implemented (Shore Up CT) to aid home and business owners to invest in flood- and wind-proofing measures in hazard-prone areas, with loans offered for a term of 15 years at an annual interest rate of 2¾ percent (Kunreuther 2015).

An obstacle for stimulating risk reduction through insurance may arise when policyholders are not able to afford the insurance premium and hence are not covered by insurance. It has been proposed that issues of affordability could be addressed while stimulating risk reduction by offering means-tested vouchers that partly cover the insurance premium in cases where low-income households residing in hazard-prone areas would otherwise face high risk-based policy premiums (Dixon et al. 2017). One prerequisite for obtaining such a voucher may be that homeowners take other climate risk adaptation measures (Kunreuther 2021).

Overall, there likely does not exist a one-size-fits-all solution with regard to insurance arrangements and other policies, as elaborated in this section, that stimulate climate change adaptation among individuals. Policy design should be judged on a case-by-case basis accounting for heterogeneity in household income, objective risk and willingness-to-purchase voluntary insurance and other climate risk reduction measures.

6 Conclusion

Climate change is expected to increase the frequency and/or severity of various extreme weather events around the world. This raises monetary losses from natural disasters as well as risks for the insurance sector insofar as these losses are covered by insurance arrangements for damages caused by extreme weather events. Increasing natural disaster risks implies that insurers should proactively adapt to climate change, for instance, by adjusting premiums, financial reserves, and reinsurance coverage to cope with higher insurance claims, and facilitate the development of minimum risk reduction requirements like building code policies. Moreover, increasing climate risk may imply reforms of natural disaster insurance arrangements. For instance, insurers stimulating policyholders to undertake risk reduction measures, e.g., by offering discounted premiums for risk reduction, may encourage homeowners to adapt to climate risk. Furthermore, introducing public–private insurance arrangements when private insurance coverage against extreme and correlated natural disaster risks becomes relatively expensive for low-income households in high-risk areas could overcome potential affordability concerns.

The basis of such adaptation strategies should be a profound understanding of the current and future development of natural disaster risks. For this purpose, catastrophe models are becoming more sophisticated for assessing local natural disaster risks under climate change scenarios. Such models have been integrated into insurance models to assess the implications of changing risks for insurance market outcomes, such as premiums and their affordability. Recent studies have integrated adaptation decisions by governments and households in these catastrophe models by developing ABMs in order to account for the risk reduction measures that are generally taken in response to changing risk, the occurrence of disasters, and decisions by other agents, among other factors. These ABMs can be an important improvement compared with standard catastrophe models that assume constant vulnerability or stylized adaptation scenarios, but a shortcoming is that the behavioral rules that guide risk reduction actions in ABMs are often designed in an ad hoc manner. Future research in this area can focus on embedding these behavioral rules in microeconomic theories and calibrating

them on the basis of surveys of actual adaptation behavior, and its determinants, and behavioral economic experiments. Further empirical research may focus on acquiring data on the evolution of behavior before and after a hazard event, which could help to improve estimates of household behavior in ABMs. It is also notable that most surveys and experiments on decision-making under climate risk have focused on single weather events. Other empirical studies may examine decisions under multihazard events due to the oftentimes high observed correlation of risk across multiple climate hazards.

References

Abdellaoui M (2000) Parameter-free elicitation of utility and probability weighting functions. Manag Sci 46(11):1497–1512

Aerts JC (2020) Integrating agent-based approaches with flood risk models: a review and perspective. Water Secur 11:1–9

Aerts JC, Botzen WW (2011a) Climate change impacts on pricing long-term flood insurance: a comprehensive study for The Netherlands. Glob Environ Chang 21(3):1045–1060

Aerts JC, Botzen WW (2011b) Flood-resilient waterfront development in New York City: bridging flood insurance, building codes, and flood zoning. Ann N Y Acad Sci 1227(1):1–82

Aerts JC, Botzen WW, Clarke KC et al. (2018) Integrating human behaviour dynamics into flood disaster risk assessment. Nat Clim Chang 8(3):193–199

Akter S, Krupnik TJ, Khanam F (2017) Climate change skepticism and index versus standard crop insurance demand in coastal Bangladesh. Reg Environ Chang 17:2455–2466

Allais M (1953) Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. Econometrica 21(4):503–546

Atreya A, Ferreira S, Michel-Kerjan E (2015) What drives households to buy flood insurance? New evidence from Georgia. Ecol Econ 117:153–161

Bajtelsmit V, Coats JC, Thistle P (2015) The effect of ambiguity on risk management choices: an experimental study. J Risk Uncertain 50(3):249–280

Botzen WW (2021) Economics of insurance against natural disaster risks. Oxford research Encyclopedia of environmental science. Oxford University Press. https://doi.org/10.1093/acrefore/9780199389414.013.712

Botzen WW, Bouwer LM (2016) Weather indicators for insured hailstorm damage to motor vehicles and potential climate change impacts. Geneva Papers Risk Insur Issues Pract 41(3):512–527

Botzen WW, van den Bergh JC (2008) Insurance against climate change and flooding in The Netherlands: present, future, and comparison with other countries. Risk Anal 28(2):413–426

Botzen WW, van den Bergh JC (2012a) Monetary valuation of insurance against flood risk under climate change. Int Econ Rev 53(3):1005–1026

Botzen WW, van den Bergh JC (2012b) Risk attitudes to low-probability climate change risks: WTP for flood insurance. J Econ Behav Organ 82(1):151–166

Botzen WW, Bouwer LM, van den Bergh JC (2010) Climate change and hailstorm damage: empirical evidence and implications for agriculture and insurance. Resour Energy Econ 32(3):341–362

Botzen WW, Kunreuther H, Michel-Kerjan E (2015) Divergence between individual perceptions and objective indicators of tail risks: evidence from floodplain residents in New York City. Judgm Decis Mak 10(4):365–385

Botzen WW, Deschenes O, Sanders M (2019) The economic impacts of natural disasters: a review of models and empirical studies. Rev Environ Econ Policy 13(2):167–188

Browne MJ, Hoyt RE (2000) The demand for flood insurance: empirical evidence. J Risk Uncertain 20:291–306

Browne MJ, Knoller C, Richter A (2015) Behavioral bias and the demand for bicycle and flood insurance. J Risk Uncertain 50(2):141–160

Charpentier A (2008) Insurability of climate risks. Geneva Papers Risk Insur Issues Pract 33(1):91–109

Chaudhry SJ, Hand M, Kunreuther H (2020) Broad bracketing for low probability events. J Risk Uncertain 61:211–244

Coronese M, Lamperti F, Keller K et al (2019) Evidence for sharp increase in the economic damages of extreme natural disasters. Proc Natl Acad Sci 116(43):21450–21455

de Moel H, Jongman B, Kreibich H et al (2015) Flood risk assessments at different spatial scales. Mitig Adapt Strateg Glob Chang 20(6):865–890

de Ruig LT, Haer T, de Moel H et al (2022) How the USA can benefit from risk-based premiums combined with flood protection. Nat Clim Chang 12:995–998

Dixon L, Clancy N, Miller BM et al (2017) The cost and affordability of flood insurance in New York City. RAND Corporation, Santa Monica
Dlugolecki AF (2008) Climate change and the insurance sector. Geneva Papers Risk Insur Issues Pract 33(1):71–90
Egli F, Schärer D, Steffen B (2022) Determinants of fossil fuel divestment in European pension funds. Ecol Econ. https://doi.org/10.1016/j.ecolecon.2021.107237
Eisner R, Strotz RH (1961) Flight insurance and the theory of choice. J Polit Econ 69(4):355–368
Etchart-Vincent N (2004) Is probability weighting sensitive to the magnitude of consequences? An experimental investigation on losses. J Risk Uncertain 28(3):217–235
Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114(3):817–868
Fehr-Duda H, Fehr E (2016) Sustainability: game human nature. Nature 530(7591):413–415
Friedl A, Lima de Miranda K, Schmidt U (2014) Insurance demand and social comparison: an experimental analysis. J Risk Uncertain 48(2):97–109
Gallagher J (2014) Learning about an infrequent event: evidence from flood insurance take-up in the United States. Am Econ J Appl Econ 6(3):206–233
Gigerenzer G (1991) How to make cognitive illusions disappear: beyond "heuristics and biases". Eur Rev Soc Psychol 2(1):83–115
Golnaraghi M (2018) Climate change and the insurance industry: taking action as risk managers and investors. The Geneva Association
Grossi P, Kunreuther H (2005) Catastrophe modeling: a new approach of managing risk. Springer, New York
Haer T, Botzen WW, Aerts JC (2019) Advancing disaster policies by integrating dynamic adaptive behaviour in risk assessments using an agent-based modelling approach. Environ Res Lett 14(4). https://doi.org/10.1088/1748-9326/ab0770
Hajimoladarvish N (2017) Very low probabilities in the loss domain. Geneva Risk Insur Rev 42:41–58
Hoeppe P (2016) Trends in weather related disasters–consequences for insurers and society. Weather Clim Extrem 11:70–79
Hossain MS, Alam GM, Fahad S et al (2022) Smallholder farmers' willingness to pay for flood insurance as climate change adaptation strategy in northern Bangladesh. J Clean Prod 338. https://doi.org/10.1016/j.jclepro.2022.130584
Hudson P, Botzen WW, Feyen L et al (2016) Incentivising flood risk adaptation through risk based insurance premiums: trade-offs between affordability and risk reduction. Ecol Econ 125:1–13
Hudson P, Botzen WW, Czajkowski J et al (2017) Moral hazard in natural disaster insurance markets: empirical evidence from Germany and the United States. Land Econ 93(2):179–208
Hudson P, Botzen WW, Aerts JC (2019) Flood insurance arrangements in the European Union for future flood risk under climate and socio-economic change. Glob Environ Chang 58. https://doi.org/10.1016/j.gloenvcha.2019.101966
Hudson P, Ruiter MC, de Ruig LT et al (2020) An assessment of best practices of extreme weather insurance and directions for a more resilient society. Environ Hazards 19(3):301–321
Irwin JR, McClelland GH, Schulze WD (1992) Hypothetical and real consequences in experimental auctions for insurance against low-probability risks. J Behav Decis Mak 5(2):107–116
Kahneman D (2003) Maps of bounded rationality: psychology for behavioral economics. Am Econ Rev 93(5):1449–1475
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decisions under risk. Econometrica 47(2):263–292
Keller C, Siegrist M, Gutscher H (2006) The role of the affect and availability heuristics in risk communication. Risk Anal 26(3):631–639
Kleindorfer PR, Kunreuther H (1999) The complementary roles of mitigation and insurance in managing catastrophic risks. Risk Anal 19(4):727–738
Kousky C (2011) Understanding the demand for flood insurance. Nat Hazards Rev 12(2):96–110
Kousky C (2017) Disasters as learning experiences or disasters as policy opportunities? Examining flood insurance purchases after hurricanes. Risk Anal 37(3):517–530
Kousky C, Michel-Kerjan E, Raschky PA (2018) Does federal disaster assistance crowd out flood insurance? J Environ Econ Manag 87:150–164
Krauss AD (2021) Effect of climate change on the insurance sector. In: Letcher TM (ed) The impacts of climate change. Elsevier
Kriesel W, Landry C (2004) Participation in the National Flood Insurance Program: an empirical analysis for coastal properties. J Risk Insur 71(3):405–420
Kunreuther H (1996) Mitigating disaster losses through insurance. J Risk Uncertain 12(2):171–187
Kunreuther H (2015) The role of insurance in reducing losses from extreme events: the need for public–private partnerships. Geneva Papers Risk Insur Issues Pract 40:741–762
Kunreuther H (2021) Improving the national flood insurance program. Behav Public Policy 5(3):318–332
Kunreuther H, Michel-Kerjan E (2015) Demand for fixed-price multi-year contracts: experimental evidence from insurance decisions. J Risk Uncertain 51:171–194

Kunreuther H, Pauly M (2004) Neglecting disaster: why don't people insure against large losses? J Risk Uncertain 28(1):5–21
Kunreuther H, Meszaros J, Hogarth RM et al (1995) Ambiguity and underwriter decision processes. J Econ Behav Organ 26(3):337–352
Kunreuther H, Novemsky N, Kahneman D (2001) Making low probabilities useful. J Risk Uncertain 23:103–120
Kunreuther H, Michel-Kerjan E, Ranger N (2013) Insuring future climate catastrophes. Clim Chang 118:339–354
Kunreuther H, Conell-Price L, Kovacs P et al (2021) The impact of a government risk pool and an opt-out framing on demand for earthquake protection. National Bureau of Economic Research. https://doi.org/10.3386/w29144
Kusev P, van Schaik P, Ayton P et al (2009) Exaggerated risk: Prospect theory and probability weighting in risky choice. J Exp Psychol Learn Mem Cogn 35(6):1487–1505
Landry CE, Turner D, Petrolia D (2021) Flood insurance market penetration and expectations of disaster assistance. Environ Resour Econ 79(2):357–386
Laury SK, McInnes MM, Swarthout JT (2009) Insurance decisions for low-probability losses. J Risk Uncertain 39: 17–44
Lo AY (2013) The role of social norms in climate adaptation: mediating risk perception and flood insurance purchase. Glob Environ Chang 23(5):1249–1257
Löfgren Å, Martinsson P, Hennlock M et al (2012) Are experienced people affected by a pre-set default option—results from a field experiment. J Environ Econ Manag 63(1):66–72
Loncy D (2023) Why climate change is making homeowner's insurance more expensive. In: Wharton Business Daily. Available via: https://knowledge.wharton.upenn.edu/podcast/wharton-business-daily-podcast/why-climate-change-is-making-homeowners-insurance-more-expensive/
Markhvida M, Walsh B, Hallegatte S et al (2020) Quantification of disaster impacts through household well-being losses. Nat Sustain 3(7):538–547
McClelland GH, Schulze WD, Coursey DL (1993) Insurance for low-probability hazards: a bimodal response to unlikely events. J Risk Uncertain 7:95–116
Meyer R, Kunreuther H (2017) The ostrich paradox: why we underprepare for disasters. University of Pennsylvania Press
Michel-Kerjan E, Kousky C (2010) Come rain or shine: evidence on flood insurance purchases in Florida. J Risk Insur 77(2):369–397
Mills E (2012) The greening of insurance. Science 338:1424–1425
Mol JM, Botzen WW, Blasch JE (2020a) Behavioral motivations for self-insurance under different disaster risk insurance schemes. J Econ Behav Organ 180:967–991
Mol JM, Botzen WW, Blasch J (2020b) Risk reduction in natural disaster insurance: experimental evidence on moral hazard and financial incentives. J Behav Exp Econ 84. https://doi.org/10.1016/j.socec.2019.101500
Mol JM, Botzen WW, Blasch JE et al (2021) All by myself? Testing descriptive social norm-nudges to increase flood preparedness among homeowners. Behav Public Policy. https://doi.org/10.1017/bpp.2021.17
National Association of Insurance Commissioners (NAIC) (2023) Parametric disaster insurance. Available via: https://content.naic.org/cipr-topics/parametric-disaster-insurance#:~:text=The%20term%20parametric%20insurance%20describes,in%20a%20traditional%20indemnity%20policy
Otto FE, Philip S, Kew S et al (2018) Attributing high-impact extreme events across timescales—a case study of four different types of events. Clim Chang 149(3):399–412
Pachur T, Suter RS, Hertwig R (2017) How the twain can meet: Prospect theory and models of heuristics in risky choice. Cogn Psychol 93:44–73
Paudel Y, Botzen WW, Aerts JC (2012) A comparative study of public-private catastrophe insurance systems: lessons from current practices. Geneva Papers Risk Insur Issues Pract 37:257–285
Petrolia DR, Landry CE, Coble KH (2013) Risk preferences, risk perceptions, and flood insurance. Land Econ 89(2):227–245
Petrolia DR, Hwang J, Landry CE et al (2015) Wind insurance and mitigation in the coastal zone. Land Econ 91(2):272–295
Poussin J, Botzen WW, Aerts JC (2013) Stimulating flood damage mitigation through insurance: an assessment of the French CatNat system. Environ Hazards 12(3–4):258–277
Raschky PA, Weck-Hannemann H (2007) Charity hazard—a real hazard to natural disaster insurance? Environ Hazards 7(4):321–329
Richter A, Schiller J, Schlesinger H (2014) Behavioral insurance: theory and experiments. J Risk Uncertain 48:85–96
Robinson PJ, Botzen WW (2018) The impact of regret and worry on the threshold level of concern for flood insurance demand: evidence from Dutch homeowners. Judgm Decis Mak 13(3):237–245
Robinson PJ, Botzen WW (2019) Economic experiments, hypothetical surveys and market data studies of insurance demand against low-probability/high-impact risks: a systematic review of designs, theoretical insights and determinants of demand. J Econ Surv 33(5):1493–1530

Robinson PJ, Botzen WW (2020) Flood insurance demand and probability weighting: the influences of regret, worry, locus of control and the threshold of concern heuristic. Water Resour Econ 30. https://doi.org/10.1016/j.wre.2019.100144
Robinson PJ, Botzen WW (2022) Setting descriptive norm nudges to promote demand for insurance against increasing climate change risk. Geneva Papers Risk Insur Issues Pract 47:27–49
Robinson PJ, Botzen WW, Kunreuther H et al (2021a) Default options and insurance demand. J Econ Behav Organ 183:39–56
Robinson PJ, Botzen WW, Zhou F (2021b) An experimental study of charity hazard: the effect of risky and ambiguous government compensation on flood insurance demand. J Risk Uncertain 63(3):275–318
Schoemaker PJ, Kunreuther H (1979) An experimental study of insurance decisions. J Risk Insur 46(4):603–618
Schwarze R, Schwindt M, Weck-Hannemann H et al (2011) Natural hazard insurance in Europe: Tailored responses to climate change are needed. Environ Policy Gov 21:14–30
Slovic P, Fischhoff B, Lichtenstein S et al (1977) Preference for insuring against probable small losses: insurance implications. J Risk Insur 44(2):237–258
Swiss Re (2017) Natural catastrophes and man-made disasters in 2016: a year of widespread damages. Available via: https://www.swissre.com/institute/research/sigma-research/sigma-2017-02/natcat-and-manmade-disasters-2016.html
Swiss Re (2020) Natural catastrophes in times of economic accumulation and climate change. Sigma, No 2/2020
Tesselaar M, Botzen WW, Aerts JC (2020a) Impacts of climate change and remote natural catastrophes on EU flood insurance markets: an analysis of soft and hard reinsurance markets for flood coverage. Atmosphere 11(2). https://doi.org/10.3390/atmos11020146
Tesselaar M, Botzen WW, Haer T et al (2020b) Regional inequalities in flood insurance affordability and uptake under climate change. Sustain For 12(20). https://doi.org/10.3390/su12208734
Tesselaar M, Botzen WW, Robinson PJ et al (2022) Charity hazard and the flood insurance protection gap: an EU scale assessment under climate change. Ecol Econ 193. https://doi.org/10.1016/j.ecolecon.2021.107289
Tucker M (1997) Climate change and the insurance industry: the cost of increased risk and the impetus for action. Ecol Econ 22(2):85–96
Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. Science 185(4157):1124–1131
Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain 5(4):297–323
van Aalst MK (2006) The impacts of climate change on the risk of natural disasters. Disasters 30(1):5–18
von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
Ward RE, Herweijer C, Patmore N et al (2008) The role of insurers in promoting adaptation to the impacts of climate change. Geneva Papers Risk Insur Issues Pract 33(1):133–139
Warner K, Ranger N, Surminski S et al (2009) Adaptation to climate change: linking disaster risk reduction and insurance. United Nations International Strategy for Disaster Reduction, Geneva
Zahran S, Weiler S, Brody SD et al (2009) Modeling national flood insurance policy holding at the county scale in Florida, 1999–2005. Ecol Econ 68(10):2627–2636
Zhang J, Czajkowski J, Botzen WW et al (2022) Assessing the drivers of intrinsically complex hurricane insurance purchases: lessons learned from survey data in Florida. In: Collins JM, Done JM (eds) Hurricane risk in a changing climate. Springer, Cham
Zhou F, Endendijk T, Botzen WW (2023) A review of the financial sector impacts of risks associated with climate change. Ann Rev Resour Econ. (In press)

Household Financial Resilience After Severe Climate Events: The Role of Insurance

Benjamin Collier and Carolyn Kousky

1 Introduction

Severe climate events and other natural disasters impose substantial costs globally. Aon estimates that in 2022, the costs of natural disasters totaled $313 billion around the world, with roughly 42% covered by insurance (Aon 2023). In the United States, the National Oceanic and Atmospheric Administration estimates that between 2012 and 2021, weather-related disasters cost the U.S.A. at least $1 trillion in damages (NOAA 2022). These losses are only growing; this is due to many reasons, including continued development in risky areas, climate change, urbanization, and, recently, inflation. Swiss Re Institute (2023) estimates that global insured natural disaster losses have been growing at around 5–7% per year since 1992.

As risks escalate, more households are facing negative financial shocks caused by disasters. In 2021, one in six adults in the United States was directly affected by a natural disaster (Board of Governors 2022). Disasters damage property, including homes, their contents, and vehicles. They can also impose a range of non-property costs, such as evacuation expenses, income losses, additional living expenses if the home is rendered uninhabitable, debris clean-up costs, higher commuting costs when transportation networks are disrupted or vehicles damaged, and the costs of generators and fuel when power is disrupted (You and Kousky 2024). Severely affected households need to solve a financing challenge: how to fund the tens or even hundreds of thousands of dollars of repairs and other expenses that a disaster leaves behind.

In this chapter, we review the existing literature to examine how households financially manage the economic impacts of large and increasing climate-related extreme events. Following existing research, our review primarily focuses on households in the United States. Climate extremes, of course, are a global problem, and the greatest vulnerabilities exist in low-income and emerging markets.

B. Collier (✉)
Fox School of Business, Temple University, Philadelphia, PA, USA
e-mail: collier@temple.edu

C. Kousky
Environmental Defense Fund, New York, NY, USA
e-mail: ckousky@edf.org

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_4

We begin the chapter by reviewing the research on the role of insurance in disaster recovery. Insurance is one of several risk management tools that households might employ to manage their financial risks, and it is especially effective for low-frequency, high-severity events such as disasters. It is intuitive that by providing funds to offset losses, insurance should improve recovery. And, indeed, the literature documents that insured households are more likely to experience positive outcomes (e.g., shorter rebuilding times) and less likely to experience negative outcomes (e.g., bankruptcy) following a disaster than uninsured households. However, the causal effect of insurance on recovery is difficult to identify in practice because households who are better-off along a number of dimensions, such as income and education, are more likely to be insured.

Despite theoretical and (suggestive) empirical evidence regarding the benefits of disaster insurance, the majority of catastrophe losses are uninsured. This is true globally and in the United States. We turn to discussing this protection gap and its drivers in Sect. 3. Insurance supply frictions, including the challenges of financing and modeling severe climate events, contribute to the catastrophe insurance gap. Demand frictions also play a role: Growing evidence shows that many consumers' willingness to pay for flood insurance is below their expected loss, likely driven by multiple factors including risk perceptions and ability to pay.

With many uninsured against disasters, this chapter then documents how households actually fund recovery from severe climate events, drawing on both qualitative and quantitative studies. For major hurricanes, households most frequently report using a combination of homeowners insurance, savings, and friends and family support, among other funding sources. We also review recent research on federal grant and loan programs and the populations that they serve. Some households are able to fund recovery, often through a patchwork of funding sources; however, others cannot and continue to struggle in the years ahead. Through surveys, households report that the disaster worsened their financial situation, and consumer credit reports show that disasters can materially increase severe financial distress, including mortgage foreclosure and bankruptcy.

Finally, we discuss how household vulnerability, funding sources, and recovery vary by income and race. Low-income households and racial and ethnic minorities are typically more vulnerable to climate shocks, driven by differences in their exposure to hazards through where they live and the nature of the building stock, their access to needed preparedness and emergency response resources, and finally, their disproportionately lower ability to access the funds needed for recovery. These households are less likely to be insured, receive a federal grant or loan, have significant savings, or be approved for a private loan. These findings illustrate a growing conclusion in the literature that severe climate events exacerbate inequality and racial disparities.

This chapter provides a holistic perspective of disaster recovery funding, including by showing that income and race are central to households' recovery. We conclude with a discussion of the policy implications of this research.

2 The Role of Insurance in Recovery

2.1 Risk Management and Insurance

A core goal of risk management is minimizing the total cost of risk, which often entails combining financial and physical tools based on their varying comparative advantages. The set of tools includes: financial risk transfer (e.g., flood insurance), avoidance (e.g., choosing to locate outside of the most flood-prone areas), reduction (e.g., elevating a home in a flood-prone area), and retention (e.g., holding savings to help fund losses after a flood). These tools are partial substitutes. For example, if prospective home buyers learn that the cost of flood insurance for a property is high, they may

decide to avoid living in that risky area. In particular, the frequency and severity of losses and the cost of specific tools determine which strategies are most effective for any particular context (Courbage et al. 2013). This may require using multiple tools. For example, even after investing in risk reduction measures, it may be impossible to eliminate risk in hazard-prone areas, and, as such, households must also determine how much risk to transfer and how much to retain.

Insurance is typically the most efficient risk management tool available to households for addressing the risk of *infrequent*, *severe* events.[1] Given limited finances, most households will not wish to retain much of the risk of severe damages. Indeed, without insurance, households may have limited options to fund recovery and may resort to coping strategies that delay recovery or increase the cost of the severe event.

2.2 Insurance and Recovery: Empirical Evidence

A small body of empirical research speaks to the role of insurance in disaster recovery. Scholars agree that disaster recovery is a complex process, and definitions of recovery, as well as the metrics to measure it, vary in the literature. One useful benchmark for recovery is whether households have returned to pre-disaster conditions.

That insurance would facilitate recovery is intuitive as it transfers funds to adversely affected populations in a time of need. However, isolating the effects of insurance empirically is challenging because populations that purchase insurance differ from those who do not. Better-off populations are more likely to be insured. Among households, disaster insurance coverage is highly correlated with factors such as income, education, race, and financial literacy (Armantier et al. 2023; Bolin and Bolton 1986; Peacock and Girard 1997; You and Kousky 2024). Insurance take-up is also correlated with risk and may correlate with the use of other forms of risk management. For example, in the U.S.A., many households in flood-prone areas are required to purchase flood insurance, and houses built in these areas must adhere to certain building requirements that reduce flood risk (e.g., elevating the property). Among countries, insurance coverage is also highly correlated with economic development: Countries with more insurance tend to have higher incomes, more developed public infrastructure, larger financial services sectors, larger public safety nets, and more stringent building codes (Cole and Xiong 2017; Outreville 2013; Chang and Lee 2012). A lack of available data can also exacerbate the challenge of identifying the role of insurance in recovery as insurers are generally unwilling to make policy-level information available to researchers. This limits the markets and lines of insurance that are researched, and also, without granular data on private insurance markets, studies can be prone to aggregation bias.

Despite these challenges, research using a variety of methods and datasets finds that insurance is associated with positive recovery outcomes. In a review of the literature looking at studies from the level of the household to the level of the country, Kousky (2019) finds that while quantitative, empirical studies are limited, the existing work suggested that insurance improves recovery by increasing the likelihood of rebuilding and minimizing post-disaster financial hardship. The studies also present varying degrees of evidence that lower-income or minority populations are less likely to be protected by insurance and, as a result, may have worse recovery outcomes (discussed further in Sect. 5).

Several papers assess the role of insurance at the country level. These papers regress economic output (e.g., gross domestic product) on measures of the impact of a disaster using country-year-level observations. They find that larger uninsured losses are associated with lower economic output

[1] As a contrasting example, avoidance may be most efficient for high-frequency, high-severity risks.

after the disaster, but insured losses have little impact on macroeconomic indicators (e.g., Von Peter et al. 2012; Melecky and Raddatz 2015). Nguyen and Noy (2020) examine economic recovery across regions of New Zealand following the Canterbury Earthquake Sequence. Using nighttime satellite light intensity data as an indicator of economic activity, they find that receipt of insurance payouts increases economic recovery following the event.

Household surveys and case studies of disaster recovery echo the country-level findings (Peacock and Girard 1997; Olshansky et al. 2006). Turnham et al. (2011) survey households impacted by the 2005 hurricanes and find that the most common reason for not rebuilding post-storm was lack of funding. If a property had been insured, it was 37% more likely to have been rebuilt. You and Kousky (2024) draw on a survey of those impacted by one of four U.S. hurricanes (Harvey, Florence, Ida, and Michael). After controlling for household and storm characteristics, as well as the range of storm-imposed costs, they find that those with insurance are over 80% more likely to report lower financial burdens after the storm in both the short and long run and are almost 60% less likely to have unmet funding needs. A telephone survey in Germany following a large flood found those with insurance received more total compensation than those who relied only on the government and were more likely to report being "satisfied" with the post-flood funding they received (Thieken et al. 2006). A deep sociological study of one community that was heavily flooded by Hurricane Harvey finds that lack of flood insurance drove increases in inequality: Households with flood insurance were able to rebuild and often even do better after the disaster, while those without flood insurance spiraled into deeper financial precarity (Rhodes and Besbris 2022).

Indirect evidence also speaks to the role of insurance in recovery, including specific channels by which insurance generates post-disaster financial outcomes. Gallagher and Hartley (2017) find that after Katrina, households who were flooded tended to have lower mortgage debt, which the authors infer is due to households using flood insurance payments to settle their mortgages. Similarly, examinations of credit performance after Hurricane Harvey suggest that flood insurance helped mitigate negative financial impacts such as bankruptcy (Billings et al. 2022) and that households in areas with a higher degree of insurance take-up borrowed less post-flood (del Valle et al. 2022). After Hurricane Harvey, a loan-level analysis found that flood insurance provided financial protection against loan modifications, delinquency, and default (Kousky et al. 2020). Finally, an earlier analysis of properties damaged by Hurricane Sandy found greater investments in damaged properties inside the Federal Emergency Management Agency (FEMA)-designated 100-year floodplain, which the authors attribute to greater flood insurance take-up (McCoy and Zhao 2018).

3 The Catastrophe Insurance Gap

Despite the benefits that insurance appears to provide, many households are uninsured or underinsured against disaster risks. From 2013 to 2022, an estimated 69% of natural catastrophe losses were uninsured globally (Swiss Re Institute 2023). The catastrophe insurance gap varies by region: 87% of losses were uninsured in the Asia-Pacific region versus 50% of losses in the United States. It also varies within countries and by peril: Take-up rates for earthquake insurance in California, for example, hover at a bit more than 10% (Marshall et al. 2018), and on average nationwide, only 30% of those in the FEMA-mapped 100-year floodplain have a flood insurance policy (Kousky 2018). The disaster insurance gap is smaller where there is a government program or mandate for disaster coverage to be included in standard property policies. For example, in countries like France and Spain, where full coverage against natural hazards is included in standard property policies, coverage is quite high (OECD 2021). In countries where disasters are excluded from standard property policies,

however, and the market depends on private insurers supplying and consumers demanding catastrophe insurance, the gap is much larger.

Both supply and demand constraints contribute to this insurance gap, which we briefly summarize here. Our summary focuses on constraints that we view as central to the themes of this chapter and is not comprehensive. Marcoux and Wagner (2023) provide a more extensive treatment of this topic for natural disaster insurance markets in the United States, and Botzen and Robinson (2023) discuss evolving insurance supply and demand constraints in the context of climate change.

3.1 Supply Constraints

Supply of disaster insurance in the private market faces several challenges. The first is the challenge of financing catastrophic losses. The spatially correlated and potentially very severe nature of catastrophe insurance claims means that the losses insurers face from disasters can fluctuate widely from year to year, with low-loss years and other severe loss years. Insuring disasters requires solving an inter-temporal smoothing problem to ensure sufficient claims-paying ability: Insurers must hold more capital and make greater use of reinsurance and other risk transfer mechanisms for disaster lines (Jaffee and Russell 1997). These financing costs contribute to higher loads on catastrophe insurance compared to non-catastrophe insurance.

The challenge of modeling catastrophe risks can also contribute to supply constraints. Insurers charge higher premiums when estimates of the risk are imprecise or vary substantially based on modeling assumptions (Hogarth and Kunreuther 1992). Because disasters are rare and severe, risk modeling estimates can be imprecise, resulting in relatively high pricing loads due to uncertainty or ambiguity. For example, Kunreuther et al. (1995) survey insurance underwriters who report that they would price earthquake insurance at 2–3 times the expected loss due to modeling risk.

Innovation in recent years has reduced the challenges of financing and modeling catastrophe risk. Regarding risk financing, integration of risk and capital markets, such as through catastrophe bonds and other insurance-linked securities, has created new opportunities to finance tail risks (Cummins and Barrieu 2013). Models of catastrophe risk also continue to improve, especially regarding flood risk. For example, for decades, flood models were inaccurate and at coarse resolutions. Now multiple vendors model pluvial, fluvial, and coastal flood (Golnaraghi et al. 2018). Also, open-source flood models are now available from nonprofits and researchers, such as the First Street Foundation and Fathom. Model improvements vary by peril, and some, such as wildfire, are still highly uncertain with different models producing expected losses many times different from each other (Evans and Xu 2019). Still, innovations in risk financing and modeling should help expand the supply of catastrophe insurance and help reduce the size of catastrophe loads on premiums.

Government intervention can pose additional constraints on catastrophe insurance supply. Regulatory pricing frictions will affect whether insurers decide to participate in a catastrophe insurance market. In the U.S.A., state-level regulations that limit an insurer's ability to raise its prices will affect insurers' decisions to offer coverage (Born and Klimaszewski-Blettner 2013). Also, government provision of insurance, while sometimes provided in response to reluctance from private insurers to offer coverage, could also crowd out the private sector, particularly if prices are substantially subsidized. For instance, in the U.S.A., flood insurance has been provided federally through the U.S. National Flood Insurance Program (NFIP) for more than five decades, following lack of private options at the time of its founding. Once in existence, though, such programs may constrain market development even as financing and modeling evolve. Some approaches to public–private partnership for managing flood risk provide a larger role for the private sector (e.g., in designing, pricing, and distributing products). In the UK, for example, private insurers provide flood coverage to lower-risk

properties, and cede high-risk properties to the government's FloodRe program, which is funded by assessments on all insurers (Surminski 2018). Compared with fully public insurers, such approaches may be more effective at leveraging market innovations that can improve insurance supply over time (Cummins and Mahul 2009).

3.2 Demand Constraints

Demand constraints are a sobering challenge in this context. While markets and public policies have evolved to improve the quantity and price of insurance supplied, research consistently documents low consumer demand for coverage. Flooding is the most common and damaging natural hazard in the United States and around the world (Swiss Re Institute 2023). As a result, low demand for flood insurance is especially concerning, making flood coverage perhaps the most studied form of disaster insurance. Flood insurance demand has posed a particularly long-standing puzzle because classical economic theory predicts risk-averse consumers would want to insure catastrophe risks (e.g., Kunreuther and Slovic 1978; Browne and Hoyt 2000; Kriesel and Landry 2004; Kousky 2011; Botzen and van den Bergh 2012; Browne et al. 2015; Wagner 2022). In the U.S.A., securing coverage from flooding requires purchasing a separate flood policy, available from the NFIP. The NFIP was established in 1968 in response to lack of private options, and academics almost immediately noted that take-up was much lower than standard economic models would predict (Anderson 1974; Kunreuther 1974). In response, Congress adopted a requirement that flood insurance be purchased for any federal assistance given for property in the FEMA-mapped 100-year floodplain; this includes federally backed loans and loans from federally regulated lenders. Despite this requirement and disclosures about flood risk in this area, FEMA estimates that on average nationwide roughly a third of households in the 100-year floodplain have flood insurance, while less than 2% of those outside this area has flood insurance (FEMA 2018).

Increasing evidence shows that consumers have a willingness to pay for flood insurance that is well below the expected value of insurance contracts. Again, looking at flood insurance in the U.S.A., many do not insure, even at rates that are less than the full risk cost (Wagner 2022). Similarly, a stated preference study of residents of a flood-prone neighborhood in Portland, Oregon, finds that respondents were only willing to pay 47% to 59% of the area's median flood insurance premium (Netusil et al. 2021).

A common explanation for this low willingness to pay is that households underestimate the risk (e.g., Browne et al. 2015; Bakkensen and Barrage 2021), yet demand remains low even when households are informed that their flood insurance premiums are priced below expected cost. Collier et al. (2023b) examine a high-risk group of policyholders in the NFIP, called "severe repetitive loss properties." In a letter, the program informed these policyholders that their premiums were increasing, but also that they remain subsidized and their premiums would never exceed their expected cost. Roughly, one-fourth of these policyholders stopped purchasing flood insurance in response to the premium increase. This behavior appears to align with research from behavioral economics and psychology: Even when individuals have access to needed information, many people are systematically biased in their risk-based decisions, such as being overly optimistic, which could lead them to forgo disaster insurance or other protective measures (Meyer and Kunreuther 2017).

Another common explanation for observed low demand is concerned about "affordability." We think of affordability as encompassing two (related) concepts. The first is a more stringent concern regarding binding ex ante liquidity constraints, which may reduce households' ability to insure (Armantier et al. 2023; Ericson and Sydnor 2018; Kunreuther et al. 2006). Simply put, households will choose rent and food over insurance if they cannot have both. Standard economic models show

that liquidity-constrained consumers should partially insure (e.g., lower their coverage limit or select a high deductible); however, in practice, households purchasing flood insurance typically fully insure their homes and choose a low deductible, even if they are offered high prices (Collier et al. 2022). Households are more likely to forgo their insurance altogether in response to a price increase instead of adjusting their coverage limits or deductibles (Wagner 2022; Collier et al. 2023b). A second view of affordability pertains to the opportunity cost of insuring. Insurance is a form of state-contingent savings—policyholders pay premiums today to cover a potential future loss—and savings transfer cash flows in the opposite direction of what low-income households desire (Rampini and Viswanathan 2019). These households want to shift consumption backward through borrowing to invest today in education and home ownership, which increase lifetime wealth; insurance premiums have a high opportunity cost because they compete with these investments. These theoretical arguments can help explain the observed, strong positive correlation between income and insurance and are consistent with field experiments examining how the timing of insurance premiums affects take-up among low-income, agricultural households in Kenya, as well (Casaburi and Willis 2018).

Concerns have also been raised that generous federal disaster aid could create a disincentive for households to insure, commonly referred to as "charity hazard." Findings on this effect vary by country, depending on the particular disaster aid programs (Raschky et al. 2013). In the United States, evidence has been limited, perhaps because the amount of grant assistance given to households is quite modest. As we discuss further below, households typically receive only a few thousand dollars, and only for federally declared disasters that activate the household assistance program in FEMA. In addition, households are required to purchase and maintain insurance as a condition of receiving federal aid. An empirical study of the flood insurance program in the U.S.A. finds that receipt of FEMA assistance reduces the amount of insurance purchased on average by 3% and has no impact on take-up rates (Kousky et al. 2018). Another study of disaster insurance in the U.S.A. and Germany finds no evidence of moral hazard (Hudson et al. 2017), although another German study does find evidence of this in flood-prone regions, attributing it to the German policy of providing post-flood aid to uninsured households (Andor et al. 2020).

Insuring requirements and norms also contribute to insurance take-up. For example, homeowners insurance take-up rates are much higher; around 94% of homeowners purchase homeowners insurance (Jeziorski et al. 2021). One explanation is that mortgage lenders require homeowners insurance for *all* borrowers, making it easier to identify and address noncompliance than for flood insurance since only a subset of borrowers must insure flood. Norms may also contribute to take-up: While they are not required to insure, around 90% of households *without* a mortgage also purchase homeowners insurance. Consumers are more likely purchase insurance if their friends, neighbors, or others in their social network do (Hu 2022; Meyer and Kunreuther 2017). That said, homeowners insurers have recently been increasing premiums in risk-prone areas, leading more households to forgo homeowners insurance (Dagher 2023; Sacks 2023). These dynamics may disrupt existing norms that have kept homeowners insurance take-up rates high.

4 Coping with Disasters

Given a large disaster insurance gap, how do households cope with the financial costs of a disaster? As noted in the introduction, disasters impose a range of demands on households' finances: the need to repair/replace damaged home and cars, evacuation and temporary housing costs, income disruptions, etc. These demands represent a financing problem, requiring severely affected households to organize dollar amounts of five or even six figures to fund recovery.

Understanding how households fund recovery from a disaster is challenging as they may rely on a variety of funding sources, which are not fully visible in a single dataset. Disaster recovery surveys overcome this challenge by asking households directly. For example, Hamel et al. (2018) conduct focus groups and a survey of households who were affected by Hurricane Harvey. They report that households often use a combination of funding sources, including insurance, savings, loans, government or charitable assistance, help from friends and family, and other sources. These sources vary in the amount of funds provided, how they can be used, and in the time it takes to receive them. This variation helps explain why households may need to use multiple sources of recovery funds to complement each other in use and timing.

Table 1 summarizes the findings of You and Kousky (2024) who survey households affected by one of four major U.S. hurricanes that occurred between 2017 and 2021 (Hurricanes Harvey, Florence, Michael, and Ida). They collected data from a random sample of households in each impacted area using the survey firm Qualtrics. The most common funding tools used by households in their sample, in order of frequency, are: (1) homeowners or renters insurance, (2) household savings, (3) transfers from family and friends, (4) credit cards, and (5) FEMA grants. We use this table as a guide to discuss what is known about the use of these various funding sources in more detail.

Insurance Homeowners/renters insurance was the most frequent source of funds for covering the costs of hurricane damages, with 55% of survey respondents reporting receiving homeowners/renters insurance claims payments (Table 1). Hurricanes can cause damages from winds and floodwaters, as well as follow-on hazards, such as fires from downed power lines, tree limbs and other debris hitting homes, and backed-up sewer lines. While homeowners/renters insurance does not cover flooding, it

Table 1 Funding disaster repairs (You and Kousky 2024)

Funding source	%
Savings	51
Insurance	
Homeowners/renters insurance	55
Flood insurance	21
Credit	
Credit cards	19
Formal loan from private bank or lender	8
Federal assistance programs	
FEMA grant	19
Federal disaster loan	7
Other transfers	
Family & friends	29
Charity, nonprofit, or community group	10
Employer	8
Local government	5

Note: This table presents survey responses to the question "Which of the following sources provided funds to help pay for the costs of repairing/rebuilding your home or for the costs of replacing items inside your home? (check all that apply)." Data are from You and Kousky (2024) and reproduced with the authors' permission

typically covers wind damages and the related impacts.[2] Homeowners insurance is fairly widespread, partially because mortgage lenders require it, so its frequent use following these hurricanes may reflect a combination of high take-up and the variety of losses from a major hurricane.

Given the demand constraints described in Sect. 3.2, it is perhaps not surprising that only 21% of survey respondents received flood insurance claims payments. As discussed above, take-up for flood insurance can be low, although policies are heavily geographically concentrated in coastal communities; roughly three quarters of all policies are in only 3% of U.S. counties (Kousky et al. 2020). Several studies have examined the determinants of flood insurance demand. As would be expected, purchase of flood insurance is greater in areas with higher risk and also in areas of higher home values and higher levels of education (Kousky 2011; Landry and Jahan-Parvar 2011; Brody et al. 2017). Take-up is also greater among those who anticipate greater flood damage (Petrolia et al. 2013) and who are at higher risk: Prior to rating reform of this program in 2021, research found adverse selection in the program, with those at higher risk more likely to voluntarily insure (Bradt et al. 2021).

Insurance payments tend to be much larger than any other sources of recovery funds. For example, for Hurricane Harvey, the average NFIP claims payment in Texas was $121,000, the average federal disaster loan was $79,000, and the average FEMA grant was $7400 (Billings et al. 2022). The magnitude of claims clearly depends in part on the severity of the event. Analysis of NFIP claims between 1978 and 2012 finds that less than 1% of claims are at the maximum amount of coverage (and the overwhelming majority of those were associated with Hurricane Katrina) and that half of building claims are for less than 10% of a building's assessed value (Kousky and Michel-Kerjan 2017).

While insurance can provide much larger payments than other sources of disaster funding, there are limitations on what is covered and funds may take months to reach policyholders. For example, NFIP policies do not include costs for temporary housing for cases in which the home is too badly damaged to be occupied. NFIP policies also have very limited coverage for basements and do not always provide replacement cost coverage, which can result in claims payments that are insufficient to cover repairs. Standard homeowners policies may have sublimits for certain types of losses, like mold or hail damage to roofs, and whether or not certain items are covered, such as temporary living or sewer backups, typically depends on whether policyholders are willing to pay more to add these coverages to their policy.

Savings Over half of survey respondents reported using some of their savings to fund repairs. We might think about two forms of savings: liquid and illiquid. Liquid savings are a typical component of a prudent risk management strategy as these reserves can complement other risk financing tools. For example, insurance experts often recommend that consumers choose the highest deductible that they can afford as it is more cost-effective to manage small to moderate losses through liquid savings. In practice, though, households appear to save very little for a rainy day. Nearly a quarter of consumers have no savings to draw on for emergencies and close to 40% have less than one month of income saved for emergencies (Ratcliff et al. 2022). Thus, while households frequently turn to savings to fund hurricane repairs, their liquid savings may cover only a small share of damages.

Households might also use less liquid savings, such as their retirement savings, to fund repairs. These accounts typically receive preferential tax treatment, and consumers incur a penalty if withdrawing funds from them for other uses. Like liquid savings, households' less liquid savings may also be limited: Around a quarter of working adults have no retirement savings and among

[2] In states at risk of hurricanes, public sector programs, often referred to as beach plans or wind pools, also provide wind-only or full homeowners insurance when consumers find it difficult to find a policy in the private market.

those with retirement savings, two-thirds reported that they were not saving enough for retirement (Brevoort et al. 2021). Federal policies have facilitated using retirement savings following major hurricanes. For example, the federal government temporarily removed withdrawal penalties on tax-preferred retirement accounts for households who were affected by Hurricane Katrina. Analyzing these households' tax returns, Deryugina et al. (2018) find that many drew down their retirement savings during this time, withdrawing $2500 on average.

The implications of these retirement withdrawals for households' financial well-being are unclear. Life-cycle models, which examine how households can maximize consumption over their lifetimes, would typically prescribe smoothing consumption by drawing down these savings in response to a large wealth shock (e.g., Gourinchas and Parker 2002; Kaplan et al. 2014). These dynamics may work especially well for younger households since they have a longer horizon over which to smooth consumption, though they also have smaller retirement accounts and so may not have enough to cover losses. The consequences for older households may be direr since they have less time to adjust. To our knowledge, the literature has not examined how using retirement savings affects the long-term financial health of older populations.

Federal Household Grants Despite political rhetoric to the contrary, federal disaster grants for households are typically quite limited. Severe disasters can result in a Presidential Disaster Declaration. These declarations trigger one or both of two assistance programs housed in the FEMA: Public Assistance (PA), which provides federal funding to assist state and local governments (e.g., to clean debris or repair damaged infrastructure), and/or Individual Assistance (IA). IA provides grants and other disaster assistance directly to households. Most declarations are focused on public, not individual, assistance: While over 95% of declarations provide PA, just over 40% of presidential declarations include IA. Small, localized disasters will not be severe enough to generate a disaster declaration and so no federal assistance will be forthcoming for these events.

IA includes grants to repair individuals' damaged homes, replace personal property, and/or cover other expenses (e.g., temporary housing or medical expenses). These grants are only available for losses that are not covered by insurance or other assistance, and unlike insurance, the intent of assistance for home repairs is to make the home safely habitable, not repair it to pre-disaster conditions. The maximum FEMA IA award adjusts with inflation and was $37,900 in 2021 (FEMA 2021); however, the typical household award is only a few thousand dollars. Collier and Ellis (2024) examine IA grants from 2005 to 2018 and find that the average grant award was $4500 (in 2018$). To further discourage the use of federal assistance over insurance, households located in the FEMA-mapped 100-year floodplain who receive disaster assistance related to flooding must obtain a flood insurance policy and maintain it in order to be eligible for federal assistance associated with future floods.

FEMA coordinates federal assistance in the local community. It establishes local tents or offices where affected households are encouraged to register for potential assistance. Households with incomes below 125% of the federal poverty line are referred directly to the IA grant program. Better-off households are directed to the federal disaster loan program (discussed next) and are only eligible for an IA grant if they are denied a loan or their loan is insufficient to cover their losses. Only 24% of grant applicants were approved for assistance over the period 2004 to 2018 (Collier and Ellis 2024). This low approval rate may reflect the detailed documentation (e.g., financial records, onsite inspections) required to receive an IA grant (Raker 2023), which can deter some households.

Gallagher et al. (2023b) examine how FEMA grants affect consumer debt and migration following severe tornadoes. They use triple-differences specifications in which they compare individuals who were in areas that received IA versus those that did not following a severe tornado. Whether a county receives IA depends on several criteria, including aggregate casualties and physical damages. The

premise of their analysis is that households' credit outcomes are a function of the tornado damages and that these other county-level criteria generate plausibly exogenous variation in households' access to cash grants. Their regressions can be interpreted as estimating, for a given level of damage, how FEMA grants influence households' credit outcomes. They find a decline in credit card debt and an increase migration for households in areas that receive cash grants and were severely affected by tornadoes. They also find that for credit-constrained households pre-disaster (e.g., those with low credit score or high credit card utilization), cash grants reduce bill delinquencies.

Federal Disaster Loans Disaster recovery loans are the primary form of post-disaster federal assistance. These loans are administered by the Small Business Administration (SBA), but applying households do not need to be associated with a small business. These loans are subsidized, typically offered at an interest rate of about half of the prevailing 30-year, fixed-rate mortgage interest rate and are long term, up to 30 years. Households can borrow up to $240,000 to repair uninsured damages, though the median borrower only takes a loan of around $25,000 (Office of Disaster Assistance 2018; Collier and Ellis 2024). This program balances two objectives: (1) facilitating recovery and rebuilding through access to low-cost credit and (2) minimizing the cost to taxpayers by ensuring that borrowers are in a position to repay. As a result, the program underwrites these loans based on the applicants' credit score and monthly cash flows, concluding that many households are not financially equipped to take on additional debt. Over half of federal disaster loan applicants are denied.

Collier et al. (2024) examine the effects of disaster recovery loans on households' finances by merging disaster loan applications with consumer credit reports. Applicants whose monthly debt payments are below 40% of their monthly income are significantly more likely to be approved. The authors use this threshold in a regression discontinuity design and find that disaster recovery loans reduce the likelihood of consumer bankruptcy and credit delinquency on credit cards, mortgages, and auto loans. These effects are persistent, visible even 3.5 years after the disaster. While we might expect that the private debt balances of declined applicants to increase—the declined might turn to private credit after being rejected for a subsidized loan—the authors do not find this effect, concluding that declined applicants may instead be delaying repairs or funding them through savings or assistance from friends and family. Additionally, they find that in the years following the disaster, recovery loan recipients are more likely to take out new auto loans, suggesting that their finances had recovered sufficiently to take on additional debt to fund spending on durable goods.

Despite these apparent benefits of disaster recovery loans, two other papers conclude that household demand for these loans is limited. Disaster recovery loans subsidize borrowers through two channels: interest rates and collateral rules. Collier and Ellis (2024) examine how the program's interest rates affect disaster loan take-up. Using discrete jumps in the program's interest rate over time, they estimate an aggregate demand curve for disaster recovery loans. This curve is downward sloping with a semi-elasticity of -0.26: A 1 percentage point (pp) increase in the offered interest rate reduces take-up by 26%. This estimate falls between the estimated semi-elasticities on other credit products: It is lower than that for credit cards (typically around 1, Gross and Souleles 2002) but higher than first mortgages (which is near 0, DeFusco and Paciorek 2017). An interesting feature of their demand curve is that consumers are sensitive to rate changes even when the offered rate is extremely low. For example, around 10% of applicants would accept a disaster recovery loan if offered a rate that is 4 pp below the prevailing mortgage rate but not at 3 pp below that rate. This sensitivity is surprising because the marginal utility of consumption is likely quite high in the aftermath of a disaster as households have so many immediate financing needs. The authors show that demand varies by the size of monthly payments, suggesting that consumers' concerns regarding fulfilling future payments reduced demand. Demand also varied with applicant credit quality: Higher-credit-quality applicants (e.g., those with higher credit scores) were less likely to accept the loan at any offered interest rate.

Given the extremely low rates that some high-credit-quality applicants reject, this reluctance may reflect a form of debt aversion.

Collier et al. (2023a) examine how the disaster loan program's collateral rules affect borrowing. Only disaster loans exceeding a certain dollar amount ($25,000 as of 2022) must be collateralized, with the damaged home serving as collateral. This rule has a huge effect on loan amounts: About one-third of all borrowers bunch at this threshold, choosing the largest loan possible without posting collateral. The authors estimate that borrowers would have increased their loan amounts by 65% on average if the program did not require collateral. Even borrowers with little or no home equity tend to avoid posting collateral, which the authors conclude indicates that attachment to the home may play an important role—while a borrower with an underwater mortgage has no equity to lose, they would be required to leave the collateralized home if they default.

The authors additionally leverage changes in the collateral threshold to examine the effects of collateral on the likelihood of default. Households with disaster damages near the threshold are more likely to bunch than households with damages well above the threshold. For a given amount of damages, changes in the threshold over time (e.g., from $14,000 to $25,000) create exogenous variation in the likelihood that a household will bunch at the threshold, which the authors use as an instrument. They find that supplying collateral reduces the likelihood of default by one-third (reducing it from 11% to 7% of borrowers). Their finding aligns with classical theories regarding moral hazard and lending (e.g., Bester 1985): Collateral reduces the effects of asymmetric information that increase the cost of lending. In this setting, specifically, it indicates that for a given amount of taxpayer dollars to fund disaster recovery loans, collateral requirements expand the amount that the program can lend.

In summary, the research on disaster recovery loans poses an interesting puzzle: While recovery loans offer persistent benefits to consumers, consumers appear reluctant to use them since take-up, on both the extensive and intensive margins, is quite sensitive to the terms. A potential explanation in this context is population heterogeneity. The immediate liquidity provided by disaster recovery loans may be especially helpful for less creditworthy households, yet these households often have little capacity to take on additional debt. Whether it is advisable to further relax the program's underwriting standards to lend to more of these households is unclear: Lending to families who cannot repay the loan can result in additional hardship if they ultimately default. Credit risks are lower for better-off households, but these households frequently have other means to fund repairs, as we discuss below. One policy motivation for providing subsidized loans to better-off households is the potential of positive spillovers in terms of community recovery from expanded home repairs. Thus, benefits of disaster recovery loans may vary by population, which is a possibility that the existing research has not fully examined.

Privately Provided Credit Almost 20% of respondents reported using credit cards to fund repairs (Table 1). The implications of card utilization in this context for households' financial well-being are ambiguous. Credit cards provide immediate access to emergency liquidity but have high interest rates. Relatedly, Morse (2011) and Dobridge (2018) find that payday borrowing, an even costlier form of credit, increases after disasters and appears to help smooth consumption. The concern is that if consumers carry credit card balances across months or years, the high interest rates on cards compound the cost of the disaster.

If consumers pay their full card balances each month, however, they would not incur cards' high interest charges. del Valle et al. (2022) examine card utilization among households flooded by Hurricane Harvey. They use flood inundation depths as a severity measure in a difference-in-differences design. They find that flooded households increased credit card spending following Harvey. In particular, they find that flooded households often opened new cards at promotional, 0% interest rates, spent on these cards, and paid down card balances before the promotional rates expired.

This savvy strategy, which appears available primarily to better-off households, allows the household to benefit from the emergency liquidity provision of credit cards without incurring their high costs.

Only 8% of the respondents funded repairs through a formal loan from a private lender. This relatively low level may reflect credit supply constraints (e.g., lenders may be unwilling to accept a damaged home as collateral), consumer substitution toward lower-interest rate federal disaster recovery loans, and/or consumer debt aversion.

del Valle et al. (2022) also document that households affected by Harvey leveraged mortgage forbearance in a way that would likely minimize borrowing costs after the storm. Many lenders provided forbearance, allowing consumers to delay repaying their mortgages for several months without incurring penalties or adverse effects on their credit reports. The loan continued to accrue interest during the forbearance period. Thus, by delaying mortgage payments and directing those funds toward repairs, flooded households were effectively financing repairs at their mortgage interest rate, which is typically the lowest interest rate available to consumers in private credit markets.

Other Transfers While there is a body of literature on social capital and the role of social networks in disaster preparedness and emergency response, there is less research on the financial role family and friends, charities, and others play in the economic recovery of households. Twenty-nine percent of survey respondents funded repairs with the assistance of family and friends (Table 1). Hamel et al. (2018) similarly find that help from family and friends plays an important role: Among households who were displaced from their home by Hurricane Harvey, 57% borrowed money from friends or family to make ends meet. Research from COVID finds that households that struggled to save before the pandemic were three times more likely to draw on support from family and friends and that borrowing from friends and families reduces the likelihood of financial hardship (Robb and Warren 2021). Remittances also increase after a disaster in lower- and middle-income countries, providing more funding for recovery (Bettin and Zazzaro 2018). Such transfers can provide an important safety net, but it is also unlikely that family and friends, many of whom may also have been hit by the same disaster if geographically proximal, could provide funding at the level necessary to cover a huge spike in expenses.

Regarding organizations in the community, Hamel et al. (2018) find that almost 30% of the household respondents affected by Harvey received help from a local or national charity. Post-disaster, employers may also offer support to impacted employees or the impacted area and relief groups more broadly. This has been documented in news reporting, but research is limited. For example, in the aftermath of Hurricane Harvey, CNN reported that companies had pledged over $157 million in relief dollars, many choosing to donate to places like the Red Cross, and also encouraging employees to donate through matching programs, or mobilizing volunteers and in-kind donations (Yurieff 2017). A survey of businesses following devastating flooding from Hurricane Floyd in North Carolina in 1999 found that independent, local businesses and locally owned franchises were more likely to support employees post-disaster than local branches of national chains, as were those with a charitable giving plan in place pre-disaster or where managers already had ties with local civic or religious groups (Bin and Edwards 2009).

Local and state governments may also establish recovery programs for households when they receive funding for such efforts from the federal government. Since the early 1990s, following very large disasters, Congress often authorizes and appropriates funds to the Community Development Block Grant—Disaster Relief program managed by the Department of Housing and Urban Development (HUD). This program can provide large sums for state and local governments—billions of dollars—to use quite flexibility after a disaster. Since this is not a standing program, each time it is activated, HUD issues rules about eligibility and use of funds, and state and local governments prepare Action Plans as part of their application as to how they will use the funds. Applicants must typically target a

certain percentage of funds at lower-income and otherwise disadvantaged areas. The funds are often used for housing, public facilities, economic reconstruction, and investments in hazard mitigation. While this program can, when activated, provide substantial funding, some of which could make its way to households, it is uncertain and takes a very long time to be spent in a community. Recent research found that on average, housing activity from CDBG-DR funds after Hurricane Sandy began 2.3 years after the disaster declaration (Martín et al. 2022).

Persistence In summary, disasters can generate a sudden, large financing challenge, leading households to piece together funds from a variety of sources. In some cases, households cannot organize enough funding to cover all their expenses. Hamel et al. (2018) report that among surveyed households whose homes were still unrepaired one year after Harvey, over half cited that they could not afford the repairs. A participant in their focus groups clarified: "I did get several [bids from contractors], anywhere from $60,000 to $80,000 or more. And the amount that I got from FEMA was not even close. So I stretched it out as far as I could, to take care of home repairs, as well as trying to replace a vehicle, because my vehicle was flooded. And I've charged up credit cards as well and pulled money from savings accounts which I'll probably be taxed on, and my house is still not finished."

Without the necessary funds, timely received, costs can cascade. Some individuals who cannot replace a lost vehicle, for example, and have no other way to reach their employer, lose their jobs (Hamel et al. 2018). Or, as another example, if mold remediation is not done quickly, it can spread, causing larger damages and health problems. Households sometimes cope by selling personal possessions, spending less on medical care, or falling behind on bills (Sweeney et al. 2022; Farrell and Greig 2018; You and Kousky 2024). Hurricane Harvey reduced college enrollment and graduation rates among young adults from affected areas (Gallagher et al. 2023a). In addition to federal disaster assistance, Deryugina (2017) finds that hurricanes increase the use of other government safety nets including income maintenance, family assistance, disability insurance, unemployment insurance, and public healthcare provision, signaling the heightened financial demands and difficulties of households in affected areas.

Unmet financing needs and these coping responses would seem to explain research findings that natural disasters can lead to declines in home ownership and increases in loan delinquencies, mortgage default, and bankruptcy rates. Using consumer credit reports, Billings et al. (2022) examine the effects of Harvey-related flooding on measures of household financial distress: credit delinquency and bankruptcy. Following Gallagher and Hartley (2017) who examine credit report data following Hurricane Katrina, Billings et al. (2022) estimate difference-in-differences models in which the intensity of treatment is measured by the depth of flooding in the census block. They consider heterogeneity by examining sub-populations who are likely to vary in their access to recovery funds. Specifically, they use whether households live in a high-risk flood zone as a proxy for flood insurance take-up because of mandatory flood insurance requirements in these areas. They also construct an ex ante financing constraints measure that combines the individual's credit score and block-level income. Like Gallagher and Hartley (2017), they find little evidence of financial distress in the aggregate. Rather, distress is concentrated among areas with high home ownership rates, low flood insurance take-up, and high ex ante financing constraints. Roughly, we think about these homeowners as having the potential for large hurricane losses and a limited capacity to fund them. Harvey increased bankruptcy rates by 20% for these households.

5 Recovery Varies with Income and Race

A large body of research finds unequal disaster recovery across income and race (e.g., Fothergill and Peek 2004; Hallegatte et al. 2020; Hamel et al. 2018). A Federal Reserve (2022) study, for example, found that individuals earning less than $50,000 were more than twice as likely post-disaster to experience lost income, longer-term displacement, or know someone who was injured or killed from the disaster (Board of Governors 2022). The study also finds that Black or Hispanic adults were more likely to be affected by natural disasters than White or Asian adults (both overall and within income and education categories, Board of Governors 2022). Modeling reveals that while lower-income households may suffer a lower share of absolute asset losses, they suffer a disproportionate impact on their well-being (Markhvida et al. 2020). In this section, we revisit the risk management and recovery research above through the lenses of income and race, examining how these factors associate with exposure and vulnerability to hazards, as well as access to and use of the various financing sources.[3]

Identifying the causal role of racial discrimination on exposure to and recovery from climate events is extremely difficult, and effectively all of the existing research uncovers associations rather than causal effects. Current racial discrimination may reduce the risk management tools available to minorities (e.g., affecting credit approval, Howell et al. 2024). Yet, the legacy of historical discrimination, especially against Black individuals, also contributes to racial disparities in income, wealth, educational attainment, location, and a host of other socioeconomic factors that, in turn, impede recovery. For example, "redlining"—the practice in which lenders refused to provide mortgages in minority and low-income communities—in the 1930s continues to affect neighborhoods today, including their racial composition and features of the housing stock (Aaronson et al. 2021). As a result, the underlying causes of observed correlations between race and risk outcomes are difficult to isolate. An insurer might charge a Black policyholder higher premiums than a White policyholder because of direct discrimination (Becker 1957); however, higher premiums may instead (or may also) reflect the greater risk of the minority-owned property due to its characteristics (called "statistical discrimination" or indirect discrimination). That is, while it is commonly held that insurers are not typically engaged in any direct discrimination—many states ban use of protected classes in pricing—race could be correlated with the underlying hazard, investment in protective measures, the safety of homes, and other drivers of claims payouts and thus could lead average premiums to vary by race (Medders et al. 2021; Schwarcz 2019). Similarly, observed variation in access to post-disaster financing resources based on race or other demographic characteristics could similarly be driven by multiple underlying causes.

Exposure and Vulnerability One explanation of observed inequities in post-disaster outcomes could be lower-income households or households of color being in more vulnerable housing or at higher risk of the hazard itself. A review of sociological research finds that low-income populations are more vulnerable to natural disasters due to a range of factors including exposure to hazards, type of homes and building construction, correlation with special needs, and locations (e.g., highly rural areas) that are difficult for emergency response services to access (Fothergill and Peek 2004). For example, lower-income households and households of color were more vulnerable to and had more

[3] Household income and wealth are highly correlated, and wealth may be more important for understanding recovery since wealthy households could sell their assets or borrow against them. However, wealth is difficult to measure, and income is a metric used by private lenders and government programs (e.g., Medicaid). Research primarily examines income heterogeneity so we focus our discussion on it.

difficult recoveries from Hurricane Katrina, which hit New Orleans in 2005 (e.g., Masozera et al. 2007; Pastor et al. 2006).

Flooding, as the most common and costliest natural disaster, has been the most researched as to any equity dimensions in distribution of the underlying peril. The relationship between flood risk and income or race often depends on existence of water-related amenities correlated with flood risk. Roughly, some properties are exposed to flood risk because they are in desirable locations near water (e.g., beaches and river fronts), while others lack water amenities but are in flood-prone areas (e.g., low-lying parts of a city). Flood risk depresses the value of the latter properties, leading lower-income (and often racial minority) households to sort into them (Bakkensen and Ma 2020). Especially in the U.S. South, racial and ethnic minorities are more likely to live in low-lying areas with higher flood risk but few amenities (Ueland and Warf 2006). Similarly, in Miami, better-off households locate in properties with access to the beach, while lower-income minorities (e.g., Black, Colombian, and Puerto Rican individuals) lived in amenity-scarce neighborhoods that are prone to inland flooding (Montgomery and Chakraborty 2015). National flood modeling of the risk today and under climate scenarios suggests that while a disproportionate share of flood damages currently are borne by lower-income, White communities, escalating coastal flood risk in the future will disproportionately impact communities with a larger share of Black residents (Wing et al. 2022).

Insurance Household insurance data that also has demographic information is typically unavailable. As such, research in this area tends to rely on either surveys or aggregate data linked to Census information. Baker and McElrath (1996) examine the effects of income and race on property insurance take-up using a 1993 survey of over 3400 employees of the University of Miami following Hurricane Andrew. In regression models, they find that having property insurance is positively related to home ownership, income, and education. Despite including these socioeconomic factors, they also find that race—indicators for whether the respondent is Black, Hispanic, or another racial/ethnic minority—is a statistically significant predictor of having property insurance in their regressions. Another study of Hurricane Andrew found that almost twice as many African Americans and non-Cuban Hispanics as Whites indicated that their insurance company was not providing sufficient funds for repairs even after controlling for income and damage levels, which could be due to Whites being more likely to have coverage from "big name" firms (Peacock and Girard 1997).

Several studies have relied on the racial composition at the ZIP code or census tract level as a proxy for the policyholder's race. Klein and Grace (2001) use ZIP code-level data to examine homeowners insurance for evidence of pricing that reflects racial discrimination. In models including controls for risk-related features of the housing stock (e.g., home age) and the neighborhood (e.g., crime statistics), they do not find evidence of race-related pricing differences. Niehaus and Terry (1993) come to a similar conclusion using ZIP code-level data when assessing for racial discrimination in the auto insurance market.

Cheung et al. (2015) examine the pricing of wind insurance policies in the Citizen's Property Insurance Corporation of Florida. Their rich policy-level data include premiums and the property's predicted average annual loss based on a widely used catastrophe model. They assess how premiums vary based on census tract measures of income and race in regression models that control for features of the property and contract (e.g., the coverage limit). They find that policyholders in census tracts with lower income or a larger share Hispanic population are charged higher premiums relative to their expected loss. Cheung et al. (2015) believe that this disparate pricing is unintentional and may reflect unobserved differences (e.g., between the factors used to model the risk and those used to determine premiums, which may be restricted by regulation).

Apart from pricing, race and ethnicity do appear to influence the claims process. In their survey following Hurricane Andrew, Baker and McElrath (1996) find that Hispanic policyholders who filed

claims experienced more delays in receiving claims payments than non-Hispanics. They attribute this difference to the discretion given to claims adjusters: Adjusters worry about insurance fraud and may influence the speed, documentation required, and size of claims payments based on the extent to which they find the policyholder trustworthy. Differences in language, culture, and institutional knowledge likely increase the difficulty of navigating the claims process for ethnic minorities. Chan (1999) examines consumer complaints to the Missouri Department of Insurance regarding homeowners and fire insurance. Most complaints relate to claims denials or settlement offers but can also pertain to policy cancellations, premium increases, and so on. In regression models, he finds that policyholders in ZIP codes with a larger share of racial and ethnic minorities are more likely to file complaints, including valid complaints (those that ultimately lead to concessions by insurers). More recent work also finds differences in claims treatment. Lin et al. (2022) find evidence of racial discrimination in the adjudication of earthquake insurance claims in Oklahoma triggered by oil and gas drilling. Using models with a large set of control variables, they find that claims filed by policyholders in ZIP codes where a larger share of the population is Black are less likely to be approved. Additionally, when policyholders in these ZIP codes receive a claims payment, the amount is smaller.

Savings and Privately Provided Credit Existence and amount of savings are clearly related to income and also race. Minorities are less likely to have retirement savings; around 20% of working age White adults do not have any retirement savings, versus around 40% for Black or Hispanic, working age adults (Brevoort et al. 2021). Moreover, only about a quarter of Black or Hispanic adults with retirement savings report that they are saving at an appropriate rate.

Racial disparities in approval for private credit are also large. In 2020, 81% of White applicants who applied for some form of credit were approved, versus 49% of Black applicants (Brevoort et al. 2021). Presumably, such disparities persist during post-disaster borrowing, as well.

Federal Household Grants Recent research has raised questions about equity in federal disaster programs. Since Individual Assistance grants are targeted to those with costs not covered by other sources, such as insurance or loans, it is not surprising that the majority of federal disaster grant applicants and recipients are in the lowest quartile (i.e., bottom 25%) of the income distribution (Collier et al. 2024). That said, only 24% of all grant applications are approved, and the approval process appears to favor better-off applicants. Smaller grant amounts are awarded to the most socially vulnerable groups, and some neighborhoods with lower average income experience both lower average amounts and lower probabilities of receiving assistance (Billings et al. 2022; Drakes et al. 2021). One explanation for this finding is that struggling households are less able to navigate the grant application process, which requires substantial documentation and restricts how the money is used. In focus groups of those impacted by Hurricane Harvey, many participants noted confusion with how to receive disaster assistance, ranging from being unaware of what was available to describing challenges navigating the process (Hamel et al. 2018). A review of both academic and nonacademic literature on equity in access to federal resources finds that lower-income and minority households face barriers in navigating the application process and are more likely to be denied (Wilson et al. 2021).

Raker (2023) assesses over 5 million hurricane-related FEMA grant applications from 2005 to 2016, analyzing the effects of income and race at each step of the application process. The first step in the process includes an onsite inspection of damages. Using multivariate regression models, he finds that among applicants, those with higher relative incomes that reside in White, low-income ZIP codes are more likely to receive an inspection. On inspection, applicants in communities of color were more likely to be deemed ineligible, for example, due to determinations that the damage did not make the home uninhabitable or an inability to document that the damage was disaster-

related (i.e., not preexisting). Similar to the findings of Baker and McElrath (1996) regarding insurance claims, he concludes that steps in the grant approval process that are intended to reduce fraud, which rely heavily on the perceptions of onsite inspectors, work against racial and ethnic minorities. Among approved applicants, higher-income households in Whiter communities received larger grants. Among approved homeowners, those in Black communities received smaller grants than those in White communities, which appears to be explained by differences in the inspectors' assessed damage amounts. Higher-income households receive more generous rental assistance than lower-income households, a difference which is not explained by damage amounts. He also finds statistically significant interactions between income and race that increase the disparity in grant amounts between lower-income households in communities of color and higher-income households in White communities.

Federal Disaster Loans Federal disaster recovery loans are more likely to go to those with a greater repayment ability. Collier and Ellis (2024) show that while applicants come from throughout the income distribution, households with above-median incomes are more likely to be approved. This regressive pattern appears due to the underwriting process: The disaster loan program approves loans based on credit score and the applicant's debt-service-to-income (DTI) ratio, both of which are correlated with income. Billings et al. (2022) similarly find that households who are more likely to be credit constrained prior to the disaster are less likely to receive disaster loans. Begley et al. (2024) show that access to disaster loans varies with race as households from areas where a larger share of the population is a racial or ethnic minority are less likely to be approved. The authors conclude that this racial disparity does not reflect racial discrimination by loan officers, using a test from Becker (1957).[4] Instead, the disparity may emerge from the correlation between race and measures of credit risk, such as income and credit score.

Other Transfers Low-income households and Black households are more likely to turn to family and friends for assistance following disasters. Hamel et al. (2018) find that half of Black respondents (vs. 28% of White respondents) reported borrowing money from family and friends to make ends meet. Similarly, they find 45% of households below 200% of the federal poverty line borrowed from family and friends.

State and local programs funded by HUD can vary substantially in design and implementation. Some local governments may prioritize aid to disadvantaged households and communities. Others may fail to do so, despite program requirements mandating some amount of targeting (Gotham 2014). Miller and Simile (1992) describe assistance from local governments, businesses, and churches to low-income households following Hurricane Hugo and provide an interesting insight on the relationship between poverty and recovery. Relief workers and volunteers were shocked by the living conditions of low-income households in their communities and had a difficult time differentiating disaster damages from preexisting disrepair. While volunteers from churches and businesses sometimes took a broader view, many relief workers were careful to address only hurricane damages. Government workers appeared especially concerned about "free riders" who might use hurricane assistance to address disrepair. Universally, workers and volunteers left when they believed that the task of repairing hurricane damages was complete.

[4] The premise of the test is that in the presence of discrimination, minorities who are approved for a loan would have a lower default rate because minority borrowers must be of better quality than nonminority borrowers to overcome discrimination in approvals. However, Begley et al. (2024) find no racial difference in loan performance among approved applicants, suggesting that racial disparities are not due to this discrimination channel.

Persistence Greater exposure and vulnerability to hazards, as well as more limited access to and use of arm-length financial services, seem to culminate, increasing the likelihood that lower-income households and people of color will experience persistent consequences of a severe climate event (Billings et al. 2022; Rhodes and Besbris 2022). Howell and Elliott (2019) use the Panel Study of Income Dynamics (PSID) to follow individuals from 1999 to 2013, assessing how living in an area affected by a natural disaster affects inequality. In contrast to almost all other datasets used in this literature, the PSID collects information on household wealth. Disaster damages are measured at the county level and taken from the Spatial Hazard Events and Losses Database of the U.S.A. Using multivariate regression models, they find a positive relationship between wealth and disaster damages for White households and a negative relationship for households of color. These relationships hold in models with a large set of demographic controls, including annual insurance premiums paid, marital status, home ownership, education, age, and neighborhood and county demographic and socioeconomic variables. They find similar patterns of divergence based on education and home ownership: The relationship between wealth and disaster damages is negative for households with less education and who rent and positive for their better-off counterparts. Their results align with a repeated research finding that severe climate events ultimately worsen inequality and racial disparities (e.g., Ratcliffe et al. 2020; Rhodes and Besbris 2022).

6 Discussion and Conclusion

In summary, in this chapter we (1) provide a more complete perspective of how households fund recovery from severe climate events, noting the specific role of insurance, and (2) highlight the centrality of income and race in understanding household recovery.[5] We conclude by discussing the potential implications of this literature for policies that might reduce the financial consequences of severe climate events for households.

Households' difficulties funding recovery highlight the importance of disaster insurance. A severe climate event can impose a large, urgent financing challenge on households, requiring them to garner tens or even hundreds of thousands of dollars to repair property and manage additional financial demands (e.g., temporary housing, lost wages). Insurance plays a unique role in this context as it tends to provide larger sums than any other alternative source of funds. Without access to as large an amount of funding, the uninsured appear to especially struggle to overcome the financing challenge of severe climate events. As catastrophe insurance primarily focuses on property damages, however, it may not cover the full costs of a natural disaster. As such, even insured households often need several sources of funding to manage a severe event.

A long-standing concern among academics, policymakers, and pundits is that households may treat insurance and ex post funding sources as substitutes, such that government grants and loans could meaningfully reduce insurance take-up. Empirical research shows that, *across households*, insurance and recovery funding look like complements, not substitutes: Better-off households are not only more likely to be insured, but also more likely to receive federal loans and larger grants, have retirement savings that they can withdraw (potentially without penalty), have promotional 0% interest credit card offers, and so on. All of these are less frequently available and tend to be given in smaller amounts to lower-income and minority households creating a pattern of haves and have-nots in disaster recovery.

[5] Research in other disciplines, especially sociology, has long been interested in how recovery outcomes vary by income and race, mostly examining such disparities through field interviews and surveys (e.g., Fothergill and Peek 2004). Recent research has advanced this effort by assessing administrative data to quantify recovery and differential outcomes.

This pattern highlights the importance of financing frictions not only in insurance and credit markets but also in the current design of federal recovery programs.

These findings stress the need to improve targeting in the range of government disaster-related safety nets, including insurance, through means testing. Existing subsidies in the NFIP, federal disaster loan program, and FEMA grant program could be redirected through improved means-based targeting. There has already been extensive work on how to design a means-tested assistance program for the NFIP (FEMA 2018; Council et al. 2015). Additionally, opportunities likely exist for these programs, which are currently housed in different places in the federal government, to collectively provide a more cohesive disaster safety net. This pattern that higher-income households have many sources of recovery funding and lower-income households have few or none alleviates certain charity hazard concerns that have limited the generosity of government disaster assistance for lower-income households. Since lower-income households do not typically buy disaster insurance in the current system, we can be less worried that improving their recovery options will distort their insurance choices.[6] Improving disaster recovery through means testing would potentially help address some of the persistent effects of disasters, such as mortgage foreclosure and personal bankruptcy. In addition, new forms of insurance, such as parametric microinsurance and meso-insurance models (in which an intermediary institution helps secure coverage on behalf of a group), are expanding beyond their initial piloting in the global south to reach vulnerable populations in many geographies. Such innovative approaches could become complements to more standard, indemnity-based, property insurance policies.

Race and ethnicity appear to influence households' recovery in ways that extend beyond income. A growing body of evidence shows that the community in which a person grows up affects their economic mobility and financial well-being as an adult (e.g., Chetty et al. 2016; Chetty and Hendren 2018), contributing to a broader interest in place-based policies (Neumark and Simpson 2015). Disasters are place-based risks, and racial and ethnic minorities live in communities that can increase their vulnerability. Place-based targeting of disaster prevention and recovery funds may be important for these under-served communities. Such policies could be conditioned on community-level characteristics (e.g., education, poverty rates, access to financial services, climate exposure), allowing communities to graduate as conditions improve. While existing response efforts are already location-dependent because of the nature of disaster risk, the current system largely allocates resources based on county damages. Direct consideration of community characteristics, especially at a smaller scale, might greatly improve the targeting of these programs. An additional concern is how to modify operational processes focused on fraud detection since processes relying on assessors' discretion and overly burdensome documentation appear to work against households of color, those with less education, non-native English speakers, and generally under-resourced households.

Finally, losses from extreme events continue to grow due to climate change and ongoing development in at-risk areas. Households are more likely to experience a natural disaster than in the past, and when they do, the costs of recovery are larger. Insurance is fundamental for financial resilience of severely affected households. Yet a focus on insurance alone offers a highly incomplete view of how households fund recovery and the degree to which the consequences of a severe event persist. This more comprehensive view is essential to designing markets and policies that minimize the costs of severe climate events.

Acknowledgments All errors are our own.

[6] The other concern regarding charity hazard reflects households' location decisions, but here also, research suggests that it is higher-income households who elect to live near beaches and riverfronts, while affordable housing draws lower-income households into (amenity poor) flood zones.

References

Aaronson D, Hartley D, Mazumder B (2021) The effects of the 1930s HOLC "redlining" maps. Am Econ J Econ Policy 13(4):355–392

Anderson DR (1974) The national flood insurance program. problems and potential. J Risk Insur 579–599

Andor MA, Osberghaus D, Simora M (2020) Natural disasters and governmental aid: is there a charity hazard? Ecol Econ 169:106534

Aon (2023) Weather, climate and catastrophe insight. Aon, Impact Forecasting

Armantier O, Foncel J, Treich N (2023) Insurance and portfolio decisions: two sides of the same coin? J Financ Econ 148(3):201–219

Baker T, McElrath K (1996) Whose safety net? Home insurance and inequality. Law Soc Inquiry 21(2):229–264

Bakkensen LA, Barrage L (2021) Going underwater? Flood risk belief heterogeneity and coastal home price dynamics. Rev Financ Stud 35(8):3666–3709

Bakkensen LA, Ma L (2020) Sorting over flood risk and implications for policy reform. J Environ Econ Manag 104:102362

Becker GS (1957) The economics of discrimination. University of Chicago press

Begley TA, Gurun UG, Purnanandam A, Weagly D (2024) Disaster lending: "fair" prices, but "unfair" access. *Management Science*

Bester H (1985) Screening vs. rationing in credit markets with imperfect information. Am Econ Rev 75(4):850–855

Bettin G, Zazzaro A (2018) The impact of natural disasters on remittances to low- and middle-income countries. J Dev Stud 54(3):481–500

Billings SB, Gallagher EA, Ricketts L (2022) Let the rich be flooded: the distribution of financial aid and distress after hurricane Harvey. J Financ Econ 146(2):797–819

Bin O, Edwards B (2009) Social capital and business giving to charity following a natural disaster: an empirical assessment. J Soc.-Econ 38(4):601–607

Board of Governors (2022) Economic well-being of U.S. households in 2021. Board of Governors of the Federal Reserve System

Bolin R, Bolton P (1986) Race, religion, and ethnicity in disaster recovery. Institute of Behavioral Sciences, University of Colorado

Born PH, Klimaszewski-Blettner B (2013) Should I stay or should I go? The impact of natural disasters and regulation on us property insurers' supply decisions. J Risk Insur 80(1):1–36

Botzen W, Robinson P (2023) Climate risk and insurance, forthcoming, Handbook of Insurance

Botzen WW, van den Bergh JC (2012) Risk attitudes to low-probability climate change risks: WTP for flood insurance. J Econ Behav Organ 82(1):151–166

Bradt JT, Kousky C, Wing OE (2021) Voluntary purchases and adverse selection in the market for flood insurance. J Environ Econ Manag 110:102515

Brevoort K et al (2021) Economic well-being of US households in 2020. Board of Governors of the Federal Reserve System

Brody SD, Highfield WE, Wilson M et al (2017) Understanding the motivations of coastal residents to voluntarily purchase federal flood insurance. J Risk Res 20(6):760–775

Browne MJ, Hoyt RE (2000) The demand for flood insurance: empirical evidence. J Risk Uncertain 20:291–306

Browne MJ, Knoller C, Richter A (2015) Behavioral bias and the demand for bicycle and flood insurance. J Risk Uncertain 50:141–160

Casaburi L, Willis J (2018) Time versus state in insurance: experimental evidence from contract farming in Kenya. Am Econ Rev 108(12):3778–3813

Chan TS (1999) Service quality and unfair racial discrimination in homeowners insurance. J Risk Insur 83–97

Chang CH, Lee CC (2012) Non-linearity between life insurance and economic development: a revisited approach. Geneva Risk Insur Rev 37:223–257

Chetty R, Hendren N (2018) The impacts of neighborhoods on intergenerational mobility I: childhood exposure effects. Q J Econ 133(3):1107–1162

Chetty R, Hendren N, Katz LF (2016) The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. Am Econ Rev 106(4):855–902

Cheung R, Cole CR, Macpherson DA et al (2015) Demographic factors and price distortions in insurance. Risk Manag Insur Rev 18(1):1–28

Cole SA, Xiong W (2017) Agricultural insurance and economic development. Ann Rev Econ 9:235–262

Collier B, Ellis C (2024) A demand curve for disaster recovery loans. Econometrica 92(3):713–748

Collier BL, Schwartz D, Kunreuther HC et al (2022) Insuring large stakes: a normative and descriptive analysis of households' flood insurance coverage. J Risk Insur 89(2):273–310

Collier BL, Hartley DA, Keys BJ, Ng JX (2024) Credit when you need it (No. w32845). National Bureau of Economic Research
Collier BL, Ellis C, Keys BJ (2023a) The cost of consumer collateral: evidence from bunching. Tech. rep., National Bureau of Economic Research
Collier B, Huber T, Jaspersen JG, Richter A (2023b) Homeowners' willingness to hedge flood risks as prices increase. Available at SSRN 4635177
Council NR et al (2015) Affordability of national flood insurance program premiums: Report 1. National Academies Press
Courbage C, Rey B, Treich N (2013) Prevention and precaution. Handbook of insurance, pp 185–204
Cummins JD, Barrieu P (2013) Innovations in insurance markets: hybrid and securitized risk-transfer solutions. Springer, Berlin
Cummins JD, Mahul O (2009) Catastrophe risk financing in developing countries: principles for public intervention. World Bank Publications
Dagher V (2023) Americans are bailing on their home insurance. https://www.wsj.com/personal-finance/americans-are-bailing-on-their-home-insurance-e3395515
DeFusco AA, Paciorek A (2017) The interest rate elasticity of mortgage demand: evidence from bunching at the conforming loan limit. Am Econ J Econ Policy 9(1):210–240
Del Valle A, Scharlemann T, Shore S (2022) Household financial decision-making after natural disasters: evidence from Hurricane Harvey. J Financ Quant Anal 1–27
Deryugina T (2017) The fiscal cost of hurricanes: disaster aid versus social insurance. Am Econ J Econ Policy 9(3):168–198
Deryugina T, Kawano L, Levitt S (2018) The economic impact of Hurricane Katrina on its victims: evidence from individual tax returns. Am Econ J Appl Econ 10(2):202–233
Dobridge CL (2018) High-cost credit and consumption smoothing. J Money Credit Bank 50(2-3):407–433
Drakes O, Tate E, Rainey J et al (2021) Social vulnerability and short-term disaster assistance in the United States. Int J Disaster Risk Reduction 53:102010
Ericson KM, Sydnor JR (2018) Liquidity constraints and the value of insurance. Tech. rep., National Bureau of Economic Research
Evans CDavid. D; Webb, Xu EJ (2019) Wildfire catastrophe models could spark the changes California needs. https://www.milliman.com/en/insight/wildfire-catastrophe-models-could-spark-the-changes-california-needs
Farrell D, Greig F (2018) Weathering the storm: the financial impacts of hurricanes Harvey and Irma on one million households. Available at SSRN 3135266
FEMA (2018) An affordability framework for the National Flood Insurance Program
FEMA (2021) Notice of maximum amount of assistance under the Individuals and Households Program, Federal Emergency Management Agency, FEMA–2021–0001
Fothergill A, Peek LA (2004) Poverty and disasters in the United States: a review of recent sociological findings. Natural Hazards 32:89–110
Gallagher J, Hartley D (2017) Household finance after a natural disaster: the case of hurricane Katrina. Am Econ J Econ Policy 9(3):199–228
Gallagher E, Billings SB, Ricketts L (2023a) Human capital investment after the storm. Rev Financ Stud. https://doi.org/10.1093/rfs/hhad003
Gallagher J, Hartley D, Rohlin S (2023b) Weathering an unexpected financial shock: the role of disaster assistance on household finance and business survival. J Assoc Environ Resour Econom 10(2):525–567
Golnaraghi M, Nunn P, Wood RM et al (2018) Managing physical climate risk: leveraging innovations in catastrophe risk modelling. Geneva Association-International Association for the Study of Insurance...
Gotham KF (2014) Reinforcing inequalities: the impact of the CDBG program on post-Katrina rebuilding. Housing Policy Debate 24(1):192–212
Gourinchas PO, Parker JA (2002) Consumption over the life cycle. Econometrica 70(1):47–89
Gross DB, Souleles NS (2002) Do liquidity constraints and interest rates matter for consumer behavior? Evidence from credit card data. Q J Econ 117(1):149–185
Hallegatte S, Vogt-Schilb A, Rozenberg J et al (2020) From poverty to disaster and back: a review of the literature. Econ Disasters Climate Change 4:223–247
Hamel L, Wu B, Brodie M et al (2018) One year after the storm: Texas Gulf Coast residents' views and experiences with hurricane Harvey recovery
Hogarth RM, Kunreuther H (1992) Pricing insurance and warranties: ambiguity and correlated risks. Geneva Papers Risk Insur Theory 17:35–60
Howell J, Elliott JR (2019) Damages done: the longitudinal impacts of natural hazards on wealth inequality in the United States. Soc Probl 66(3):448–467
Howell ST, Kuchler T, Snitkof D et al (2024) Lender automation and racial disparities in credit access. J Financ 79(2):1457–1512

Hu Z (2022) Social interactions and households' flood insurance decisions. J Financ Econ 144(2):414–432
Hudson P, Botzen WW, Czajkowski J et al (2017) Moral hazard in natural disaster insurance markets: empirical evidence from Germany and the United States. Land Econ 93(2):179–208
Jaffee DM, Russell T (1997) Catastrophe insurance, capital markets, and uninsurable risks. J Risk Insur 205–230
Jeziorski W, Ramnath S et al (2021) Homeowners insurance and climate change. Chicago Fed Lett 460:1–6
Kaplan G, Violante GL, Weidner J et al (2014) The wealthy hand-to-mouth. Brookings papers on economic activity, p 77
Klein RW, Grace MF (2001) Urban homeowners insurance markets in Texas: a search for redlining. J Risk Insur 581–613
Kousky C (2011) Understanding the demand for flood insurance. Natural Hazards Rev 12(2):96–110
Kousky C (2018) Financing flood losses: a discussion of the National Flood Insurance Program. Risk Manag Insur Rev 21(1):11–32
Kousky C (2019) The role of natural disaster insurance in recovery and risk reduction. Ann Rev Resour Econ 11:399–418
Kousky C, Michel-Kerjan E (2017) Examining flood insurance claims in the United States: six key findings. J Risk Insur 84(3):819–850
Kousky C, Michel-Kerjan EO, Raschky PA (2018) Does federal disaster assistance crowd out flood insurance? J Environ Econ Manag 87:150–164
Kousky C, Palim M, Pan Y (2020) Flood damage and mortgage credit risk: a case study of hurricane Harvey. J Housing Res 29(sup1):S86–S120
Kriesel W, Landry C (2004) Participation in the national flood insurance program: an empirical analysis for coastal properties. J Risk Insur 71(3):405–420
Kunreuther H (1974) Disaster insurance: a tool for hazard mitigation. J Risk Insur 287–303
Kunreuther H, Slovic P (1978) Economics, psychology, and protective behavior. Am Econ Rev Papers Proc 68(2):64–69
Kunreuther H, Meszaros J, Hogarth RM et al (1995) Ambiguity and underwriter decision processes. J Econ Behav Organ 26(3):337–352
Kunreuther H, Pauly M et al (2006) Insurance decision-making and market behavior. Found Trends® Microecon 1(2):63–127
Landry CE, Jahan-Parvar MR (2011) Flood insurance coverage in the coastal zone. J Risk Insur 78(2):361–388
Lin X, Browne MJ, Hofmann A (2022) Race discrimination in the adjudication of claims: evidence from earthquake insurance. J Risk Insur 89(3):553–580
Marcoux K, Wagner K (2023) Fifty years of U.S. natural disaster insurance policy, forthcoming, Handbook of Insurance
Markhvida M, Walsh B, Hallegatte S et al (2020) Quantification of disaster impacts through household well-being losses. Nature Sustain 3(7):538–547
Marshall D et al (2018) An overview of the California earthquake authority. Risk Manag Insur Rev 21(1):73–116
Martín C, Teles D, DuBois N (2022) Understanding the pace of HUD's disaster housing recovery efforts. Housing Policy Debate 32(1):102–127
Masozera M, Bailey M, Kerchner C (2007) Distribution of impacts of natural disasters across income groups: a case study of New Orleans. Ecol Econ 63(2–3):299–306
McCoy SJ, Zhao X (2018) A city under water: a geospatial analysis of storm damage, changing risk perceptions, and investment in residential housing. J Assoc Environ Resour Econ 5(2):301–330
Medders LA, Parson JA, Thomas-Reid M (2021) Gender x and auto insurance: is gender rating unfairly discriminatory? J Insur Regul 40(7)
Melecky M, Raddatz C (2015) Fiscal responses after catastrophes and the enabling role of financial development. World Bank Econ Rev 29(1):129–149
Meyer R, Kunreuther H (2017) The Ostrich Paradox: why we underprepare for disasters. University of Pennsylvania Press, Philadelphia
Miller KS, Simile C (1992) They could see stars from their beds: the plight of the rural poor in the aftermath of Hurricane Hugo, University of Delaware, Disaster Research Center
Montgomery MC, Chakraborty J (2015) Assessing the environmental justice consequences of flood risk: a case study in Miami, Florida. Environ Res Lett 10(9):095010
Morse A (2011) Payday lenders: heroes or villains? J Financ Econ 102(1):28–44
Netusil NR, Kousky C, Neupane S et al (2021) The willingness to pay for flood insurance. Land Econ 97(1):17–38
Neumark D, Simpson H (2015) Place-based policies. In: Handbook of Regional and Urban Economics, vol 5. Elsevier, Amsterdam, pp 1197–1287
Nguyen CN, Noy I (2020) Measuring the impact of insurance on urban earthquake recovery using nightlights. J Econ Geogr 20(3):857–877
Niehaus G, Terry A (1993) Evidence on the time series properties of insurance premiums and causes of the underwriting cycle: new support for the capital market imperfection hypothesis. J Risk Insur 60(3):466–480

NOAA (2022) U.S. billion-dollar weather and climate disasters, U.S. National Oceanic and Atmospheric Administration. https://doi.org/10.25921/stkw-7w73
OECD (2021) Enhancing financial protection against catastrophe risks: the role of catastrophe risk insurance programmes. www.oecd.org/daf/fin/insurance/Enhancing-financial-protection-against-catastrophe-risks.htm
Office of Disaster Assistance (2018) Standard operating procedure, Disaster Assistance Program, Small Business Administration
Olshansky RB, Johnson LA, Topping KC (2006) Rebuilding communities following disaster: lessons from Kobe and Los Angeles. Built Environ 32(4):354–374
Outreville JF (2013) The relationship between insurance and economic development: 85 empirical papers for a review of the literature. Risk Manag Insur Rev 16(1):71–122
Pastor M, Bullard RD, Boyce JK et al (2006) In the wake of the storm: environment, disaster and race after Katrina. Russell Sage Foundation
Peacock WG, Girard C (1997) Ethnic and racial inequalities in hurricane damage and insurance settlements. Hurricane Andrew: ethnicity, gender and the sociology of disasters, pp 171–190
Petrolia DR, Landry CE, Coble KH (2013) Risk preferences, risk perceptions, and flood insurance. Land Econ 89(2):227–245
Raker EJ (2023) Stratifying disaster: state aid, institutional processes, and inequality in American communities. Social Forces 102(2):430–453
Rampini AA, Viswanathan S (2019) Financing insurance, NBER Working Paper 22293
Raschky PA, Schwarze R, Schwindt M et al (2013) Uncertainty of governmental relief and the crowding out of flood insurance. Environ Resour Econ 54:179–200
Ratcliff C, Middlewood B, Knoll MA et al (2022) Emergency savings and financial security: insights from the making ends meet survey and consumer credit panel. Consumer Financial Protection Bureau Office of Research Reports Series (22-1)
Ratcliffe C, Congdon W, Teles D et al (2020) From bad to worse: natural disasters and financial health. J Housing Res 29(sup1):S25–S53
Rhodes A, Besbris M (2022) Soaking the middle class: suburban inequality and recovery from disaster. Russell Sage Foundation
Robb H, Warren A (2021) When does financial support from family, friends play the role of emergency savings?
Sacks B (2023) Climate disasters make it harder to insure your home. Here's what to know. Washington Post. https://www.washingtonpost.com/climate-environment/2023/06/07/insurance-coverage-loss-climate-change/
Schwarcz D (2019) Towards a civil rights approach to insurance anti-discrimination law. DePaul L Rev 69:657
Surminski S (2018) Fit for purpose and fit for the future? An evaluation of the UK's new flood reinsurance pool. Risk Manag Insur Rev 21(1):33–72
Sweeney K, Wiley H, Kousky C (2022) The challenge of financial recovery from disasters: the case of Florida homeowners after hurricane Michael
Swiss Re Institute (2023) Sigma explorer. https://www.sigma-explorer.com
Thieken AH, Petrow T, Kreibich H et al (2006) Insurability and mitigation of flood losses in private households in Germany. Risk Anal Int J 26(2):383–395
Turnham J, Burnett K, Martin C et al (2011) Housing recovery on the Gulf Coast, phase II: results of property owner survey in Louisiana, Mississippi, and Texas. Washington, DC: US Department of Housing and Urban Development, Office of Policy Development and Research
Ueland J, Warf B (2006) Racialized topographies: altitude and race in southern cities. Geograph Rev 96(1):50–78
Von Peter G, Von Dahlen S, Saxena SC (2012) Unmitigated disasters? New evidence on the macroeconomic cost of natural catastrophes
Wagner KR (2022) Adaptation and adverse selection in markets for natural disaster insurance. Am Econ J Econ Policy 14(3):380–421
Wilson B, Tate E, Emrich CT (2021) Flood recovery outcomes and disaster assistance barriers for vulnerable populations. Front Water 3:752307
Wing OE, Lehman W, Bates PD et al (2022) Inequitable patterns of US flood risk in the Anthropocene. Nat Climate Change 12(2):156–162
You X, Kousky C (2024) Improving household and community disaster recovery: evidence on the role of insurance. J Risk Insur 91(2):299–338
Yurieff K (2017) Businesses donate over $157 million to Harvey relief efforts. CNN

Insuring Wildfires Globally

Mary Kelly and Zhe Peng

1 Introduction

Wildland or landscape fires, a general term describing any non-structure fire that occurs in vegetation such as trees, grasses, and shrubs, occur globally. Not all landscape fire is dangerous or unwanted: fires in natural environments are common occurrences, burning approximately 4% of the world's surface annually (Sigmund et al. 2021), and are critical to the healthy functioning of many ecosystems.[1] Some species of trees require fire for seed germination, and for other species, landscape fires accelerate the decomposition of organic matter and help create room for trees to thrive. Landscape fires may also kill insects and diseases that harm forests, paving way for new growth of grasses and other plants, which in turn, offer food and shelter for animals and birds (Parks Canada 2022).

Whether caused by nature or humans, when landscape fires burn out of control, they become wildfires. Consistent with the IPCC's (2020 p 4) definition of risk, wildfire risk arises from the frequency and severity of the wildfire hazard, the exposure of systems (e.g., people, communities, agricultural and ecological environments), and the vulnerability of the system (e.g., design of infrastructure, land use planning). In particular, fires that occur in or near the wildland-urban interface (WUI) pose a greater threat to human and economic activity, create greater economic losses, and are usually more aggressively suppressed than fires occurring in areas with no human settlement.[2] Fires in the WUI can be fueled both by vegetation (natural or cultivated) and by human-built structures.

[1] In contrast, structural building fires or vehicle fires are not critical to the healthy function of any ecosystem and occur typically in urban environments.

[2] The U.S. Fire Administration (2022) defines the wildland-urban interface as "the zone of transition between unoccupied land and human development. It is the line, area or zone where structures and other human development meet or intermingle with undeveloped wildland or vegetative fuels."

M. Kelly (✉)
Lazaridis School of Business & Economics, Wilfrid Laurier University, Waterloo, ON, Canada
e-mail: mkelly@wlu.ca

Z. Peng
Property and Casualty Insurance Compensation Corporation (PACICC), Toronto, ON, Canada
e-mail: jpeng@pacicc.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_5

This combination can create fire conditions that are uniquely different than an isolated wildfire or structural fire (FireSmart Canada 2023).

Fire was the initial peril covered by property insurance contracts, and the property-casualty insurance industry has been providing protection to property owners against fire losses for over 400 years. In Sect. 2, we provide a brief history of fire insurance and elaborate on the coverage for wildfire losses and the depth of insurance market penetration for wildfire losses globally.

Wildfires can cause significant human fatalities and injuries, agricultural and property losses, and associated indirect revenue losses. These losses are discussed in Sect. 3. Wildfire creates substantial agricultural losses, including loss of crops, timber, and farming equipment. Structures and infrastructure can also be damaged, as wildfire has the potential to disrupt transportation, telecommunications, energy, and water supply. Wildfires often produce significant revenue losses, including tourism losses, business interruption losses, a decrease in property values, and a potential erosion of the tax base. In addition to the fire itself, wildfires also produce significant pollution, posing threats to both physical and mental health.

There are tools to manage wildfire risk. In addition to suppressing wildfire, community and public fire mitigation and prevention aims to reduce the frequency of ignition, manage the fuels available to burn, and undertake land use planning to protect people and property. Property owners also have a role to play in creating defensible spaces and hardening homes within the WUI. There are also postfire actions to reduce future flood risk that may follow from wildfire. These activities are discussed in Sect. 4.

Increasing wildfire risk is largely propelled by climate change, land use, land management practices, and demographics (UNEP 2022). Both economic and insured losses from wildfires are on the rise and are expected to grow in the future (Munich Re 2019). Drivers of growing wildfire risk and empirical evidence of this growth are presented in Sect. 5.

The growth in wildfire risk raised considerable insurability concerns, especially in California and Australia. Private insurers have a role to play in providing tools to property owners to reduce wildfire risk. Insurance market solutions, which leverage the capital markets and potentially involve governments, are also available and are yet to reach their potential in enhancing the availability and affordability of coverage for wildfire losses. These topics are discussed in Sect. 6.

A brief conclusion is given in Sect. 7. As the science of wildfire may be unfamiliar to many insurance academics, who may have better knowledge of structural fire, we provide a concise primer on global wildfire risk in Appendix A.

2 Insuring Wildfire Losses

Commercial fire insurance, of which structural fire was the original insured peril, has existed for over 400 years. We briefly outline the history of fire insurance and then focus specifically on coverage for wildfire losses before providing a brief global view on insurance market penetration for wildfire losses.

2.1 A Brief History of Fire Insurance

Fire insurance—of which the initial focus was structural fire—started in Europe, where records of its operations date back to Hamburg, Germany, in 1591 (Evans 1987). The 1666 Great Fire of London kindled the awareness of and a private market for fire insurance in the United Kingdom, after similar attempts were aborted due to the 1642–1646 Civil War (Carlson 2005). In 1680, the Fire Office in

London became one of the first insurers to offer coverage for property damage from fire, and their contracts contained a clause granting insurers the right to rebuild the damaged structure (Rosevear 1932). In 1718, fire insurance was offered in Scotland by the Friendly Society (Evans 1987).

Meanwhile, private insurers actively engaged in fire prevention and firefighting by running fire brigades, as both measures helped reduce the claims payouts. As the British Empire grew, British companies expanded from offering coverage in London to providing protection in its colonies and foreign territories, based on the same principles that underpinned the original British coverage.

2.2 Insurance for Losses from Wildfire

Although typically insurers no longer provide firefighting services, modern insurance coverage against fire losses went beyond the original coverage.[3] Today, fire risk is unique among all perils insured: in many jurisdictions, property losses from fire are insured even if the peril that created the fire (e.g., terrorist attack or earthquake) is excluded from the underlying contract.[4] Despite increases in wildfire risk, the current ability of property owners to purchase insurance in the voluntary market has not been impacted in most jurisdictions.[5] In general, there is greater penetration of property insurance for commercial structures than for residential properties because coverage is mandatory in many jurisdictions, and premiums are tax deductible as an allowable business expense.

Unlike structural fire, wildfire can cause significant agricultural losses, which will be discussed in Sect. 3.1. In North America and the EU, government-level crop insurance programs stabilize farm revenues by protecting against declines in crop yields and/or revenues caused by multiple perils, including fire. Similar to standard property coverages, even if fire ignition was caused by an uninsured peril, the resulting losses from fire would still be covered. In other jurisdictions, such as Australia, private insurers compete to offer single peril coverages for broadacre crops. Agricultural operations can purchase coverage for hail or fire, with the majority of claims relating to hail losses.

Insurance may also indemnify timber losses due to wildfires. In North America, New Zealand, Australia, and the EU, some insurers offer forestry insurance for private forests, which covers the destruction of trees from perils such as fire, hail, and earthquake, and also provides indemnification for replanting costs, firefighting expenses, debris removal, and plantation infrastructure.[6] The level of insurance penetration varies by country. Brunette and Couture (2023) noted that 95% of private forests are insured in Sweden, but less than 4% of forests are insured in France, Spain, and Germany.

Similar to losses arising from other catastrophes, not all damage caused by wildfires is covered by insurance. For example, infrastructure damage, such as damage to transportation, power and gas services, and drinking water sources, is also common in wildfires. In most developed and developing nations, public infrastructure may not be insured, and repairs to infrastructure are funded by taxation and potentially by international relief such as the EU Solidarity Fund. Longer-term significant revenue losses, such as declines in housing prices, losses due to highway closures, and decreased tax base, are also not covered by insurance.

[3] In response to catastrophic wildfire losses in high-value neighbourhoods in the United States and Canada, some insurers offer prefire and postfire mitigation services and structure protection services to high-net-worth policyholders.

[4] One exception is damage to the insured property caused by contamination by radioactive material resulting from fire, which is a typical fire exclusion. Similarly, any damage intentionally caused by the insured, e.g., arson, is not insurable either.

[5] California is a significant exception to this statement. Associated Press (2023) reported that seven of the top 12 insurers have either paused or restricted new business in the state and that eight companies had requested rate increases exceeding 20%. Insurability issues in California are discussed in Sect. 6.1.

[6] Other perils, such as insect damage and carbon losses, may also be insurable (Brunette and Couture 2023).

2.3 Insurability of Wildfire Losses

When discussing insurance coverage for potentially catastrophic events, such as wildfire, the concept of insurability is of critical importance. Private insurance works most efficiently when the following conditions, related to the concept of insurability, are met.

1. The event must be random. The timing and location of the event must be unpredictable and independent of the behaviour of the insured.
2. The event must be assessable. The likelihood and severity of the event must be estimable and quantifiable before the event occurs, allowing for the setting of a premium.
3. The insurance solution (both premium and coverage provision) must be economically viable for both the insured and the insurer.[7]

Depending on the underlying catastrophic event, all three conditions may be violated. Although the behaviour of the insured does not cause a natural disaster, it can be argued that the timing and location of some catastrophic events are fairly predictable (e.g., flooding on flood plains). For some catastrophic events (e.g., geomagnetic storms), the second criterion may not be satisfied—there may not be enough historical information to estimate the size or the impact of an occurrence. Advancements in the modeling sophistication of catastrophes have alleviated such concerns somewhat, but improved estimation may also result in significant premium increases for some risks, potentially violating the criteria for economic viability. The accumulation of losses may be so great that insurers cannot afford to bear that risk, making coverage economically unviable in the private market.

Even if coverage is affordable and readily available, because insurance contracts were originally designed for idiosyncratic losses, policy provisions that work well for idiosyncratic events (i.e., structural fire) may not provide sufficient coverage for losses arising from catastrophic events (i.e., wildfire). We discuss three specific issues below.

In some jurisdictions, such as in North America and most jurisdictions in the EU (Switzerland being the exception), residential coverages for dwellings and contents are bundled together. Since the amounts payable for contents, external structures, and landscaping are tied to the amount of insurance purchased for the dwelling, it is possible that the maximum amounts payable may not be sufficient to replace external structures and landscaping destroyed by wildfire. There may also be exclusions in what is covered in a standard residential policy. After the 2019/2020 bushfire season in Australia, individual rural landowners were responsible for repairs and replacement of underground infrastructure for water storage and lines, electricity cables between outbuildings, and sewage lines. These replacement costs, not covered by residential insurance policies, were borne by landowners (Australia, The Senate 2021).

In many jurisdictions, such as in North America and Australia, residential policies offer coverage for additional living expenses incurred during both evacuation and the subsequent rebuilding process.[8] However, there is usually an upper limit on the length of time that coverage is available, and in the event of widespread devastation, the policy limit could be easily breached. For example, after the 2019/2020 bushfires in Australia, many individuals were displaced from their houses for more than 1 year, but most insurance contracts limited the insurer's liability to pay for temporary accommodations to 12 months (Australia, The Senate 2021). In recognition of delays faced when trying to rebuild after

[7] This definition of insurability is from Swiss Re (2017). While other sources define these characteristics slightly differently, the overall requirements are consistent.

[8] Furthermore, additional living expenses during evacuation will be covered, even if the underlying property suffers no damage.

a disaster, in 2020, legislation was passed in California (California Insurance Code Section 2060 (b)(1)) requiring insurers to provide for up to 36 months of additional living expenses in the event of a covered loss related to a state of emergency. Concerns also exist that the wording in policies may lead to an exclusion in coverage if the dwelling is habitable, despite the infrastructure required for habitation not yet being restored.[9]

Many of the concerns related to policy provisions in residential property coverages may also apply to commercial property coverages, and in particular, coverage provided for business interruption losses may not be sufficient after a catastrophic loss. Moreover, revenue lost due to evacuation is typically excluded from business interruption coverage, which requires damage to the insured property, or nearby properties. Nevertheless, many commercial policies have an extension for civil authority coverage that would cover the loss of business income, rental value, and extra expenses incurred due to denied access or forced evacuation.

Contract provisions, interestingly, can be at odds with Build Back Better options. Build Back Better is a postdisaster recovery approach to reduce vulnerability of communities to future disasters. The United Nations Office for Disaster Risk Reduction defines Build Back Better as "the use of the recovery, rehabilitation, and reconstruction phases after a disaster to increase the resilience of nations and communities through integrating disaster risk reduction measures into the restoration of physical infrastructure and societal systems, and into the revitalization of livelihoods, economies, and the environment".[10] As such, the key tenet of Build Back Better is to rebuild stronger, safer, and more disaster-resilient structures, infrastructure, and economies. Although replacement cost coverage provisions in insurance contracts provide for building back to the original condition of the property, additional expenses associated with Build Back Better mitigation measures, such as flame-resistant shingles, are typically not covered. Similarly, "same site" requirements, in which coverage only applies to rebuilding in the same location, may conflict with land use planning regulations if the location is now deemed to be high-risk.

2.4 Comparing Insurance Coverage of Wildfire Losses to Other Catastrophic Losses

Market penetration of private insurance for wildfire coverage is substantially higher than coverage for other natural disaster losses. And insurance penetration globally is the greatest with 63% of losses covered by insurance (OECD 2021).[11] Perhaps this is not surprising: unlike earthquake and flood, fire, and wind are not excluded perils in standard residential and commercial property coverages.

However, the proportion of wildfire losses covered by insurance varies significantly across jurisdictions. According to the OECD (2021), the share of insured losses for wildfires is less than 20% in many countries, such as Greece (9%) and Portugal (10%). In countries that have experienced substantial wildfire losses in the past, insurance penetration is notably higher: Canada (66%), the United States (71%), and Australia (74%). Insurance penetration for residential properties is also a function of whether there exists a mandatory offer or purchase provision within the jurisdiction (such

[9] This is of particular concern in California. In July 2021, Senate Bill 872 instructed insurers that they could not restrict additional living expenses if a home is rendered uninhabitable due to damaged power or water lines, even if the damage is not to the property itself (California Department of Insurance 2021a).

[10] Other governments and organizations have similar definitions.

[11] Over the same time period, 18% of earthquake losses and 29% of flood-related losses were insured.

as New Zealand or Belgium) or if mortgage lenders require proof of insurance before extending a mortgage loan (such as in North America, many countries in the EU, and Australia).[12]

Globally, between 1990 and 2019, economic losses from wildfire are significantly smaller than economic losses arising from earthquake, storm, or flood. Because of the relatively small size (to date) of wildfire losses, in the majority of jurisdictions worldwide, the risk of wildfire losses satisfies insurability criteria and thus is insured entirely within the private insurance market.[13] Correspondingly, most government interventions that support the provision of insurance for catastrophic losses—either in the primary insurance or reinsurance market—have not included or considered wildfire losses.[14] In Spain, wildfire is not considered an extraordinary peril, and losses arising from wildfires are covered by private insurers and not by Consorcio de Compensación de Seguros (CCS), the government insurer of extraordinary risks.[15] A similar situation exists in France, where the public-private Nat Cat scheme provides broad coverage against many natural disasters (flood, earthquake, tsunami, landslide, mudslide, avalanche, subsidence, and high winds) but excludes wildfire.

Although many jurisdictions have separate residual pools for high-risk properties (such as flood coverage for simple risks in Belgium, storm, and flood risk in many states in the United States, and earthquake in Türkiye), there are no separate high-risk pools specific for wildfire risk. In both Canada and Australia, where some communities face significant wildfire risk, there is no residual or involuntary market for high wildfire risk properties.[16] In the United States, FAIR plans provide coverage for owner or tenant occupied dwellings, commercial buildings, including wineries and farms, but typically exclude coverage for crops and livestock that cannot purchase insurance from the private market.[17] In California, Liao et al. (2022) reported that the FAIR plan market share of insureds in the highest wildfire risk zip codes has increased steadily between 2006 and 2018, whereas market shares in lower risk zip codes have remained relatively flat or have declined. A further and more detailed discussion on wildfire insurability concerns in both California and Australia is provided in Sects. 6.1 and 6.2.

As with all catastrophic events, the expectation of ex post disaster assistance reduces the demand for private insurance. Browne and Hoyt (2000) refered to this phenomenon as "charity hazard".[18] Several researchers (see, for example, Kousky 2017, Kousky et al. 2018, and Davlasheridze and Miao

[12] In contrast, in China, where residential insurance is not a mortgage requirement, residential insurance accounted for 0.7% of property-casualty premiums and 0.01% of China's GDP (Zhang et al. 2020).

[13] The one exception is Switzerland, where, in 19 of 26 cantons, a cantonal monopoly insurance company provides coverage for buildings (although in most cantons, private insurers compete to provide coverage for contents).

[14] Instead, most government interventions focus on earthquake/earth movement losses (such as Natural Catastrophe Insurance of Iceland and New Zealand's Toka Tū Ake EQC) or flood (National Flood Insurance Program in the United States and Flood Re in the United Kingdom).

[15] The CCS is the sole provider of coverage against extraordinary risks (extraordinary floods, earthquakes, tidal waves, volcanic eruptions, atypical cyclones and the falling of astral bodies and meteorites, violence, and incidents caused by the armed forces). The CCS, however, does cover health costs for individuals injured while working to suppress wildfires. Given the early arrival of the wildfire season in Spain (https://www.euronews.com/green/2023/03/27/the-era-of-mega-forest-fires-has-begun-in-spain-is-climate-change-to-blame), it remains to be seen if wildfire will be defined as an extraordinary risk.

[16] The Australian Cyclone Reinsurance Pool commenced operations in 2022 to provide reinsurance for primary insurers that write residential and small commercial business insurance, primarily in northern Australia. It only covers cyclone and related flood damage. Similarly, in Canada, in 2022, the federal government announced that it would create a flood insurance program for high-risk properties and allocated 23.56 million USD to the development of the program in 2023. Whether the two programs will be expanded to offer wildfire coverage is unknown.

[17] FAIR plan is a named peril policy and covers losses from fire, lightning, and explosion, although additional perils may be added for an additional cost.

[18] The expectation of ex post government assistance also reduces the amount spent on mitigation (Kelly and Kleffner 2003).

2019) found that, in the United States, the amount of residential flood insurance purchased decreased when counties received public assistance grants from the government following a flood event. Similar outcomes were noted in Austria and Germany by Raschky et al. (2013). The presence of government relief programs reduced the demand for flood insurance with a stronger impact in Austria, which had guaranteed partial relief, than in Germany, which had complete but uncertain funding.

However, interventions, such as subsidized insurance coverage and mitigation activities that would further reduce premiums, could increase the demand for coverage. Specifically related to fire coverage, Sauter et al. (2016) concluded that while public support reduced the willingness to pay for fire insurance, suggesting the existence of charity hazard, subsidized insurance premiums did increase the willingness to pay, implying an increased demand for coverage.

3 Losses Arising from Wildfires

The largest fires, in terms of acreage burned, are in the boreal forests, where many of these fires propagate with no human intervention. The costliest fires, measured by insured losses, as shown in Table 1, occur in regions with a substantial WUI.

Total economic losses (from 2000 to 2022) and insured losses (from 2000 to 2021) highlighting jurisdictions with significant fire risk are given in the first two panels in Fig. 1, and the total number of people injured killed or left homeless is given in the last graph in this figure. Compared to fatalities arising from other natural disasters, since the advent of modern firefighting, direct fatalities from wildfires are much less significant.[19] Globally, between 2000 and 2022, Ritchie et al. (2022) record roughly 1800 direct fatalities from wildfire.

Direct losses arising from wildfires include physical and mental health impacts, damage to structures, infrastructures, timber and agricultural production, and short-term environmental impact. Indirect losses include short-term and long-term economic and supply chain impacts, disruptions to utilities, transportation and government services, potential long-term housing market impacts, decreased tax base, loss of ecosystem services, and increased risk of future hazards (particularly flood and debris slide). Thomas et al. (2017) noted that for the United States, although the losses generated by wildfires are known to be substantial, they are largely undocumented and unknown at a national level. They estimated the annual economic losses from wildfire to range from 63.5 billion to 285 billion USD, and the total cost (which includes intervention and mitigation) to be between 71.1

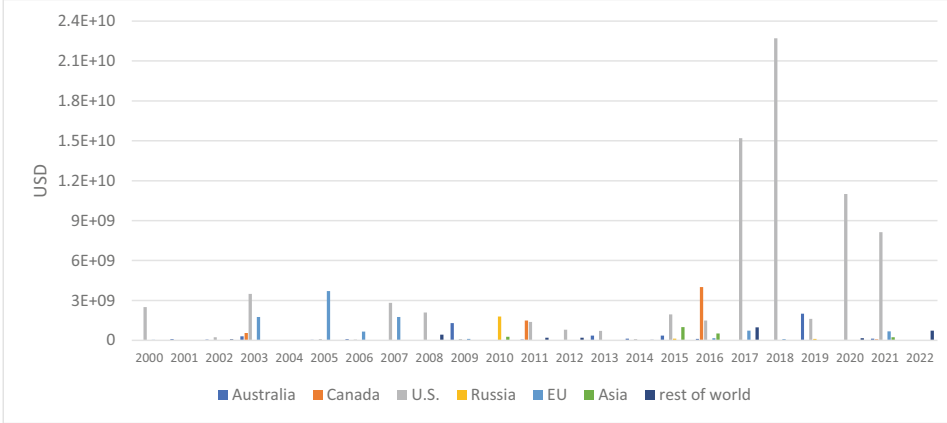
Table 1 Largest wildfires by insured losses

Fire	Location	Date	Insured losses (billions 2022 USD)
Camp fire	California, U.S.	November 2018	11.8
Tubbs fire	California, U.S.	October 2017	10.5
Woolsey fire	California, U.S.	November 2018	5.0
Oakland tunnel fire	California, U.S.	October 1991	3.7
Atlas fire	California, U.S.	October 2017	3.6
Horse Creek fire	Alberta, Canada	May 2016	3.5

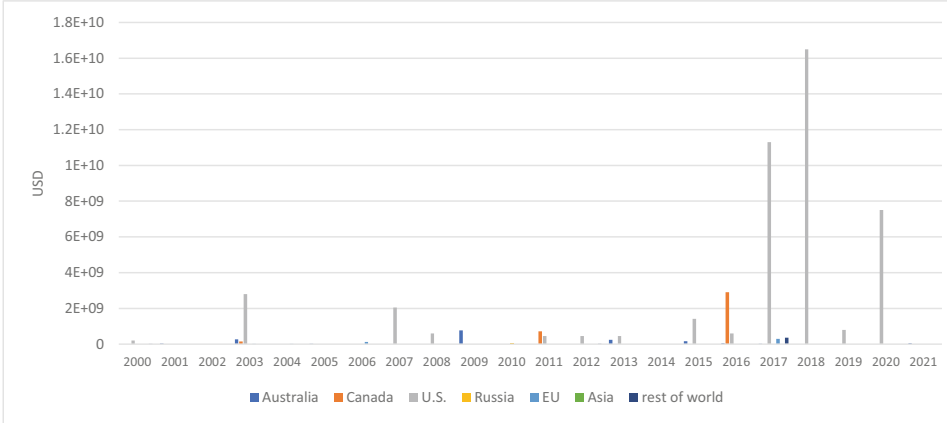
Source: Insurance Information Institute (2023a)

[19] More recent fires that have caused considerable fatalities include the Black Dragon Fire in China in May and June 1987 and Indonesian wildfires between May 1997 and March 1998, causing 191 and 240 direct deaths, respectively, and the 2009 Black Saturday fire in Australia, which caused 173 deaths (Cameron et al. 2009). More recently, wildfires in Hawaii have killed at least 97 people (Sinco Kelleher and Boone 2023).

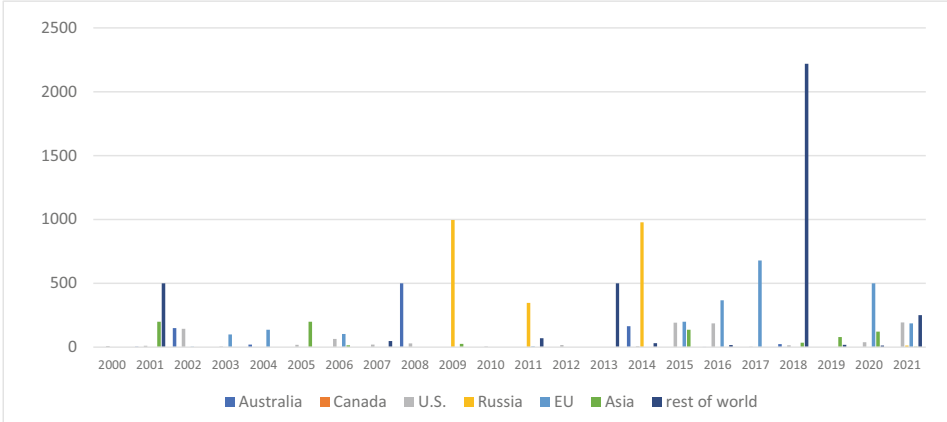
Panel A. Economic losses from wildfires for selected regions



Panel B. Insured losses from wildfires for selected regions



Panel C. Number of people injured, killed, or displaced for selected regions



Source: Ritchie et al. (2022)

Notes: In Panels A and B, all values unadjusted for inflation.

Fig. 1 Cost of wildfires globally

billion and 347.8 billion USD. Howard (2014) suggested that the total cost to the United States is between 20 and 125 billion USD, but climate change would increase these costs by 10 to 62.5 billion USD annually by 2050 (equivalent to 0.06% to 0.36% of projected U.S. GDP).

Although some estimates exist for some losses for some events, the true cost of loss from wildfires globally (similar to other natural disasters) is difficult to calculate.[20] Distinct from other natural disasters, the pollution generated by wildfires results in chronic illnesses in individuals and contributes to climate change. These costs are long term in nature and are very difficult to track.

Below we highlight some of the costs that have been calculated in the literature for specific fire events. In particular, we examine agricultural and timber losses, losses to structures and infrastructure, direct and indirect losses to tourism, and other revenue losses associated with wildfire. This is followed by a summary of the literature related to the direct and indirect health impacts of wildfires. We conclude with a discussion of the environmental impact of wildfires, highlighting the relationship between wildfires and postfire flooding.

3.1 Direct Agriculture and Timber Losses

Wildfires bring about agricultural property losses, including losses of crops, livestock, farm structures (e.g., fences), and machinery. The timber industry and associated sectors such as pulp and paper and construction are especially vulnerable to wildfire losses. Not only is timber burned, but logging infrastructure can also be damaged and destroyed. Pulp and paper and construction industries can also be impacted due to a reduction in available wood and wood products.[21] The total agricultural losses arising from specific fire events appear sporadically in news reports, however, as Thomas et al. (2017) noted, systematic estimates of both agricultural and timber losses remain unavailable for the United States. The same is true for other jurisdictions. When considering crop and livestock production only, wildfire was responsible for 1 billion USD of losses between 2008 and 2018 (FAO 2021).[22] As systematic estimates are not available, we highlight a few of the significant fire losses herein.

The 1997 Indonesian fire scorched 6.18 million acres of land designated as plantations/agriculture and 2.47 million acres of forest, causing an agricultural loss of 470.4 million USD (881.81 million in 2023 USD), a loss of timber valued at 493.7 million USD (925 million in 2023 USD), and losses of direct forest benefits (food, local medicine, and other raw materials) valued at 705 million USD (1321 million in 2023 USD) (Pettenella et al. 2022). Among the hardest hit crops were coffee and palm oil (Mongabay.com 2020). Seven years later, between June and November 2004, wildfires again burned across Indonesia, resulting in losses of 16.1 billion USD (20.5 billion in 2023 USD). Over half of the losses were sustained by the agricultural and forestry sectors (World Bank 2016). Similarly destructive were the 2022 Corrientes wildfires in Argentina, which destroyed yerba maté and rice crops jointly worth 67.8 million USD. The wildfires also caused farming operation losses, estimated to exceed 551 million USD (OCHA 2022).

Even if crops are not destroyed in wildfires, postfire yield and quality may deteriorate. After the 2017 wildfires in California, winemakers found that grapes exposed to smoke absorbed chemicals that

[20] Pettenella et al. (2022) summarized studies of wildfire losses in Costa Rica, Indonesia, Türkiye, and Italy, in which attempts were made to assess the socio-economic damages from forest wildfires.

[21] The 2009 Australian Black Saturday Fires destroyed 47.4 million USD (66.7 million in 2023 USD) worth of timber, and the 2017 wildfires in British Columbia, Canada destroyed 1 year's worth of allowable timber production (Bailey and Yeo 2019).

[22] Based on crop and livestock production only, wildfire is the least impactful of all catastrophes, being responsible for 1% of all losses to crops and livestock. The top three disasters impacting crop and livestock production systems—drought, flooding, and storms—account for 71% of all agricultural production losses (FAO 2021).

altered, if not ruined, the taste and smell of wine (Madhusoodanan 2021). The estimated economic loss was valued at 75 million USD (87 million in 2023 USD) (California State Assembly 2020). In 2020, after another year of devastating wildfires, over 325,000 tons of winegrapes valued at 600 million USD (697 million in 2023 USD) were not harvested due to concerns about wildfire smoke contamination. As a result, the U.S. Department of Agriculture paid over 200 million USD (232 million in 2023 USD) in claims to grape growers (Lee 2021).

Wine production in Australia after the 2019/2020 bushfires was similarly impacted. The devasting bushfires also destroyed over 7.41 million acres of conservation land, 3.7 million acres of plantations and forests, and 2 million acres of farmland. Additionally, more than 63,000 sheep and 8400 cattle perished in the fires (Wittwer and Waschik 2021).

3.2 Direct Structure and Infrastructure Losses

Compared with structural fire, of which property losses and potential injury or loss of life are contained within a tight radius, wildfires have broader-ranging effects. They have the potential to disrupt transportation, telecommunications, energy, and water supply. Table 2 lists some of the largest property losses, in terms of structures destroyed, from recent wildfires.

Between 2005 and 2022, over 97,156 structures were destroyed by 2475 wildfires across the United States (Headwater Economics 2022). Although there is no comprehensive database that tracks the number of structures burned, the damage to infrastructure or the value of the property damaged in the United States (Thomas et al. 2017), St. Denis et al. (2023) have created a database using United States government data that documents the acreage burned, suppression costs, and structures threatened and destroyed of 34,478 wildfires between 1999 and 2020.

3.3 Direct and Indirect Tourism Losses

A significant economic loss from wildfires is the loss in tourism dollars. Wildfires can have a significant impact on tourism, from a degradation of air quality due to smoke in surrounding regions, flight delays, and cancellations at airports adjacent to wildfires, and from direct fire losses that often impact outdoor recreational areas. Until the burned region is rehabilitated, there can be ongoing loss of tourism and recreational dollars. For example, in June 2010, the Schultz fire in Arizona damaged no structures but burned 15,000 acres of land (although, as with other wildfires, there is a wide range in estimates of the acreage burned) and injured two firefighters. It led to the closure of parts of the Coconino National Forest for the entire summer (Syed 2021). Similar to other losses, no comprehensive assessment of the overall tourism losses is available, and Thomas et al. (2017) noted that any existing studies cannot be scaled up to give such a metric. Estimates of tourism losses for selected wildfires are given in Table 3.

3.4 Other Indirect Losses

As with other natural catastrophes, the time to rebuild lost structures is significantly longer than with the replacement of property after idiosyncratic losses. This results in revenue losses, as commercial enterprises are unable to be operational immediately after a wildfire, which may result in short- and long-term losses of customer and/or employee base. For households, indirect losses arise when

Table 2 Property losses from wildfires

Year	Name and location	Acreage burned (thousands)	Total insured loss USD billion (2023 USD)	Damaged or destroyed structures	Source
1991	Oakland tunnel fire, CA	1.6	1.7 (3.77)	2900 damaged or destroyed structures	Cal Fire (2019)
2003	Old, cedar and 12 other fires, CA	273.2	3.4 (5.34)	3710 homes	Cohen and Westhaver (2022)
2016	Chimney Tops 2 fire, TN	11.4	2.0 (2.48)	2000 homes	Cohen and Westhaver (2022)
2016	Horse River fire, Alberta	1547	2.9 (3.64)	3244 damaged or destroyed structures	CatIQ database
2017	Thomas fire, CA	281.9	2.5 (3.08)	1334 damaged or destroyed structures	Cal Fire (2019) Van Leer (2018)
2017	Tubbs fire, CA	36.8	9.4 (11.54)	5643 damaged or destroyed structures	Cal Fire (2019)
2018	Woolsey fire, CA	96.9	4 (4.79)	1643 damaged or destroyed structures	Cal Fire (2019)
2018	Camp fire, CA	153	12.5 (14.98)	18,804 damaged or destroyed structures	Cal Fire (2019)
2019 / 2020	Australian bush fires	46,000	1.44 (1.63)	8800 damaged or destroyed buildings,	Wittwer and Waschik (2021)
2020	North Complex fire, CA	318.9		2342 damaged or destroyed structures	National Interagency Fire Center (2020)
2020	Glass fire, CA	67.5	2.95 (3.44)	1520 damaged or destroyed structures	National Interagency Fire Center (2020)

individuals require temporary accommodations either due to evacuation or the time required to rebuild damaged properties and infrastructure.[23]

Thomas et al. (2017) observed that there could be permanent changes in the population base as people may permanently move away after a disaster.[24] This has long-term impacts on the economic stability of a region and the tax base of local governments. Other studies (e.g., Donovan et al. 2007; Mueller et al. 2009) point to decreasing housing prices after occurrences of wildfires or when information transparency on wildfire risks increases.

[23] Even if a property is not damaged, failed infrastructure such as the lack of sewage, telecommunications, electricity, or water services could delay return to both commercial and residential properties.

[24] For example, before hurricanes Katrina, Rita, and Wilma wreaked havoc in 2005, the population of New Orleans was 455,000. In late 2007, the population was estimated to be 320,000 (Sastry 2009). Although poverty, crime, and a lack of basic services were cited to be contributing reasons to population decline in New Orleans, it has been conjectured that a deficiency of opportunities within the state combined with better opportunities found during displacement was a driving factor in the persistent population decline (Paxson and Rouse 2008).

Table 3 Tourism losses from wildfires

Year	Location	Acreage burned (millions)	Tourism loss estimated cost original USD (2023 USD)	Other reported loss metrics	Source
1997	Indonesia	12.4[a]	70.4 million (1.3 billion)		Pettenella et al. (2022)
1998	Florida, United States	0.5	138 million (255.5 million)	Reduced tourism caused estimated loss of $61 million in hotel revenues and an estimated $77.2 million in other tourism-related spending	Diaz (2012) Butry et al. (2001)
2003	San Diego, United States	0.376	32.5 million (53.3 million)		Diaz (2012)
2003	Victoria, Australia	2.71	20 million (32.8 million)	Around 1000 tourism businesses affected or closed	Cioccio and Michael (2007)
2015	Indonesia		399 million (479.6 million)		World Bank (2016)
2017	North Bay region, California, United States	1.25		Reduction in inbound passenger activity of 6000–8000 passengers; >5% decline in hotel tax collected	Bay Area Council Economic Institute (2021)
2017	Kootenay area, B.C., Canada	0.09[b]	29.6 million (36.5 million)		Kootenay Rockies Tourism (2018)
2019/2020	Australia	13	3.46 billion (4.04 billion)	Caused 60% of the booking cancellations in non-fire zones	Thiessen (2020)

[a]Official government estimates say that 1.85 million acres were destroyed, but environmental organizations say the amount is significantly higher (https://rainforests.mongabay.com/08indo_fires.htm)
[b]Acreage burned in the Southeast Fire Centre and the adjacent Kamloops Fire Centre combined

Although direct fatalities from wildfire have declined with the advent of modern firefighting, the magnitude and severity of fires and the resultant financial impact have been increasing dramatically over the last 40 years. Swiss Re (2023) noted that insured losses from wildfires have doubled between 1980 and 2022, with particularly large economic and insured losses in Alberta, Canada (Horse River fire in 2016) and in California (Tubbs, Atlas, and Thomas fires in 2017, Camp and Woolsey fires in 2018, and the Glass, CZU Lightning Complex and LNU Lightning Complex fires in 2020). Between 1981 and 2015, wildfires accounted for roughly 1.3% of insured natural catastrophe losses, and this percentage increased to 12.5% for 2016–2020 (Swiss Re 2021).[25]

[25] Wildfires are usually named for the geographic location or geographic features where the wildfire started. However, other names are often used to refer to significant fires. The Horse River fire, for example, is also known as the Fort McMurray wildfire or the Wood Buffalo fire.

3.5 Direct and Indirect Health Impacts of Wildfires

Despite declining direct fatalities, wildfires are associated with significant health impacts, both physical and mental. As the IPCC (2022 p. 11) has concluded, physical health issues are mostly related to exposure and inhalation of "wildfire smoke, atmospheric dust, and aeroallergens". This pollution not only impacts first responders and local residents, but wildfire smoke can be carried in the atmosphere for hundreds or thousands of miles, impacting populations far away from the wildfire. For example, wildfire smoke from the 2016 Horse River fire in Canada traveled as far as the United Kingdom (Vaughan et al. 2018).

An increase in smoke exposure is associated with a significant increase in overall mortality, and it is also associated with a surge in physician visits, emergency department visits, and hospitalization due to respiratory ailments such as asthma, chronic obstructive pulmonary disease (COPD), respiratory infections, pneumonia and bronchitis, and cardiac arrest (Reid et al. 2016). With respect to children, although there is insufficient evidence on the long-term impacts of wildfire smoke exposure on mortality, studies have concluded that children also experience significant short-term health impacts from smoke exposure (Gao et al. 2023).

Pollution (fine particulate matter) from wildfire smoke is "more toxic than equal doses from ambient PM 2.5" (Aguilera et al. 2021 p. 1). Exposure to these particulates exacerbates respiratory ailments such as asthma and chronic obstructive pulmonary disease (COPD) (Künzli et al. 2006; Reid and Maestas 2019), and not surprisingly, increased prescription usage for respiratory ailments (Dhingra et al. 2023). For example, in 2015, fires in Borneo resulted in an air quality index of 2000, causing one-half million people to develop significant respiratory illness, resulting in the premature deaths of 100,000 people (Sever 2020).[26] Wildfire smoke from boreal fires is particularly dangerous, as smoke from peat fires yields a higher level of fine particulate matter (Ikeda and Tanimoto 2015). The 2010 heat wave and wildfires in Moscow were estimated to be responsible for 11,000 deaths in 6 weeks. The elderly population was disproportionately impacted, with the majority of deaths arising from cardiovascular diseases (Shaposhnikov et al. 2014).

The resultant health costs associated with wildfires and wildfire smoke are not insignificant. Using data from the 2009 Station Fire in California, estimates suggest that health costs per individual ranged between 87 and 95 USD per symptom day (122 and 133 in 2023 USD) (Richardson et al. 2013). In another study, increased hospital admissions from wildfires in California in 2018 resulted in an estimated 7.8 billion USD (9.38 billion in 2023 USD) in healthcare costs across the Bay Area (Bay Area Council Economic Institute 2021). Similarly, the health costs associated with the 2019/2020 Australian bushfires were estimated to be 1.33 billion USD (1.07 billion in 2023 USD) (Hitch 2020). In Indonesia, the immediate direct health costs incurred in 2015 amounted to 151 USD million (192.3 million in 2023 USD) (World Bank 2016).

The costs attributed to wildfire smoke exposure alone are also significant. Kochi et al. (2010) calculated that the total cost of health effects from wildfire smoke exposure ranges from 200,000 USD to 1.2 billion USD (275,929 to 1.66 billion in 2023 USD) per fire. In the United States, long-term exposure to wildfire smoke is estimated to cost 76–130 billion USD annually (Bailey and Yeo 2019).

Fire suppression itself also causes casualties and fatalities. Although no global estimates exist on injuries, data on wildland firefighter fatalities have been tabulated by several sources. The Centers for Disease Control (2022) calculated over 400 wildland firefighter deaths in the United States between 2000 and 2019, and a further 19 wildland firefighters died in 2020 and 2021. Alexander

[26] An air quality index value below 50 is considered good, and values above 300 create hazardous conditions for everyone.

and Buxton-Carr (2011) tabulated 165 wildland fire suppression-related fatalities in Canada between 1941 and 2010. In 2023, Canada's worst wildfire season ever, news outlets reported at least four firefighter deaths by the end of July (see, e.g., Aziz 2023). In Australia, between 2010 and 2020, 30 firefighters died during firefighting efforts, and another nine died enroute to fight bushfires (Coates 2020). Firefighting fatalities are more common in developing nations due to a lack of training and equipment. In 2019, 30 firefighters died while suppressing a fire in Muli County, China (Zhang 2019).

In addition to physical ailments, mental health impacts arising from wildfires are also significant. Eisenman and Galway (2022) reviewed 19 articles examining how wildfire smoke specifically affects mental health and well-being. Although this area of research is in its infancy, the study identified potential pathways in which mental health and well-being may be impacted. These diverse impacts included loss of nature, evacuation and relocation, isolation from others, sleep disturbances, food insecurity, physical symptoms, and reduced physical activity.

After the 2016 Horse River fire in Canada, contacts related to mental health climbed from 1000 annually to 44,000 between May 2016 and January 2018 (Beamish 2018). Felix and Afifi (2015) surveyed residents who survived three consecutive wildfires in California between July 2008 and May 2009, and found that repeated evacuation experiences triggered greater stress levels.

After a wildfire, victims may develop post-traumatic stress disorder (PTSD), depression, or other anxiety disorders. PTSD has been studied extensively in the literature, and civilian rates of PTSD after a disaster range from 20% to 60% among those impacted by the disaster, and rates of PTSD for firefighters range from 13% to 20%. Using estimates in the United States for the cost of treating PTSD for military personnel, Thomas et al. (2017) calculated that the annual cost for treating PTSD in firefighters is 30.17 million in 2023 USD. Because there are no estimates on the number of civilians exposed to wildfires and given the wide range of the percentage of civilians that could potentially suffer from PTSD, it is not possible to calculate a similar cost for civilians.

3.6 Direct Environmental Impact and Increased Future Flood Risk

The first immediate loss from wildfires is the damage to the impacted land, changing its vegetation, drainage system, and topology (Shakesby 2011). Wildfires force wild animals to flee their natural habitat, potentially increasing contact with domesticated animals and humans. This could lead to increased risk of zoonoses like COVID-19 (Kemper 2020; Albery et al. 2021). Although there is no accurate assessment of the number of animals that die in wildfires annually (Garcês and Pires 2023), it is estimated that the Australian bushfires of 2019/2020 killed or displaced nearly 3 billion animals (Australia, The Senate 2021). Albery et al. (2021) observed that slower-moving species (like turtles) and very young and elderly animals are most likely to die in wildfires. The loss of nesting and breeding areas, shelter, and food sources have longer-term impacts on the ability of wildlife to survive and reproduce. Aquatic animals may also be impacted as wildfires can raise water temperature, and toxins can settle in bodies of water nearby.

Burned biomass—including vegetation—also results in the mass release of greenhouse gases, notably carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), among others (Urbanski 2014). These emissions create a feedback loop, where more gases contribute to higher temperature, which in turn, contributes to more wildfires thus increasing greenhouse gases. This poses a significant challenge to the "zero-carbon" initiative aimed to combat global warming. Emissions due to wildfires vary vastly by time, the type of vegetation, and jurisdiction. For example, Vicente et al. (2014) estimated that wildfire-related CO_2 emissions in Mexico from 1999 to 2010 were approximately 112 Tg, equivalent to one-third of the country's annual carbon emission (around 400 Tg CO_2 equivalent). And among different vegetation types, burned rainforest contributed to 60% of the annual wildfire-

related CO_2 emission. By Aug 1, 2023, Canada's wildfire-related carbon emissions reached 290 Tg, tantamount to 40% of the country's total carbon emission in 2021 (670 Tg) (Copernicus 2023).

Additionally, unlike most natural disasters, not only does the underlying peril create significant damage, so does the response: wildfire suppression actions can also damage the landscape. Drainage systems can be impacted, soil may be eroded, additional fuels for future fires are created, protected areas may be exposed by new roads and fire guards, and the landscape (and possibly infrastructure) may be damaged by the influx of equipment needed to suppress the fire.

There are significant long-term impacts because wildfires increase the vulnerability of the local ecological system to future losses. The burned vegetation, if not cleared away, creates additional fuels for future fires. After losing vegetation to wildfire, the topsoil of the impacted area can become a water-repellant layer, and flood remains a significant risk until vegetation is restored, a process which could take several years. The soil degradation and loss of vegetation after a wildfire increases the risk of future debris (land, rock, and mud) slides (Staley et al. 2016). In unburned areas, trees and shrubs break the force of the rain; but after a wildfire, rain hits the damaged soil with full force. Shakesby (2011), in examining Mediterranean wildfires, observed that depending on the topography and rainfall, significant runoff and sediment flows can exist for up to 5 years following a fire. Touma et al. (2022) noted that there is a risk of flash flooding for up to 5–8 years after an extreme wildfire.

Postfire flooding can cause greater damage than the wildfire itself. Heavy rainfall in the days following the June 2010 Schultz wildfire created a significant flood event and 85 homes were damaged or destroyed in communities up to 4 miles away, and one person died in the flooding (Syed 2021). Similarly, the catastrophic flooding and mudslides in southern British Columbia in fall 2021 were exacerbated by the region's worst wildfire season in summer 2021 (Clague 2021).

In Colorado, the Grizzly Creek fire burned 32.6 thousand acres in Glenwood Canyon between August and December 2020 but caused relatively little damage. However, it was predicted that significant debris flows were possible in the region if rain intensity were to exceed 24 mm per hour for a 15 min interval. A storm in late June 2021 was anticipated to exceed that threshold, prompting authorities to close surrounding roads in advance of the storm. As a result, no lives were lost, but highway damage was estimated to exceed 100 million USD (Palmer 2022). Thomas et al. (2017) observed that if the debris slide and run-off end up in a watershed, there could be short- and long-term effects on the aquatic ecosystem, as well as substantial water treatment costs.

4 Managing Wildfire Risk

Wildfire seasons are becoming longer, the frequency and severity of wildfires are growing, and the population and assets at risk are increasing. However, landscape fires cannot and should not be avoided. So, the success in reducing wildfire losses is often linked with the ability to manage fires to protect assets, which involves suppression, mitigation, and postfire activities.

While insurance remains an important tool in supporting economic recovery after a wildfire, preventing extreme wildfires may be a more cost-effective solution.[27] Extreme wildfire weather conditions, increased fuel loads, and ignitable structures lead to firefighting resources being overwhelmed. Fire suppression has been shown to be ineffective in battling extreme wildfires, which typically involve a high rate of spread and frontal fire intensity, crowning, prolific spotting, large fire whirls and well-established convection columns (Cohen and Westhaver 2022). As such wildfire mitigation is also necessary to reduce the risk of wildfire on communities. Community and public actions focus

[27] Private insurers also play a critical role in promoting and incenting mitigation, education, and risk reduction as discussed in Sect. 6.3.

Table 4 Key stages in fire suppression

Defensive suppression	Fire behaviour is beyond the control capacity of the suppression resources
Offensive suppression	The suppression resources are making "gains" on the fire. Plans are emerging and being executed
Containment achieved	A control line has been established along the entirety of the sector or division, and subsequent mop-up activities are expected to hold the fire at this perimeter
Mop-up	Complete mop-up of some depth (typically a 100-foot perimeter)
Patrol and rehabilitation	The fire is still smouldering in the interior or in isolated hotspots on the perimeter. Perimeter mop-up is almost complete, and resources can be freed up for rehabilitation or demobilization

Source: Simpson et al. (2019)

on reducing anthropogenic ignition, managing fuel loads, and land use planning. Individual property owners, both residential and commercial, have a role in reducing fuel loads on private property and reducing the flammability of structures. Finally, because of future flood concerns following wildfires, debris removal and ecosystem restoration are important steps in the recovery process.

4.1 Wildfire Suppression

Wildfire suppression operations include extinguishing a wildfire, preventing, or modifying the movement of unwanted fire, or managing a fire when it provides benefits like fuel reduction or improved wildlife habitat (United States, Department of the Interior 2022b). Suppression costs, generally the largest line item in government wildfire management budgets, cover the cost of airplanes, machinery, firefighters, incident command personnel, as well as the infrastructure to house first responders and incident command personnel.

Simpson et al. (2019) examined 10 bushfires (out of an original sample of 75) between 2011 and 2015 in Victoria, Australia, and found that wildfire suppression involved 20 distinct tasks. These tasks can be categorized into five separate stages as given in Table 4. Normally, wildfires progress in order through each of the stages. But in some cases, wildfires do not trigger a defensive stage, and there may be an escape that causes the process to revert to an earlier stage. Depending on the size of the wildfire, stages may not occur concurrently across all locations.

Mattioli et al. (2022), in their review of 44 articles on wildfire suppression costs, found that suppression expenditures were most readily available in North America, and there is a general lack of data on European suppression costs.[28] Simpson et al. (2019) reported overall wildfire operation costs for Australia. Even where only anecdotal evidence is available, the trend is clear: increasing extreme wildfire risk has led to significantly increasing fire suppression costs globally, and these increases are unsustainable for many jurisdictions.

Despite the growth in wildfire severity, most jurisdictions have not seen corresponding growth in fire suppression personnel. Between 2009 and 2020, the United States Department of Interior has experienced a reduction of almost 1000 full-time staff in fire preparedness and fuels management and is staffing 70 fewer wildland fire engines (United States, Department of the Interior 2021). Most

[28] The National Interagency Fire Center in the United States tallies suppression costs for both DOI agencies and the forest service at https://www.nifc.gov/fire-information/statistics/suppression-costs. These data do not reflect the total cost spent on suppression in the United States, as states are responsible for responding to fires on nonfederal lands. Data for Canada are available at https://natural-resources.canada.ca/climate-change/impacts-adaptations/climate-change-impacts-forests/forest-change-indicators/cost-fire-protection/17783.

municipal fire departments, especially those protecting rural populations in the United States, are responsible for wildland firefighting. However, there is a significant lack of resources, personnel, and training for these departments (NFPA Research 2021).[29] In Australia, the number of paid firefighters increased from 17,278 in 2009 to 20,692 in 2019, but the number of volunteer firefighters declined from 169,766 to 152,798 (Karp 2020).

In the event of an extreme wildfire, no jurisdiction has the personnel resources to combat the fire. In North America, there exists a reciprocal cooperative agreement between Canada, the United States, and Mexico for resource sharing during peak fire activities. The EU sponsors a national disaster management system across 34 countries, and the EU Civil Protection Mechanism coordinates and pays for the majority of deployment costs. The disaster response may also be deployed globally. During the Australian Black Summer bushfires of 2019/2020, New Zealand, Papua New Guinea, Singapore, United States, and Canada sent additional personnel and equipment to fight the bushfire. In summer 2023, Canada's worst fire season on record, crews from Australia, New Zealand, South Africa, United States, Spain, France, Portugal, Chile, and Costa Rica joined force with Canada to fight fires.

4.2 Community and Public Fire Mitigation and Prevention

Even countries with sophisticated technical firefighting capacity cannot respond adequately to extreme wildfires, and, as such, the World Bank (2020) stressed the necessity that countries broaden their focus from suppressing wildfires alone to both suppression and prevention efforts. Investments in fire suppression, therefore, should be balanced with investments in measures that prevent and reduce the severity of wildfires.[30]

While reducing the frequency of losses focuses on human activities that ignite fires, reducing the potential impact of landscape fires focuses largely on strategies to cut the fuel load available and protect economic assets within the WUI. Compared to the cost of fighting wildfires, mitigation is relatively inexpensive. The Insurance Information Institute (2022) notes that the per acre cost to fight wildfires is 50,000 USD, while the cost of mitigation is merely 1500 USD. Porter et al. (2021) estimate that new housing built to the National Research Council's WUI guidelines have a benefit-cost ratio of 30:1 and retrofitting existing housing could see savings of $14 for every dollar spent.

Ignition Prevention

The majority of wildfires in the United States and Europe are caused by human activity (Sjöström and Granström 2023; Congressional Research Service 2023). As a result, governments have placed a great emphasis on reducing anthropogenic ignition, often by prohibiting activities, such as open fires. In Russia, where 72% of fires can be ascribed to arson or carelessness (Hesseln 2018), the government has banned the burning of dry vegetation on farmland and around other important infrastructure (e.g., railways) (Agence France-Press 2015). Other interventions include the cessation of mechanical crop harvesting during particularly dry conditions and developing programs to raise community education and awareness.

[29] Most local departments have formal or informal agreements to obtain assistance at regional or state levels if necessary.

[30] However, in many countries, the budget for wildfires encompasses all activities related to wildfire management, including preparedness, suppression, fuels management, firefighting facilities, postfire rehabilitation, and investments in fire science. As a result, in years when suppression costs are high, less may be allocated towards wildfire prevention and preparedness.

Fuels Management

An increase in available fuels also increases wildfire risk. Reducing fuels that feeds a wildfire improves the efficiency and safety of wildfire suppression as well as reducing the potential of devasting wildfire impacts (United States, Department of the Interior 2022a). Fuels can be reduced by prescribed burns, thinning forested areas by hand or with equipment, reducing the quantity of grass by animal grazing, and chemically treating areas to remove invasive plants. Establishing and maintaining fuel breaks—designated areas at selected locations where combustible vegetation is intentionally reduced—is an important fuels management technique.

A prescribed burn deliberately uses fire in a specific area under certain weather conditions for forest and wildlife management and to reduce fire risk. Prescribed burns reduce fuels buildup and alleviates fuel load and can prevent a small fire from becoming out of control. In Germany, sparks from train brakes caused more than 800 fires in the 2003 fire season, creating massive railway disruptions, and putting nearby properties at risk. In response, Germany began undertaking prescribed burns near train tracks to reduce fire risk (Hetzel and Goldammer 2004).

Prescribed burns near communities tend to be more costly, as more resources are typically allocated to ensure that fires stay under control (Florec et al. 2020).[31] Firebreaks, which can be natural barriers like water bodies or constructed ones like two-track dirt roads, are used to minimize the likelihood of prescribed fire escaping (Weir et al. 2012). However, the benign effect of prescribed burning diminishes over time, and it becomes less effective when fire weather conditions become more extreme (Tolhurst and McCarthy 2016).

Cultural burning, used by indigenous communities around the world, implements controlled burning for cultural purposes (e.g., promoting vegetation for food and medicinal purposes, clearing the understory of vegetarian, creating sustainable cohabitation with wild animals).[32] Although both cultural and prescribed burnings create the same outcomes, such as fuel load reduction and/or habitat enhancement, cultural burning is based on the distinct cultural values, perspectives, or practices of indigenous communities. Indigenous fire stewardship has largely been undervalued after the arrival of colonization, and only recently has become an accepted practice for land management.[33]

Another practice for reducing wildfire severity is mechanical thinning. This involves decreasing the density of vegetation within an area by removing some trees, pruning the lower branches of some trees, or creating fuel breaks. Fuel is reduced, reducing the likelihood of a catastrophic event, and helps to maintain and restore ecosystems. The U.S. Forest Service (n.d.), for example, suggests that mechanical treatment can be used in places and conditions when it may not be safe to carry out prescribed burnings.

Mechanical thinning and prescribed burning are costly to implement and maintain long term (Maestas et al. 2016, among others) and these practices may have significant environmental impacts (Davies et al. 2023). A less conventional method of fuels management is targeted grazing, where animals remove ladder fuels from the environment, thus reducing the risk of a surface fire becoming a canopy fire (see, e.g., Taylor Jr. 2006, and Davies et al. 2022). Sheep, goats, cattle, or wildlife such as deer are used to manage vegetation in a variety of settings, especially on landscapes that are too steep or rocky for other human-led management techniques (e.g., mowing or chemical treatment). Since

[31] Outcomes can be catastrophic when prescribed burns get out of control. The 2000 Cerro Grande Fire in New Mexico, United States, scorched 48,000 acres instead of the planned 900, destroying 235 homes and displacing 400 people, resulting in an estimated loss of 1 billion USD (1.75 billion in 2023 USD) (Wyland 2020). Nevertheless, cases where a prescribed fire escaped and caused significant damage have remained rare (Kobziar et al. 2015).

[32] One example is the Owl Creek Cultural Burn in July 2022 in British Columbia, Canada, which aimed to restart berry growth for both the Lil'wat Nation community and grizzly bears (https://prescribedfire.ca/cultural-burning/).

[33] See, for example, National Park Service (n.d.), and Cultural Burning and Prescribed Fire (n.d.).

targeted grazing does not require ignition, it does not bring the large fire risk attached to prescribed burning. Clark et al. (2023) find that it can effectively create and maintain fuel breaks on highly flammable rangelands. However, enough fuels may not be removed to be effective if targeted grazing starts too late in the season or ends too soon; if it is difficult to keep the livestock in the designated grazing area; or if there are scheduling issues that conflict with the timing and duration of the targeted grazing.

Finally, although fuels management significantly reduces fire risk, UNEP (2022) cautioned that communities cannot solely rely on fuels reduction to prevent extreme wildfires.

Land Use Planning

While eradicating extreme wildfires is not possible, as Moritz et al. (2014 p 64) have suggested, communities can achieve "a sustainable coexistence with wildfire" by improved land-use planning. Land use planning encompasses landscaping regulations and subdivision design standards. To reduce wildfires around the WUI, land use planning emphasizes strict requirements on allowable vegetation, zoning, and building codes. Building codes and standards in the WUI are designed to reduce the risk of wildfire damage and discourage development that negatively impacts fire-prone ecosystems. Land use planning tools also ensure adequate water supply, road widths, and access for firefighting equipment. In turn, this creates a safer environment for fighting fires and results in fewer structures being destroyed (Gude et al. 2008).

Promoting fire-smart land use planning requires not only well-designed policies but also economic incentives that encourage appropriate landscape management and use for residential, commercial, agricultural, and recreational activities. The Bay Area Council Economic Institute (2021) proposed that municipalities can apply differential tax rates to new constructions in the WUI, which could account for future costs of providing fire protection to these buildings, thus potentially disincentivizing high-risk construction.

4.3 Property Owner Mitigation Strategies

The destruction of homes accounts for the majority of structural losses. As Cohen and Westhaver (2022) have noted that, with respect to the destruction of communities, the crux is not just a "wildfire problem" but one of easily ignitable structures and homes. After all, property owners have no control over wind and weather factors that exacerbate wildfire risk, but they do have significant control over the vegetation on their property and the ignitability of their homes. Consequently, the top priority should be reducing the ignition potential of structures and infrastructure.

The two elements to consider are creating a defensible space around dwellings by managing vegetation and reducing ignitability by home hardening. Taken together, these actions can significantly decrease wildfire losses (Institute for Catastrophic Loss Reduction 2019).

To create a defensible space, the ground area near the dwelling should consist of hard or non-combustible materials. Vegetation, including tree branches, should not encroach upon buildings. Further away from the dwelling, the focus is on reducing the fuel load and preventing surface fires from climbing into the crowns of trees. Embers, which can be blown miles ahead of the fire, are responsible for the overwhelming majority of house destruction during a wildfire. As such, removing possible ignition sources or fuels around the house is the first defense against wildfires (National Fire Protection Association n.d.; Mell et al. 2010; Westhaver 2017).

The second element, home hardening, is the process of preparing a home (and other household structures such as sheds or garages) to reduce the vulnerability to embers and heat. Exterior building surfaces—roofs, walls, eaves, soffits, decks, and fencing close to a building—should be ignition

resistant and non-combustible. Eaves and rain gutters should be kept clear of debris, and windows should be double-paned and have screens to increase ember resistance. Flammable objects, such as patio furniture and BBQ tanks, should be moved away from structures during a wildfire.

Organizations such as FireSmart™ in Canada and Firewise USA® provide wildfire mitigation advice to homeowners. Although the focus of such programs is on residential property owners, similar mitigation activities can be adopted by commercial property owners.

4.4 Postfire Activities

With respect to postloss activities associated specifically with wildfires, various restorative practices are often necessary. The land burned by wildfires needs to be rehabilitated in order to repair ecosystems, reduce the likelihood of postfire flooding, and rebuild recreation infrastructure. Restoration activities may also include reforestation, ensuring that peatlands are not drained or changed, and invasive species are effectively managed (World Bank 2020).

Soil stabilization, the practice of replanting trees and shrubs and reseeding native grasses and plants, will help control erosion and reduce the likelihood of future flood events. Reseeding native plants also helps to reduce the threat of invasive species. Watersheds and water habitats also need to be rehabilitated, and this can be done by planting trees and shrubs, reducing soil erosion, restoring waterways to their natural shape, and using wildlife, such as beavers, to restore habitats. Landscapes may also be transformed to minimize the impact of future runoffs.

As with all catastrophic losses, the disaster recovery process also involves restoring and redeveloping impacted WUIs: land use decisions are revisited, and structures and infrastructure are repaired and replaced. Delays in postdisaster reconstruction can lead to persistent negative consequences and hinder recovery in the long term (Cambridge Centre for Risk Studies 2020).

To improve the resiliency of communities after a catastrophic wildfire, it is important to build more sustainable communities that go beyond immediate disaster relief. However, Wibbenmeyer et al. (2019) argued that spending funds on fire prevention, specifically fuels management, in communities that have already experienced a significant wildfire loss could be considered a misallocation of resources. If debris after a wildfire has been properly removed, the immediate fire risk is much smaller as there is less fuel to burn.[34] As such, Wibbenmeyer et al. (2019) contend that the funds might be better spent on fire prevention activities in communities of similar risk profiles but have not yet experienced a wildfire.

4.5 Improve the Collection and Analysis of Wildfire Data

To improve the understanding of wildfire causes and to identify gaps in prevention and response capacities, wildfire data collection and analysis are much needed. As World Bank (2020 p. 19) recommended, governments should "systematize and operationalize fire danger rating and early warning systems, including the improved integration of weather data in fire forecasting". With quality data and proper modeling, governments and communities can review wildfire management practices to identify best practices.

[34] Fires caused by lightning within recently burned areas were less likely to lead to large fires than in comparable areas without recent fires (Halofsky et al. 2020).

However, data on the economic impact of wildfires remain unsystematically collected. Even in the United States, loss estimates are dispersed across news reports and scholarly reports, and systematic estimates of many losses remain unavailable (Thomas et al. 2017). The same is true for all nations.

Compared with insurance models for other catastrophes (e.g., earthquakes and hurricanes), the modeling of potential insured losses from wildfire is much less well developed. This can be attributed to three factors. First, human-induced activities underpin the majority of wildfires. The National Park Service (2022) notes that nearly 85% of all wildfires in the United States are caused by human factors.[35] Incorporating human behaviour into a model for wildfire is more difficult than that for earthquakes or windstorms, as the latter is not affected by human actions. Second major modeling firms only cover a limited number of geographic regions. For example, Moody's Risk Management Solutions mainly models wildfire in North America; Verisk has a wildfire model for the United States, covering 13 western states where the risk is greatest, and a bushfire model for Australia. ZestyAI has recently developed a community-level and a property-level score for communities in the United States. Last, successful wildfire risk modeling requires the integration with climate modeling, which is further complicated by climate change in recent decades.

There are intricate causal relationships among factors that may lead to increased wildfire risk. For example, lightning strikes are likely to cause wildfires only if vegetation is dry (Macias-Fauria et al. 2011). There is a lack of definite causal impact of precipitation and temperature, as the two factors are region-specific and may confound with other factors like the type of vegetation (Westerling and Bryant 2008). Undoubtedly, recent advances in remote-sensing techniques, lightning detection systems, and the use of meteorological variables have advanced the prediction of lightning. Despite such development, wildfire projections by different scientists differ considerably, and it remains uncertain how sensitive such projections are to climate change.

5 Wildfires in the Future

Wildfire risk is increasing globally. Tyukavina et al. (2022) document increasing forest loss due to fire across all biomes, with the greatest loss in boreal subregions of North America and Eurasia, followed by subtropical forest losses in Australia and Oceania.

5.1 Reasons for Growing Insured Losses

The key drivers of the growth in insured (and economic) losses from wildfire are land use, economic growth and the accumulation of assets, and climate change (Pandey et al. 2023, among others). We discuss each of these drivers in turn and then provide empirical evidence of the increase in frequency and/or severity of wildfire losses globally.

Land Use

Land use is defined as the "direct human alteration of terrestrial ecosystems" (UNCCD 2017 p. 30) and encompasses where and how land is used for commercial, residential, agricultural, and recreational activities. One of the most significant drivers of insured wildfire losses is the increase in the number of communities near wildland areas (Cappelletti 2019). Because of increasing

[35] In contrast, 40% of fires in British Columbia, Canada are caused by human induced activity (https://www2.gov.bc.ca/gov/content/safety/wildfire-status/wildfire-response/what-causes-wildfire), and the proportion exceeds 95% in Europe (https://www.bbc.com/news/58159451).

urbanization, countries around the globe have experienced significant growth in the WUI, where most fires are caused by human activity and fueled by surrounding vegetation. In the United States, from 1990 through 2010, the size of the WUI has increased from 143.57 to 190.27 million acres, with the number of dwellings increasing from 30.8 to 43.3 million (Radeloff et al. 2018). Between 1981 and 2021, the WUI in central-western Patagonia in Argentina grew by 80% (Godoy et al. 2022). The WUI covers roughly 7.5% of Europe, but within individual countries, the size of WUI ranges from less than 1% in Malta and Ireland to 23% in Liechtenstein. In the Mediterranean countries, which accounted for 86% of burned areas in Europe between 1990 and 2019, the size of the WUI ranged from 4.5% in Spain to almost 15% in Italy (Bar-Massada et al. 2023).[36] The largest WUI is in Asia, measuring 1.6 million km^2 (Chen et al. 2022).

Besides WUIs, other land use decisions can exert a direct impact on the severity, frequency, and intensity of wildfires through increasing fuel loads and ignition opportunities. For example, forestry incentives such as large-scale plantations, postfire salvage logging, and the planting of commercial species such as eucalyptus have led to increased wildfire risk (World Bank 2020). The use of above-ground utility power lines is representative of another land use decision. Utility power lines, a significant ignition source in California, are responsible for six of the state's most destructive wildfires. There are more than 40,000 miles of exposed power lines in wildfire-prone areas, but little progress has been made to reduce the risk of ignition, such as making equipment more fire-resistant (California State Auditor 2022).

The change of the ecosystem through the construction of large-scale infrastructure (e.g., railways and power station) can also exacerbate fire risk. One example is the building of the Irkutsk Hydroelectric Power Station—located on Angara River and upstream of Lake Baikal—in Siberia in 1956. The power station disrupted the natural flow of the river, causing low water levels and deficiencies in water flow out of the Lake, thereby contributing to increasing peat fire incidents in the Selenge River delta (Wang et al. 2018).

Active fire suppression has been a major contributor of extreme wildfires in many parts of the world including Canada (Parisien et al. 2020) and the western United States (DellaSala et al. 2022). The historic practice of extinguishing wildfires immediately (or not implementing prescribed burns) has resulted in an overgrowth of shrubs and debris on forest floors, and a greater encroachment of trees on grasslands providing fuels, especially during dry seasons, for surface fires. The absence of landscape fires also negatively impacts forest health and biodiversity, increasing susceptibility to pests and diseases that also increase fuel load.

Economic Growth and Accumulated Assets

Another driver of increasing insured losses is economic growth and an increase in accumulated assets. Economic growth, characterized by urbanization and rising populations, leads to increased insured losses in areas susceptible to natural disasters (see, e.g., Kunreuther and Michel-Kerjan 2013). For example, the Cape Coral–Fort Meyers metropolitan area—where Hurricane Ian caused a landfall in 2022—saw its population grow by 620% since 1970 (Swiss Re 2023). Although the year 2022 was an average year in terms of the number of hurricanes, it turned out to be the third costliest hurricane season on record.[37] Most of the losses resulted from Hurricane Ian, which created 52.5 billion USD in insured losses (Insurance Information Institute 2023b).

[36] Different mapping techniques produce different estimates of the size of the WUI. Chen et al. (2022) calculated that the WUI in Europe covers 13% of the total land mass.

[37] There were 14 named storms in the 2022 Atlantic hurricane season, and the 30-year average is 14.4 storms (Swiss Re 2023). Hurricane Ian became the third costliest natural catastrophe event on record after Hurricane Katrina in 2005 and Hurricane Harvey in 2017.

High rates of inflation have increased the nominal value of property, buildings, and other insurable assets. Since the start of the COVID-19 pandemic, property insurance exposures have grown at a faster rate than real GDP, adding further pressure to insured losses by increasing the nominal value of buildings and other insurable assets. In 2022, inflation rates have averaged 7% in developed economies and 9% in emerging economies (Swiss Re 2023). High inflation and disruptions in the global supply chain have also put pressure on rebuilding costs after a large catastrophe. For example, building replacement costs in Australia after the February and March 2022 flooding rose by more than 20% (Swiss Re 2023).

Climate Change

Climate change produces higher ambient air temperature and less rainfall, creating drought conditions. Such conditions increase the quantity of dry fuels, which subsequently leads to an increase in the acreage burned. Hotter temperatures in California, for example, have contributed to record-breaking droughts over the past decade. In turn, over 147 million trees died in the state between 2010 and 2018 (Bay Area Council Economic Institute 2021), creating fuels for future fires. Jones et al. (2020) provided a concise summary of the impact of climate change on wildfire risk globally. IPCC (2022) notes that burned areas have increased in North America, the Amazon, Australia, and parts of Africa and Asia, which are consistent with the consequences of climate change. Other country-specific evidence can be found in the United States (Westerling and Bryant 2008), Canada (Le Goff et al. 2009), the Mediterranean (Moriondo et al. 2006), Australia (Di Virgili et al. 2020), India (Mina et al. 2023), China (Du et al. 2021), among other locations.

Climate change also increases the variability of weather patterns over time. Touma et al. (2022) have predicted a significant increase in the number of extreme fire weather events with an extreme rainfall event in the following year across the western coast of the United States. This gives rise to an increased likelihood of postfire flood events. The fire season, characterized by dry and windy weather that helps flames to spread, has lengthened (Jolly et al. 2015; IPCC 2022). According to OECD (2023), the duration of the annual fire weather season has increased by 27% between 1979 and 2019. Also increased is the severity of convective storms, which facilitate ignition by lightning (Macias-Fauria et al. 2011).

However, the intricate causal links between wildfires and weather conditions make future wildfire prediction difficult. First, wildfire has a reciprocal relationship with climate change. Climate change, as noted above, has exacerbated wildfire risk but wildfires are also significant contributors to climate change: when forests burn, CO_2, along with other greenhouse gases, is released back into the atmosphere, contributing to climate change. Second, wildland fire activities are strongly influenced by the combination of fuels, ignition agents, and weather (Flannigan et al. 2005). In locations where peak lightning occurs during rainy seasons, wildfires are less likely (Macias-Fauria et al. 2011). Also, impacts of precipitation and temperature are region-specific and are confounded with other factors such as the type of vegetation (Westerling and Bryant 2008; Macias-Fauria et al. 2011). Based on Australian data, Wang et al. (2022) find that landscape characteristics (e.g., forest, water, or urban cover), vegetation and the types thereof (which affect fuel loads), and fire indexes also explain the geographic variations of bushfires.

Although climate change has not been the prominent driver of increased losses to date, its importance is anticipated to increase in the future (Swiss Re 2023; IPCC 2022). Of all weather-related perils impacted by climate change, Swiss Re (2023) projected that wildfire perils will see the second highest growth rate behind coastal flooding.

5.2 Global Evidence of Growing Losses

Growth in wildfire incidents has been a global reality. Global Forest Watch (n.d.), a wildfire tracker, found that tree cover loss from fires has increased in the past two decades, rising from 6.2 million acres in 2001 to 23 million in 2021. According to Chaucer, a specialty reinsurance group, "the land mass burned by wildfires in Europe jumped by 75% from 2019 to 2020" (Evans 2022). Recent wildfires have generated significant insured losses that have far exceeded earlier ones. As shown in Table 1, with the exception of the Oakland Tunnel fire in 1991, the costliest wildfires have all occurred since 2016. Between 2018 and 2022, global wildfire losses were 69 billion USD, with 39 billion USD in insured losses (Munich Re n.d.).

Looking into the future, Bailey and Yeo (2019) have placed southern Canada, western United States, southern Europe, and southern Australia into the "Negative Outlook" category, suggesting an increase in their wildfire risk; meanwhile, eastern United States, eastern Europe, central Asia, China, and western South America appear on the "Watch" list, indicating heightened concerns. We briefly highlight the increase in wildfire risk in these different regions below.

Boreal Forests

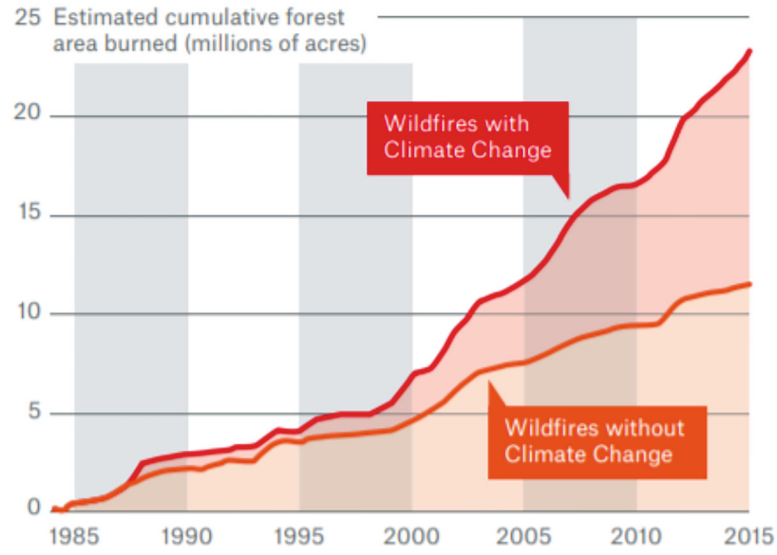
Holdover fires have increased in frequency and severity because of climate change: annual temperatures in the Arctic and subarctic have increased approximately twice the increase in average global temperatures, with the most significant change occurring after 1990 (Baltzer et al. 2007, among others). The return period of significant wildfires in Siberia has increased from 1-in-100-years in the late 1850s to 1-in-65-years by 2004. Between 2000 and 2010, 50% more acreage has burned annually in the Arctic than in any decade in the previous century, and the acreage burned increased again between 2010 and 2020 (Borunda 2021). Tomshin and Solovyev (2022) estimated that the average annual burned area in Siberia from 2001 to 2020 was 16.06 million acres, and the size is likely to remain stable. However, underlying this stability are asymmetric impacts to different parts of Siberia: West Siberia is likely to see a decrease in burned areas, but southern and central Eastern Siberia are likely to record more catastrophic fires.

North America

In North America, the number of large grassfires (burning at least 1000 acres) across the Great Plains increased from 33 per year in 1985 to 116 per year by 2014, and the total acreage destroyed grew by 400%. The increase in the total area burned is greatest in the High Plains, Southwest Tablelands, the Northwest Great Plains, and the Edwards Plateau ecoregions (Donovan et al. 2017).

Wang et al. (2015) projected that over the next century, climate change will increase wildfires across all of Canada. The fire season is expected to be significantly wetter by 2080, but the increase in precipitation will not offset the impact of the expected temperature increase. Instead, weather conditions will make it easier for fires to ignite, survive, and spread. The impact is anticipated to be felt first in more southern regions, with the number of days in which fire could spread (days in which the weather is conducive to wildfires) increasing by 200% to 300% in eastern Canada and more than 50% in western Canada. Evidence as of 29 September 2023 supports this projection, as Natural Resources Canada (2023) recorded that 17.8 million hectares have burned in the 9 months, exceeding the 10-year average of 2.7 million hectares by more than sixfold.

In the United States, Abatzoglou and Williams (2016) demonstrated that the amount of area burned in the western states between 1985 and 2015 was roughly twice as large as it would have been without climate change, as shown in Fig. 2. The impact of climate change, as projected by Munich Re (2019), is that severe fire years will become the new normal in the western states of the United States.



Source: Abatzoglou and Williams (2016) as reported by Munich Re (2019).

Fig. 2 Historical wildfire risk in western United States attributed to climate change

Europe

For the 27 European jurisdictions that reported fire-related statistics in 2016 and 2021, the total number of fires decreased by 7665 but the area burned increased by 3700 acres (calculated from JRC 2017, 2022). However, fire risk, as measured by the Canadian Forest Fire Weather Index System, has increased significantly.[38] The European Environment Agency (2021) credited improvements in mitigation as the reason for the reduction in the number of fires.

Since 2016, Greece has contributed the highest burned area in Mediterranean Europe (European Environment Agency 2021). In August 2021, unprecedented wildfires ravaged multiple cities in Greece, burning down a total of 0.32 million acres (Xanthopoulos et al. 2022). While an estimate of the economic losses is not available, the government released a relief package of 500 million EUR (550 million USD) to residents in Evia and the Attica Region (Papadopoulos 2021).

Asia

An abrupt surge in fire risk in recent years applies to some parts of Asia, and the reasons are a combination of high temperature, drought, and human-caused ignitions (Yang et al. 2021). In 2002, Uttarakhand—a state in India—reported 922 wildfire events and this increased to 41,600 by 2019 (Mina et al. 2023). The same situation has been observed in other parts of Asia: From August 2021 to August 2022, incidents of wildfires in Mongolia increased by 73% (Ankhtuya 2022).

Australia

Australia's growth in wildfire losses is startling. Before 2019, the three largest fires occurred during 2002/2003 (Victorian bushfires), 2006/2007 (Great Divide bushfires), and 2008/2009 (Black Saturday

[38] The index is based on daily observations of temperature, relative humidity, wind speed, and 24 h precipitation, which impact the moisture content of the forest floor and fire spread, fuel available for combustion, and fire intensity (https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi). This index is the most widely used fire weather system in the world.

bushfires), scorching 3.2, 1.3, and 1.1 million acres, respectively. All three are one magnitude smaller than the 2019/2020 Black Summer bushfires—by far the largest in Australia—which encompassed 183 postal codes across four states, killed 445 people, and burned more than 42 million acres of land.

The increasing fire peril has also been confirmed by researchers. Applying a spatial—temporal model, Valente and Laurini (2021) decomposed the number of fire occurrences in Australia into trend, cycle, and seasonal components. Using data from 2003 to 2020, the study showed an upward trend of wildfire occurrences since 2011, which coincides with an upward trend in temperature.[39]

6 Growing Concerns About Insurability and Potential Solutions

Given the evidence on growing wildfire losses in the previous section, it is not surprising that there are concerns about the future insurability of wildfire losses. As the frequency and severity of wildfires grow, there will be a point where premiums become unaffordable, and insurers are unwilling to offer coverage.[40] This is already the case in California, where some insurers have been refusing to renew residential property coverage in high-risk neighborhoods (see, e.g., California Department of Insurance 2021b). In Australia, the Climate Council (2022) estimated that by 2030, roughly 4% of properties will be uninsurable due to climate change. As insured losses from wildfires grow globally, more jurisdictions may face the same issues as the two countries.

6.1 Insurability of Wildfire Losses in California

As Table 1 shows, four of the costliest insured losses from wildfires occurred in California. The 2017 and 2018 wildfires, in particular, had a profound impact on the property-casualty insurance market. First, more properties were classified as high-risk. Owners of residential and commercial properties categorized as high-risk were faced with cancellations and non-renewals. Many property owners whose policies that were not canceled saw their premiums increase by over 400% (California State Assembly 2020).[41] Second, given that wildfire losses also caused reinsurers' loss ratios to exceed 125% in both years, there were concerns about future reinsurance availability in the state, putting more pressure on primary insurers (Moody's 2023). As a result, the number of policies insured by the California FAIR plan, the state's insurer of last resort, grew by 240% between 2017 and 2020.

[39] Wang et al. (2022) also obtained the result of increasing risk of wildfires, linking Australia's fire-conducive climate change to three climate modes, the El Niño Southern Oscillation, the Indian Ocean Dipole, and the Southern Annular Mode.

[40] Insurability issues arise both from primary insurance market and reinsurance market conditions. The ability of primate insurers to offer coverage for potentially catastrophic losses such as wildfire is largely linked to reinsurance availability. Because primary insurers cannot hold enough surplus or collect enough premiums annually to pay for a catastrophic loss, the fundamental determinant of a domestic market's capacity is its ability to transfer the risk at an affordable price with reasonable terms, of which the international property catastrophe reinsurance markets play a significant role. In high-income OECD countries, roughly 15% of gross written property premiums are ceded to unaffiliated reinsurers, and there is a positive correlation between the level of catastrophe exposure and the use of property reinsurance (OECD 2018).

[41] Liao et al. (2022) document average premium increases between 2003 and 2018 and find that in the voluntary market, very high-risk zip codes had seen an average increase of 24.3% compared to a 4.5% increase in low-risk zip codes and a slight decrease in very low-risk zip codes. For FAIR plans, the very high-risk zip codes saw an average increase of 57% compared to 46.2% in the lowest-risk zip codes.

Pricing Models

California's strict pricing guidelines required that insurers price coverage based on retrospective information only. And they were forbidden from incorporating information on current or future anticipated events, such as expected future wildfire losses due to climate change, into pricing models. In response, many insurers placed restrictions on new policies or ceased writing in high-risk zip codes. This was changed in September 2023 with the introduction of the California Sustainable Insurance Strategy (California Department of Insurance 2023). This legislation has allowed for the incorporation of climate modeling in pricing models along with the requirement that the rating methodology also incorporates expected savings arising from property and community-level mitigation efforts. Additionally, in order to use the new pricing models, insurers were also required to write no less than 85% of their statewide market share in distressed areas as identified by the Department of Insurance.

Meanwhile, pressure from insurers to place limits on coverage from wildfire losses is unlikely to be accepted by legislators. In 2018, the California legislature introduced a 1-year moratorium on canceling or non-renewing residential insurance policies in areas within or adjacent to a fire perimeter, as defined by zip code, and the legislation was reenacted again in 2019, 2020, and 2021 (California Department of Insurance n.d.). This intervention proved successful: Liao et al. (2022) noted that insurer-initiated non-renewals in protected zip codes dropped by 20% between 2019 and 2020, compared with a 2.5% drop in other parts of the state.

Inverse Condemnation

There is one more key issue that complicates the insurability of wildfire risk in California: inverse condemnation. Inverse condemnation is a legal concept that entitles property owners to compensation if their property is damaged by public use. In particular, if utility companies are found to be responsible for a wildfire, they are liable for the subsequent damages. This is relevant for California, where roughly 10% of wildfires are started when dry vegetation surrounding power lines catches fire from sparks and discharges from the lines (McCormick 2021).

Jacoby (2022) claimed that California's inverse condemnation rule is unique in the following two ways. First, it is applied under the doctrine of strict liability, which has two distinct impacts. On the one hand, since utility companies are held completely responsible for wildfire losses, individuals and municipalities lack the incentive to undertake actions to reduce wildfire risk. On the other hand, under the strict liability regime, the mitigation costs—which must be raised from investors—render utilities in California more costly to operate than those in other jurisdictions. Second, the definition of public use has been extended to private entities that offer public services (e.g., a utility company). For example, the electrical transmission lines belonging to Pacific Gas & Electric (PG&E) were responsible for the ignition that caused the 2018 Camp Fire in California. The application of inverse condemnation led PG&E to file for Chapter 11 bankruptcy.[42]

There have been proposals to reduce the liability of utilities, arguing that the concept of proportionate fault or the application of a reasonableness test should also apply. It is also argued that privately owned utility companies should be excluded from inverse condemnation as regulations restrict the ability of these companies to spread the cost of losses across their customers.

The solution to the costs associated with inverse condemnation has been the creation of an insurance pool that covers the claims of participating utility companies (San Diego Gas & Electric, Southern California Edison, and PG&E). In the year following the 2018 Camp Fire, the state established the California Wildfire Fund to provide approximately 21 billion USD in claims paying capacity. It is funded by contributions from utility companies and ratepayers and administered by the California Earthquake Authority (CEA). Participating utility companies are responsible for the first 1

[42] PG&E also pleaded guilty to 84 felony counts of involuntary manslaughter and was fined four million USD.

billion USD of losses with amounts in excess of that will be paid from the fund (California Wildfire Fund 2023).

Except for a very small purchase of reinsurance in its inaugural year of operation, the fund has not purchased reinsurance since noting that the cost and structure of coverage available would not improve the fund's durability (Evans 2020a, 2020b). As of December 2022, the fund has received 12 billion USD in capital and has the ability to issue debt backed by surcharges on ratepayers. As of April 2023, no claims have been paid by the fund (California Catastrophe Response Council 2023).

6.2 Insurability of Wildfire Losses in Australia

In Australia, a significant portion of the losses from the 2019/2020 bushfires were covered by reinsurance. Even though loss ratios of primary insurers were high (71% in 2019/2020 versus 63.5% in 2016/2017), thanks to the provision of reinsurance, the industry still posted 1751 million USD (2042 million in 2023 USD) in profits (KPMG 2020). Nevertheless, the payout pressure on reinsurers led Swiss Re to comment that Australian insurers have consistently failed to predict the true cost of catastrophic events (Fernyhough 2020).

Looking into the future, the likelihood and severity of catastrophic events in Australia are expected to increase under the backdrop of climate change. Although riverine flooding is the key peril faced by property owners, bushfire risk is also significant. In the federal electorates of Richmond and Page in New South Wales, more than 5% of properties will be considered high risk due to bushfire. The Climate Council (2022) estimated that by 2030, roughly 4% of properties will be uninsurable, and a further 9% will be subject to annual expected damage costs that equal 0.2% to 1% of replacement cost.[43] The percentage of uninsurable properties varies by state and territory, with estimates ranging from 1.3% of the properties in the Australian Capital Territory (ACT) to 6.5% in Queensland.

Increasing risk will lead to increasing premiums. The Institute of Actuaries of Australia (2022) estimated that under IPCC's emissions scenarios, premiums required to cover bushfire losses could grow between 7% and 33%. The expected increases are not uniform, with current high-risk coastal areas that have already experienced significant bushfire losses anticipated to see the greatest increase.

Following the devastating bushfires of 2019/2020, the Australian government recommended that an inquiry into the availability and affordability of general insurance for both commercial and residential policyholders in bushfire-prone zones be undertaken (Australia, The Senate 2021).

One key recommendation is to embed the potential value of mitigation into insurance premiums. Historically, insurers have not been able to monitor or measure the impact of mitigation improvements undertaken by property owners to reduce bushfire risk, and as such, insurance premiums have not reflected mitigation activities undertaken. The development of a national Disaster Resilience Star Rating (DRSR) program by the Bushfire Building Council of Australia could provide a mechanism by which insurers can recognize mitigation activities in underwriting models. For example, the DRSR app uses scientific algorithms to help residents quantify the value of their mitigation activities, such as assessing how the surrounding environment and the structure of the property would affect future fire ratings.

In 2021, the Australian government created a reinsurance pool backed by a limited government guarantee for cyclone and flood damage in northern Australia to provide readily available and

[43] According to the report's definitions, a property is considered uninsurable if the expected damage cost annually exceeds 1% of the property's replacement cost. Even if insurers were willing to offer coverage, the premium would be unaffordable for most property owners.

affordable coverage in high-risk areas. It is reasonable to consider that if significant supply concerns arise due to bushfire risk, this pool could be extended to cover wildfire.

Underinsurance and uninsurance of residential properties are of significant concern in Australia, and government funding mechanisms for emergency services exacerbate the problem.[44] Much of this funding historically came from levies and taxes on home insurance policies, with tax rates varying between 10% and 30% depending on the state. The Insurance Council of Australia (ICA 2021) noted that the total taxes will rise from 1 billion USD in 2020/2021 to 1.29 billion USD by 2024/2025.

6.3 Potential Insurance Company Responses to Growing Wildfire Losses

The availability of insurance creates moral hazard as it may not disincentivize individuals from living in areas that are at high risk of wildfire losses (Hazra and Gallagher 2022), especially if future wildfire risk cannot be priced in the policy, as was the case in California until recently (Associated Press 2023).

In addition to paying claims, insurers have a role to play in assisting property owners in reducing the frequency and severity of wildfire losses. For example, insurers can offer discounts for community-wide mitigation, such as community designation of Firewise or Shelter-in-Place in the United States, and for individual home-specific mitigation such as home hardening or maintaining a defensible space (see, e.g., California Department of Insurance 2022). However, insurance savings alone are typically insufficient to incentivize property owners to engage in loss mitigation (Kunreuther and Michel-Kerjan 2013) and monitoring costs may be too high for insurers (Kunreuther et al. 2013). Mitigation is likely to be undertaken if properties are deemed to be at such high risk that insurers simply reduce or stop providing coverage in certain vulnerable regions until communities and property owners adopt sufficient loss mitigation practices. As noted in Sect. 2, in response to catastrophic wildfire losses in high value neighbourhoods in the United States, some insurers, notably Chubb and AIG, offer prefire and postfire mitigation services and structure protection services to high-net-worth policyholders (Chiglinsky and Ballentine 2018). Since insurers are directly offering the services, this alleviates the high cost of monitoring.

Insurers also contribute to education and research. The Insurance Institute for Business and Home Safety (https://ibhs.org/) in the United States and the Institute for Catastrophic Loss Reduction (https://www.iclr.org/) in Canada are industry-sponsored scientific research and communications organizations. These organizations provide the industry with technical information about building codes, effective land use codes, new building designs, and retrofit applications, with the goal of creating more resilient communities. Some insurers also sponsor academic research by creating chairs and postdoctoral fellowships at universities.

6.4 Potential Insurance Market Solutions

The availability and affordability concerns for property coverage in wildfire-prone areas, combined with the hardening of reinsurance markets in 2023 and the expected future growth in climate-induced catastrophic events are putting and will continue to put pressure on private insurers. In this section, we discuss possible interventions, including nontraditional insurance products and government

[44] Eleven percent of households in Australia do not purchase insurance (Institute of Actuaries of Australia 2022).

interventions in both primary insurance and reinsurance markets.[45] All these interventions could support the resiliency of private insurance markets and ensure coverage for future wildfire losses.

Catastrophe Bonds and Parametric Coverage

Given the increasing severity of extreme wildfires (as well as other weather-related natural disasters), there is a growing lack of capacity in traditional insurance markets to cover losses and increasing concerns that traditional indemnity-based products may not provide sufficient coverage for indirect losses (Hedberg and Franco 2022). Two nontraditional solutions to expand capacity and provide sufficient compensation are indemnity-based and parametric catastrophe bonds.[46] Indemnity-based solutions provide a payout when losses exceed a predefined value and parametric coverage pays out once the parameter exceeds a predetermined value. Li and Su (2023) used publicly available data to model wildfire risk in the United States and found that both indemnity-based and parametric catastrophe bonds can substantially mitigate tail risk arising from large fires.

Catastrophe bonds expand capacity by spreading catastrophic risk borne by insurers and reinsurers to the capital market. Typically, wildfire perils are covered by multiperil catastrophe bonds that also target storms, earthquakes, and severe weather. Fire-only catastrophe bonds are relatively uncommon. There were three launched in 2021 which cover California wildfire risk (ARTEMIS 2023a), in which two of the bonds were sponsored by utility companies, and the third by an insurer. The total issue amount for the three series was 260.7 million USD. More recently Randolph Re, Aon's specialty reinsurer, offered a 25 million USD bond in 2022 (Evans 2022) and a 45 million USD bond covering California wildfire in 2023 (ARTEMIS 2023b).

Parametric coverage expands capacity by covering risks that are not fully covered by traditional insurance and/or by offering customized solutions. Key parameters for wildfire coverage are burned acreage and/or burn severity, combined with vegetation salvage, fire weather conditions, climate change, and location. Since the loss assessment process is objective and technology-driven, moral hazard, such as exaggerating the size of the insured loss or delaying the reporting to ask for higher compensation and fraud can be effectively reduced. More importantly, the claim payout process can be expedited by parametric coverage. Finally, Hedberg and Franco (2022) noted that parametric models tend to be updated more frequently than traditional data sources used in the insurance industry and, as such, may provide for more accurate pricing solutions.

Although forest ecosystems will often rebound on their own following a loss, an earlier release of funds provided by parametric insurance products will allow for immediate restoration of watersheds, stabilization of soil, and removal of debris, reducing the probability of flooding in upcoming seasons, and providing habitats for at-risk and threatened wildlife species. Funds can also be used for active replanting in burned areas, reducing the likelihood of encroachment by invasive species. Removing debris quickly after a wildfire reduces the fuel load and mitigates the possibility of another extreme wildfire. Repairing recreation infrastructure such as roads, trails, and campground facilities reduces tourism revenue losses.

At the time of this writing, insurers or brokers that offer parametric wildfire cover include Liberty Mutual Re, where the parameter is the burned area detected by satellite image (Liberty Mutual Re 2021), and Descartes Underwriting, where the parameter is a blend of a Large Fire Potential (LFP) index and a combustion index.[47] Given the sparsity of products available on the insurance market, parametric coverage is yet to achieve its full potential (Martinez et al. 2021). For example, Brèteau-

[45] The discussion on government intervention in primary insurance and reinsurance is summarized from Kelly and Kleffner (2023).

[46] Cummins and Barrieu (2013) and Barrieu et al. (2025) provide an in-depth discussion on securitized risk-transfer solutions and catastrophe bonds, respectively.

Amores et al. (2021) developed an insurance policy that covers the risk of drought-induced forest dieback based on a drought index. The policy could be integrated into an index-based cover for wildfires.

Community-Based Catastrophe Insurance

Insurance plays a critical role in recovery from natural disasters. However, many assets that can be damaged by wildfire may be uninsured or underinsured, and many indirect losses arising from natural disasters also fall outside the scope of traditional insurance products. This protection gap can slow economic recovery.

Bernhardt et al. (2021) suggested that a community-based catastrophe insurance (CBCI) program could be used to close the protection gap and improve community resiliency after a natural disaster such as a wildfire. CBCI policies are purchased by community entities and provide coverage for multiple or communal properties. A CBCI can be used for single or multiple perils and can be designed to incentivize mitigation measures to increase resiliency against catastrophic losses. In the case of wildfire, CBCI funds can be used to finance individual household losses, such as extended evacuation costs or Build Back Better initiatives, which are not covered by insurance. These funds can also be used to rehabilitate communal landscape areas such as parks and green spaces. Providing funds to commercial entities to cover uninsured losses would help municipal governments to maintain tax receipts thereby protecting against credit downgrades (Bernhardt et al. 2021).

Although not yet used for wildfire risk, the Center for NYC Neighborhoods (CNYCN) purchased a parametric risk transfer policy to protect low-income households against flood risk. Cash payments of 1.1 million USD will be made to CNYCN following a major flood event, which, in turn, will provide grants of up to 15,000 USD to qualified households (Snape 2023).

Public Primary Insurance Models

Public intervention in insurance markets is common in jurisdictions worldwide to address availability and affordability concerns for the provision of insurance for catastrophic losses such as earthquakes and other earth movements, floods, and terrorism.

A public entity may operate as a sole mandatory insurer to provide coverage for specific perils for all risks (such as the Danish Natural Hazard Council and the National Catastrophe Insurance of Iceland) or as an optional insurer competing with private insurers to provide coverage for specific perils (such as the CEA). A mandate requiring coverage be provided to all insureds would effectively alleviate availability concerns but may introduce affordability issues if coverage was actuarially priced.

Coinsurance agreements in which a government insurer either provides protection for a subset of assets, such as structures but not contents (e.g., Toka Tū Ake EQC in New Zealand and the cantonal monopoly insurers in Switzerland) or pays a portion of catastrophic losses for all insureds (e.g., the case of Belgium), would allow for private market participation. Since many government insurers have flat pricing models (with the CEA being the exception to the above-mentioned list), the actuarial pricing of private insurers would still provide a signal of the level of risk.

[47] McGowan Risk Specialist, an insurance broker, offers coverage based also based on burned areas, see https://www.mcgowanrisk.com/products/cat-specialty-property-products/parametric-insurance/parametric-first-dollar-wildfire-insurance/. For the products offered by Descartes Underwriting, see https://ircai.org/top100/entry/parametric-wildfire-insurance/

Given that wildfire is generally not considered a peak peril,[48] and to date, insurers have been willing to provide coverage except in the most wildfire-prone areas, risk-sharing pools—with or without government participation—may be an effective way to provide coverage. However, if coverage is actuarially priced, as with the FAIR plans in the United States, affordability concerns may still persist. In Belgium, the *Bureau de Tarification—Catastrophes Nautrelles* establishes the maximum premiums that can be charged, and the risk of insuring high-risk properties at a subsidized price is shared across all insurers in proportion to their market size.

One of the problems with catastrophic losses is the inability of insurers to transfer losses over time. Mechanisms like the Norwegian Natural Perils Pool (NNPP), which provides multiperil coverage for all properties (and not just high-risk properties), may solve this problem. Insurers collect the catastrophe premiums from each insured and use the funds to cover their share of claims, reinsurance, and administration costs associated with the pool. If the premiums earned exceed the insurer's relative share of claims mandated by the NNPP, the difference is allocated to the insurer's special natural perils reserve. Funds in this reserve can only be used to cover future natural perils claims. If the insurer's current year premiums fall short of paying their share of the pool's costs, the shortfall must be paid first out of the insurer's natural perils reserve and then out of its operating budget if the reserve remains inadequate.

Although wildfires, to date, have not created the level of losses of other natural disasters, in the face of growing losses from wildfires and the potential uninsurability of high-risk properties, future interventions may be required to support the provision of insurance for wildfire losses.

Public Reinsurance Models

Primary insurance market interventions in which the government creates a public insurer that underwrites risks and settle claims may be expensive. In this case, a more cost-effective solution to public intervention may be to offer reinsurance. A public reinsurer (like the Florida Hurricane Catastrophe Fund or the Caisse Centrale de Réassurance in France) would have the ability to transfer losses over time, which allows it to provide reliable and predictable coverage to primary insurers, removing concerns about the availability of reinsurance. However, if premiums are set to be actuarially fair, then affordability issues may still exist.

One example of such a reinsurer is Flood Re in the United Kingdom, established in 2013 to provide residential flood insurance to high-risk homes. Although primary insurers set premiums and policy conditions, Flood Re offers optional subsidized reinsurance to help keep premiums affordable. The subsidized reinsurance results in an annual underwriting loss, which is financed through a levy on primary insurers, proportionate to their market shares and the fixed premium insurers pay when they cede a flood risk.

7 Conclusion

Landscape fires are ubiquitous: approximately 4% of the world is on fire at any point in time (Sigmund et al. 2021). Wildfires, landscape fires that go out of control, are also everywhere. They can happen in forests, either boreal, coniferous temperate, deciduous temperate or tropical, and grasslands. As a necessary part of a dynamic ecosystem, wildfires, in general, are not as destructive as other natural

[48] Traditionally earthquake, flood, and windstorm (hurricane/typhoon) were considered peak perils. These events have severe underwriting loss potential and generally are the most researched, modelled, and understood natural catastrophes in the insurance industry. In contrast, secondary peril are events that happen relatively frequently, generate medium size losses, and can include secondary events following primary perils (e.g., tsunami). Historically, these perils have not been modeled significantly by the insurance industry. Wildfire would be considered a secondary peril.

catastrophes (e.g., earthquakes and hurricanes). Nevertheless, extreme wildfires can have significant impacts on the fauna and flora of the burned areas. Not only do they result in significant agricultural, timber, structure, and infrastructure losses, but the smoke and pollutants they generate also create long-term health impacts that can affect populations thousands of miles away.

Unlike other natural catastrophes, human-induced activity is responsible for an overwhelming portion of wildfires in most jurisdictions. Therefore, wildfire risk management entails not only wildfire suppression but also managing human activity that influences wildfire risk. These activities include wildfire mitigation by communities, the public, and property owners, postfire recovery aimed at building resiliency; and systematically collecting and analyzing wildfire-related data. Despite progress in climate modeling and satellite sensing, unpredictable anthropogenic ignition that differentiates wildfires from other catastrophes and the complex interactions between weather, topography, and vegetation all make wildfires more difficult to model than windstorms and earthquake losses. These factors pose nontrivial challenges to fire detection and prediction.

Most areas around the globe have seen increasing fire risk over the recent decades. Swiss Re (2023) showed that climate change is expected to increase wildfire risk more than any other weather-related peril except flood. Property losses from wildfire have always been covered by insurance but increases in wildfire risk in the WUI have generated concerns about the future availability and affordability of coverage, especially in California and Australia. To date, with the exception of FAIR plans in the United States, there is no mechanism to provide coverage for high-risk properties. Also, unlike other natural catastrophes, globally, there are no government interventions to support the coverage of insurance for wildfire losses. Possible solutions to provide coverage when the traditional insurance solutions either fail or are insufficient to cover losses include catastrophe bonds and parametric coverage, community-based catastrophe insurance, and government interventions into either the primary insurance or reinsurance market.

Overall, as with all catastrophes, insurance cannot be the sole solution for protecting against wildfire risks. Since wildfires cannot and should not be eliminated, mitigating ongoing losses from such risks requires the collaboration of governments, communities, and individuals. Jurisdictions must carefully balance the costs of suppression with the cost of mitigation, prevention, and preparedness activities to build fire-resistant and sustainable communities. Insurers have a valuable role to play not only in supporting recovery after a loss but also in providing incentives for community and individual property owner mitigation and investing in research to create more resilient communities.

Acknowledgements We thank the editor, Georges Dionne, and two anonymous reviewers for insightful comments on an earlier version of this chapter. Information on the Australian insurance market provided by Kevin Liu is also gratefully acknowledged.

Appendix A: Global Wildfire Risk

Wildfires can burn vegetation in and above the soil, and accordingly, can be categorized into three types: ground fires, surface fires and crown fires, and many wildfires consist of all three types concurrently (Northwest Fire Science Consortium 2019). Soil with significant organic matter provides fuel to ground fires, allowing them to smoulder for a long time before becoming a surface fire, sometimes overwintering following very hot and dry conditions. Surface fires, as the name indicates, burn dried or dead vegetation that is lying or growing just above the ground. And crown fires burn in the canopies and leaves of bushes and trees. With high wind, a crown fire spreads swiftly from one treetop to the other, making it more challenging to extinguish than ground fires.

Wildfires spread differently than structural fires as they create their own weather system. Burning debris—such as sparks, embers, pinecones, and bark—can be thrown several miles ahead of the wildfire. If these embers land on fuel sources, they in turn become spot fires, creating new fires in advance of the moving wildfire. Depending on the species and size of the trees, wind speed and direction and topography, new fires can start a significant distance ahead of the current perimeter of a wildfire.[50]

Historically, wildfires have been more prevalent in geographic regions with hot summers and dense vegetation, such as Australia, South Africa, and the dry forests and prairies of Europe and North America (Nix 2018). In recent decades, climate change has resulted in more incidents of the "perfect storm" for wildfires—low humidity, high temperature, and strong wind—in places that were historically less susceptible to fire risks (Westerling and Bryant 2008). The changing landscape of wildfire, particularly in the Arctic boreal areas, Europe, the Mediterranean basin, and South Asia, has rendered wildfire risk a global issue (Moritz et al. 2014).

We briefly outline the wildfire hazard in the four biomes susceptible to fire: boreal forests, temperate forests, tropical forests, and grasslands.

Boreal Forests

The boreal forest (i.e., taiga) is the world's largest terrestrial biome and lies just south of the Arctic Circle. The boreal forest in Russia spans both Asia and Europe, reaching from central Asia to Arctic permafrost and from Europe to the Bering Sea. In Canada and the United States, the boreal forest covers 1.5 billion acres and stretches from Alaska in the west to Newfoundland and Labrador in the east.

The Arctic receives little precipitation, and as a result, vegetation easily dries out, becoming more flammable. Over time, the slowly decomposing vegetation accumulates and becomes porous peat that is rich in organic materials and conducive to oxygen. The combustion of peat does not produce a direct flame, but rather, it smoulders at a relatively low temperature.

Historically, the fire season in boreal forests typically extended from early spring (with the lightning season and the introduction of human activity into the region) and extended to the fall. However, boreal fires can smoulder in organic soils for days, weeks, or even months until weather conditions are such that flames spread (McCarty et al. 2020).[51] These holdover or overwintering fires (also called zombie fires) are more prevalent following a season in which the summer was hotter than normal and when there was significant fire damage (Scholten et al. 2021). Both elements are necessary to create holdover fires—overwintering fires do not occur in years following cool summer seasons, even if there were large forest fires in the previous year.

Because the underlying soils contain more carbon-heavy peat, fires in boreal forests have the potential to release significantly more carbon than fires in any other ecosystem. In 2020 alone, wildfires in the Arctic released 2.5 times as much carbon dioxide (CO_2) as temperate forest wildfires in California in the same year (Borunda 2021).[52]

[50] In the 2009 bushfires in Australia, the maximum spot fire distance was 34 km. For spot fires in North America, the maximum reported distance is 20 km (Gabbert 2020). Page et al. (2018), in a U.S. study, found that 94% of spot fires occur within 500 m. Storey et al. (2020), in an Australian study, found that the median spot fire distance was 100 m with only 2% of spot fires occurring at distances greater than 3 km.

[51] Under snow cover, fires can hibernate for 7–8 months from one season to the next.

[52] The fires released almost 250 megatons of carbon dioxide between January and September 2020, with the majority of emissions from fires in Russia's Sakha Region (Copernicus 2020). For reference, in 2021, the annual CO_2 emission in the United Kingdom was 347 megatons (ourworldindata.org).

Arctic holdover fires also create a feedback loop: wildfires in peatlands release more carbon, which exacerbates global warming; as the temperature increases, more peatlands melt, and more wildfires are started. Nevertheless, historically, countries have not found it economically feasible to suppress all boreal wildfires: there is little habitation and human activity over much of the boreal forests, thus fewer resources have been allocated to fire suppression therein.[53] As such, boreal wildfires have been among the largest in the world in terms of area burned.

Temperate Forests

South of the boreal forests are temperate coniferous forests in North America, Europe, and Asia; and south of temperate coniferous forests in southern Canada, the United States, Europe, and Japan are temperate deciduous forests, characterized by broadleaf trees, shrubs, herbs, and mosses.

The risk of fire in temperate forests, as in other biomes, depends greatly on moisture content, with significantly less fire risk in moist forests than in drier forests. However, there is greater human activity in temperate forests than in boreal forests, which increases the likelihood of ignition. Additionally, activities including selective logging and active fire suppression have resulted in greater fuel buildup and combined with longer and drier fire seasons, have resulted in an increased potential for larger, high severity fires (Halofsky et al. 2020).

Because fires in temperate forests release less CO_2 into the atmosphere than fires in the boreal forest, their impact on global warming is lower. However, there is greater human habitation around temperate forests, leading to greater human and economic loss than wildfires in boreal forests.

Tropical Forests

Tropical rainforests, dominated by broad-leafed trees forming a diverse upper canopy, are situated near the equator between the Tropic of Cancer and the Tropic of Capricorn. Historically, wildfires have been infrequent in tropical rainforests because of the high humidity (see, for example, Cochrane 2003). Even in the dry season, foliage retains significant moisture, making the foliage resistant to ignition. Arrogante-Funes et al. (2022) note that moist tropical forests have not evolved to adapt to fires and, as such, are among the most ecologically vulnerable biomes to fire.

Nevertheless, due to forest degradation from logging and other human activities, wildfire risk has been on the rise, even though many fires are extinguished with the arrival of the rainy season. In recent decades, Indonesia has emerged a main contributor to tropical deforestation caused by wildfires, with the 1997/1998 and 2015 fires burning down 0.8 and 6.4 million acres of forest (around 0.86% and 7% of the country's total forest acreage), respectively. These fires were caused by slash-and-burn

[53] In 2015, Russia's Ministry of Natural Resources established control zones, where fires are allowed to burn if there is no threat to communities or economically important infrastructure (Tubridy 2022). Canada also has defined a full response zone, outside which wildfires are not actively suppressed. Much of the boreal forest lies outside of the full response zone (Canada, Natural Resources 2021). Boreal forest fires are three of the top five largest (by area) wildfires between 2000 and 2022 as given by Igini (2022): 2003 Siberian taiga fires (55 million acres) in 2003, Northwest Territories fires (8.5 million acres) in 2004 and Alaska fire season (6.6 million acres) in 2004. Two tropical fires—the 2019/2020 Australian bushfires (42 million acres) and the 2010 Bolivia forest fire (3.7 million acres)—round out the top five.

agriculture, and the resultant fire risk was exacerbated by prolonged dry seasons induced by El Niño (Edwards et al. 2020).[54]

An exception to the relative immunity of tropical rainforests against wildfire is the coastal eucalyptus forest of Australia that encircle the continent, except for the coast between Perth and Broome. Eucalyptus trees produce gummy resin and oily leaves, both of which are highly flammable. Long strips of bark that naturally hang from the tree, when ignited, can be carried by the wind for several miles resulting in multiple bushfires.[55]

Grasslands

Except for Antarctica, the grassland biome is found on every continent in the world. Temperate grasslands extend across the interior plains of North America, Northeast China, Ukraine, southeastern interior of Australia, and South America. Tropical and subtropical grasslands lie closer to the equator and are largely located in South America, Africa, and northern Australia.

Grassfires exhibit distinct characteristics compared to forest wildfires, attributable to different fuel quality and quantity. Grassfires tend to start earlier in the day than other types of fires, as grasses dry out faster than other vegetation. The moisture content of the grassland—specifically, the percentage of dead material—is a major driver of fire spread. Grassfires also travel very quickly, up to 25 km per hour (CFA 2023).[56] The lower intensity of grassfires compared to forest fires contributes to quicker recovery of grasslands than forests (Neary and Leonard 2020). Nevertheless, grassfires pose a major threat to livestock production, especially during emergency evacuations (O'Hara et al. 2021).

The fire risk of grasslands differs vastly. Some grasslands—such as the savannas in Brazil and the Okefenokee swamp in Georgia, United States—hold so much water that fire risk is notably lower, though not absent. In contrast, the Hulunbuir Grassland in northern China, which is subject to irregular precipitation and large temperature variations, experiences 340 grassfires per year with an annual burned area of 2.75 million acres, or 4.3% of the total acreage (Bian et al. 2013).

References

Abatzoglou JT, Williams AP (2016) Impact of anthropogenic climate change on wildfire across western US forests. Proc Natl Acad Sci USA 113(42):11770–11775. https://doi.org/10.1073/pnas.1607171113

Agence France-Presse (2015) Russia bans burning of dry fields after wildfires. NDTV News, 12 Nov 2015. https://www.ndtv.com/world-news/russia-bans-burning-of-dry-fields-after-wildfires-1242657. Accessed 13 Apr 2023

Aguilera R, Corringham T, Gershunov A, Benmarhnia T (2021) Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California. Nat Commun 12:1493. https://doi.org/10.1038/s41467-021-21708-0

[54] Slash-and-burn agriculture is used to clear land to grow food, by which forested land is clear cut, and the remaining vegetation is burned. Slash-and-burn agriculture not only leads to deforestation but also results in significant soil loss, landslides, water contamination, and air pollution.

[55] Eucalyptus trees are also predominant in California. They are considered an invasive species that displaces native plants, which, by altering soil conditions, further impedes the growth of native plants (Wills 2018). Eucalyptus trees, also imported into Portugal, has now become the country's most common tree species, and were blamed for the devasting wildfires in 2017 (Ames 2017).

[56] In contrast, according to The Weather Network (2023), most forest fires advance at a rate of less than 0.5 km/h, and few spread faster than 1 km/h. The fastest forest fires can burn at 6 km/h or more. Speed is a function of both moisture content and type of tree: coniferous trees burn 5–10 times faster than deciduous trees, and deciduous trees burn more slowly when in full leaf.

Albery GF, Turilli I, Joseph MB, Foley J, Frere CH, Bansal S (2021) From flames to inflammation: how wildfires affect patterns of wildlife disease. Fire Ecol 17(1):23. https://doi.org/10.1186/s42408-021-00113-4
Alexander ME, Buxton-Carr P (2011) Wildland fire suppression related fatalities in Canada: a preliminary report. In: Proceedings of 11th international wildland fire safety summit, Missoula, Montana, USA, 4–8 Apr 2011
Ames P (2017) Portugal's 'killer forest'. Politco EU. https://www.politico.eu/article/portugal-fire-eucalyptus-killer-forest/. Accessed 28 Mar 2023
Ankhtuya (2022) The number of wildfires in Mongolia rise by 73 percent, 29 Aug 2022. https://news.mn/en/797800/. Accessed 15 Apr 2023
Arrogante-Funes F, Aguado I, Chuvieco E (2022) Global assessment and mapping of ecological vulnerability to wildfires. Nat Hazards Earth Syst Sci 22(9):2981–2022
ARTEMIS (2023a) Catastrophe bond and ILS market dashboard, catastrophe bond & insurance-linked securities deal directory. https://www.artemis.bm/deal-directory/. Accessed 20 Apr 2023
ARTEMIS (2023b) Randolph Re (Series 2023–3). https://www.artemis.bm/deal-directory/randolph-re-series-2023-3/. Accessed 28 Sept 2023
Associated Press (2023) Wildfire-prone California to consider new rules for property insurance pricing, 22 Sept 2023. https://apnews.com/article/california-home-insurance-wildfire-risk-premiums-cf40911606e8e4d9c7c35ca57ca733e8. Accessed 27 Sept 2023
Australia, The Senate (2021) Lessons to be learned in relation to the Australian bushfire season 2019–20 – final report. https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Finance_and_Public_Administration/Bushfirerecovery/Report. Accessed 15 Apr 2023
Aziz S (2023) In the line of fire: firefighters face perilous conditions amid record season, 2 Aug 2023. https://globalnews.ca/news/9869441/canada-firefighter-deaths-dangers-wildfires/. Accessed 31 Aug 2023
Bailey R, Yeo J (2019) The burning issue: managing wildfire risk. Marsh & McLennan Insights. https://www.oliverwyman.com/our-expertise/insights/2019/sep/the-burning-issue-managing-wildfire-risk.html. Accessed 24 Mar 2023
Baltzer H, Gerard F, Weedon G, Grey W, Comball B, Bartholomé E, Bartalev S, Los S (2007) Coupling of vegetation growing season anomalies with hemispheric and regional scale climate patterns in Central and East Siberia. J Clim 20(15):3713–3729. https://doi.org/10.1175/JCLI4226
Bar-Massada A, Alcansena F, Schug F, Radeloff VC (2023) The wildland – urban interface in Europe: spatial patterns and associations with socioeconomic and demographic variables. Landsc Urban Plan 235:104759. https://doi.org/10.1016/j.landurbplan.2023.104759
Barrieu P, Braun A, Makariou D (2025) Catastrophe bonds. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer Books, New York: Chapter 6
Bay Area Council Economic Institute (2021) The true cost of wildfires. http://www.bayareaeconomy.org/report/the-true-cost-of-wildfires/. Accessed 23 Mar 2023
Beamish L (2018) Mental health still a priority two years after Fort McMurray wildfire. Edmonton J, 2 May 2018. https://edmontonjournal.com/news/local-news/mental-health-still-a-priority-two-years-after-fort-mcmurray-wildfire. Accessed 28 Apr 2023
Bernhardt A, Kousky C, Read A, Sykes C (2021) Community-based catastrophe insurance: a model for closing the disaster protection gap. Marsh McLennan Insights. https://www.marshmclennan.com/insights/publications/2021/february/community-based-catastrophe-insurance.html. Accessed 27 Apr 2023
Bian H, Zhang H, Zhou D, Xu J, Zhengxiang Z (2013) Integrating models to evaluate and map grassland fire risk zones in Hulunbuir of Inner Mongolia, China. Fire Saf J 61:207–216. https://doi.org/10.1016/j.firesaf.2013.09.004
Borunda A (2021) Zombie fires in the Arctic are linked to climate change. https://www.nationalgeographic.com/environment/article/zombie-fires-in-the-arctic-are-linked-to-climate-change. Accessed 28 Mar 2023
Brèteau-Amores S, Brunette M, François C, Leblois A, Martin-Stpaul N (2021) Index insurance for coping with drought-induced risk of production losses in French forests. BETA Work. Pap. https://hal.science/hal-03401881/
Browne MJ, Hoyt RE (2000) The demand for flood insurance: empirical evidence. J Risk Uncertain 20(3):291–306
Brunette M, Couture S (2023) Forest insurance for natural events: an overview by economists. Forests 14(2):289. https://doi.org/10.3390/f14020289
Butry DT, Mercer ED, Prestemon JP, Pay JM, Holmes TP (2001) What is the price of catastrophic wildfire? J For 99(11):9–17. https://www.fs.usda.gov/research/treesearch/4446
Cal Fire (2019) Top 20 most destructive California wildfires. http://large.stanford.edu/courses/2022/ph240/chunduru1/docs/calfire-24oct22.pdf. Accessed 4 Apr 2023
California Catastrophe Response Council (2023) Report to the legislature and department of finance on the formation, administration, and disposition of the wildfire fund ("statutory report"), 1 Jan 2023. https://www.cawildfirefund.com/annual-legislative-report. Accessed 26 Apr 2023

California Department of Insurance (2021a) Wildfire survivors now covered by new insurance provisions, 27 July 2021. https://www.insurance.ca.gov/0400-news/0100-press-releases/2021/release078-2021.cfm. Accessed 24 Apr 2023
California Department of Insurance (2021b) New data shows insurance companies non-renewed fewer homeowners in 2020 while FAIR Plan 'insurer of last resort' policies increased, 20 Dec 2021. https://www.insurance.ca.gov/0400-news/0100-press-releases/2021/release117-2021.cfm. Accessed 7 Sept 2023
California Department of Insurance (2022) Insurers currently offering discounts, 14 Nov 2022. https://www.insurance.ca.gov/01-consumers/105-type/95-guides/03-res/Insurers-Currently-Offering-Discounts.cfm. Accessed 28 Sept 2023
California Department of Insurance (2023) Commissioner Lara announces Sustainable Insurance Strategy to improve state's market conditions for consumers, 21 Sept 2023. https://www.insurance.ca.gov/0400-news/0100-press-releases/2023/release051-2023.cfm. Accessed 28 Sept 2023
California Department of Insurance (n.d.) Mandatory one year moratorium on non-renewals. https://www.insurance.ca.gov/01-consumers/140-catastrophes/MandatoryOneYearMoratoriumNonRenewals.cfm. Accessed 29 Mar 2023
California State Assembly (2020) The impact of wildfires on California agriculture report. Informational hearing, Sacramento, CA, 18 Nov 2020
California State Auditor (2022) Electrical system safety: California's oversight of the efforts by investor-owned utilities to mitigate the risk of wildfires needs improvement, Mar 2022. https://www.auditor.ca.gov/reports/2021-117/index.html. Accessed 27 Sept 2023
California Wildfire Fund (2023) California wildfire fund. https://www.cawildfirefund.com/. Accessed 26 Apr 2023
Cambridge Centre for Risk Studies (2020) Optimising disaster recovery. https://www.preventionweb.net/publication/optimising-disaster-recovery-role-insurance-capital-improving-economic-resilience. Accessed 26 Sept 2023
Cameron PA, Mitra B, Fitzgerald M, Scheinkestel CD, Stripp A, Batey C, Niggemeyer L, Truesdale M, Holman P, Mehra R, Wasiak J, Cleland H (2009) Black Saturday: the immediate impact of the February 2009 bushfires in Victoria, Australia. Med J Aust 191(1):11–16. https://doi.org/10.5694/j.1326-5377.2009.tb02666.x
Canada, Natural Resources (2021) Cost of wildfire protection, 6 Apr 2021. https://natural-resources.canada.ca/climate-change/impacts-adaptations/climate-change-impacts-forests/forest-change-indicators/cost-fire-protection/17783. Accessed 23 Mar 2023
Canada, Natural Resources (2023) National wildland fire situation report, 29 Sept 2023. https://cwfis.cfs.nrcan.gc.ca/report. Accessed 30 Sept 2023
Cappelletti A (2019) The increasing risk of wildfire catastrophes. [general insurance] Insights Newsletter, June 2019. https://www.soa.org/news-and-publications/newsletters/general-insurance/2019/june/gii-2019-iss-7/the-increasing-risk-of-wildfire-catastrophes/. Accessed 26 Apr 2023
Centers for Disease Control (2022) Fighting wildfires. https://www.cdc.gov/niosh/topics/firefighting/default.html. Accessed 28 Apr 2023
Carlson JA (2005) The economics of fire protection: From the great fire of London to rural/metro 1. Econ Aff 25(3): 39–44
CFA (Country Fire Authority) (2023) Grassfires: rural. https://www.cfa.vic.gov.au/plan-prepare/am-i-at-risk/grassfires-rural. Accessed 30 Aug 2023
Chen B, Wu S, Jin Y, Song Y, Wu C, Venevsky S, Xu B, Webster C, Gong P (2022) Wildfire risk for global wildland-urban interface (WUI) areas. Preprint. https://doi.org/10.21203/rs.3.rs-2147308/v1
Chiglinsky K, Ballentine C (2018) AIG, Chubb dispatch wildfire crews as California blazes spread. Bloomberg News, 12 Nov 2018. https://www.bloomberg.com/news/articles/2018-11-12/aig-chubb-dispatch-wildfire-crews-as-california-blazes-spread#xj4y7vzkg. Accessed 30 Sept 2023
Cioccio L, Michael EJ (2007) Hazard or disaster: tourism management for the inevitable in Northeast Victoria. Tour Manag 28:1–11. https://doi.org/10.1016/j.tourman.2005.07.015
Clague JJ (2021) How an 'atmospheric river' drenched British Columbia and led to floods and mudslides. The Conversation, 17 Nov 2021. https://theconversation.com/how-an-atmospheric-river-drenched-british-columbia-and-led-to-floods-and-mudslides-172021. Accessed 15 Apr 2023
Clark PE, Porter BA, Pellant M, Dyer K, Norton TP (2023) Evaluating the efficacy of targeted cattle grazing for fuel break creation and maintenance. Rangel Ecol Manag 89(1):69–86. https://doi.org/10.1016/j.rama.2023.02.005
Climate Council (2022) Uninsurable nation: Australia's most climate vulnerable places. https://www.climatecouncil.org.au/resources/uninsurable-nation-australias-most-climate-vulnerable-places/. Accessed 5 Apr 2023
Coates L (2020) Bushfire deaths in Australia, 2010–2020. PreventionWeb, 14 Feb 2020. https://www.preventionweb.net/news/bushfire-deaths-australia-2010-2020. Accessed 28 Apr 2023
Cochrane MA (2003) Fire science for rainforests. Nature 421(6926):913. https://doi.org/10.1038/nature01437
Cohen JD, Westhaver J (2022) An examination of the Lytton, British Columbia wildland-urban fire destruction. Institute for Catastrophic Loss Reduction

Congressional Research Service (2023) Wildfire statistics. In Focus 10244, 1 June 2023
Copernicus (2020) Copernicus reveals summer 2020's Arctic wildfires set new emission records. https://atmosphere.copernicus.eu/copernicus-reveals-summer-2020s-arctic-wildfires-set-new-emission-records. Accessed 23 Mar 2023
Copernicus (2023) 2023 Canada wildfires emissions have already doubled previous annual record, Press Release, 3 Aug 2023. https://atmosphere.copernicus.eu/2023-canada-wildfires-emissions-have-already-doubled-previous-annual-record. Accessed 29 Sept 2023
Cummins JD, Barrieu P (2013) Innovations in insurance markets: hybrid and securitized risk-transfer solutions. In: Dionne G (ed) The handbook of insurance, 2nd edn. Springer Books, New York, pp 547–602
Cultural Burning and Prescribed Fire (n.d.) What is Cultural Burning. https://prescribedfire.ca/cultural-burning/. Accessed 28 Sept 2023
Davies KW, Wollstein K, Dragt B, O'Connor C (2022) Grazing management to reduce wildfire risk in invasive annual grass prone sagebush communities. Rangelands 44(3):194–199. https://doi.org/10.1016/j.rala.2022.02.001
Davies KW, Johnson DD, Smith J (2023) Frequent and catastrophic wildfires in Great Basin Rangelands: time for a proactive management approach. Rangel Ecol Manag 89(1):1–2. https://doi.org/10.1016/j.rama.2022.05.001
Davlasheridze M, Miao Q (2019) Does governmental assistance affect private decisions to insure? An empirical analysis of flood insurance purchases. Land Econ 95(1):124–145. https://doi.org/10.3368/le.95.1.124
DellaSala DA, Baker BC, Hanson CT, Ruediger L, Baker W (2022) Have western USA fire suppression and megafire active management approaches become a contemporary Sisyphus? Biol Conserv 268:109499. https://doi.org/10.1016/j.biocon.2022.109499
Dhingra R, Keeler C, Staley BS, Jardel HV, Ward-Caviness C, Rebuli M, Rapazzo K, Hernandez ML, Chelminski A, Jaspers I, Rappold AG (2023) Wildfire smoke exposure and early childhood respiratory health: a study of prescription claims data. Environ Health 22:48. https://doi.org/10.1186/s12940-023-00998-5
Di Virgili G, Evans JP, Clarke H, Sharples J, Hirsch AL, Hart MA (2020) Climate change significantly alters future wildfire mitigation opportunities in southeastern Australia. Geophys Res Lett 47(15):e2020GL088893. https://doi.org/10.1029/2020GL088893
Diaz JM (2012) Economic impacts of wildfire. Southern fire exchange: a JFSP knowledge exchange consortium
Donovan GH, Patricia AC, David TB (2007) Wildfire risk and housing prices: a case study from Colorado Springs. Land Econ 83(2):217–233. https://www.jstor.org/stable/27647762
Donovan VM, Wonkka CL, Twidwell D (2017) Surging wildfire activity in a grassland biome. Geophys Res Lett 44(12):5986–5993. https://doi.org/10.1002/2017GL072901
Du J, Wang K, Cui B (2021) Attribution of the extreme drought-related risk of wildfires in spring 2019 over Southwest China. Bull Am Meteorol Soc 102(1):S83–S90. https://doi.org/10.1175/BAMS-D-20-0165.1
Edwards RB, Naylor RL, Higgins MM, Falcon WP (2020) Causes of Indonesia's forest fires. World Dev 127:104717. https://doi.org/10.1016/j.worlddev.2019.104717
Eisenman DP, Galway LP (2022) The mental health and well-being effects of wildfire smoke: a scoping review. BMC Public Health 22:2274. https://doi.org/10.1186/s12889-022-14662-z
European Environment Agency (2021) Forest fires in Europe. https://www.eea.europa.eu/ims/forest-fires-in-europe. Accessed 12 Apr 2023
Evans R (1987) The early history of fire insurance. J Leg Hist 8(1):88–91. https://doi.org/10.1080/01440368708530888
Evans S (2020a) European wildfire risk growing, but hard to adequately price: Chaucer. https://www.artemis.bm/news/european-wildfire-risk-growing-but-hard-to-adequately-price-chaucer/. Accessed 6 Apr 2023
Evans S (2020b) California Wildfire Fund shies away from reinsurance pricing at renewal. ARTEMIS. https://www.artemis.bm/news/california-wildfire-fund-shies-away-from-reinsurance-pricing-at-renewal/. Accessed 6 Apr 2023
Evans S (2022) Aon's Randolph Re issues $25m wildfire cat bond, likely for Mercury. ARTEMIS. https://www.artemis.bm/news/aons-randolph-re-issues-25m-wildfire-cat-bond-likely-for-mercury/. Accessed 28 Sept 2023
FAO (2021) The impact of disasters and crises on agriculture and food security, Rome. https://doi.org/10.4060/cb3673en
Felix ED, Afifi W (2015) The role of social support on mental health after multiple wildfire disasters. J Community Psychol 43(2):156–170. https://doi.org/10.1002/jcop.21671
Fernyhough J (2020) Climate risks 'means higher premiums'. The Australian Financial Review, 2 Mar 2020. Accessed 19 Apr 2023
FireSmart Canada (2023) The wildland urban interface (WUI). https://firesmartcanada.ca/about-firesmart/the-wildland-urban-interface-wui/. Accessed 28 Sept 2023
Flannigan MD, Logan KA, Amiro BD, Skinner WR, Stocks BJ (2005) Future area burned in Canada. Climate Change 72(1–2):1–16. https://doi.org/10.1007/s10584-005-5935-y
Florec V, Borton M, Pannell D, Kelso J, Milne G (2020) Where to prescribe burn: the costs and benefits of prescribed burning close to houses. Int J Wildland Fire 29:440–458. https://doi.org/10.1071/WF18192

Gabbert B (2020) Researchers study factors that affect long-distance spotting of wildfires. Wildfires Today. https://wildfiretoday.com/tag/spot-fire/. Accessed 4 Apr 2023
Gao Y, Huang W, Yu P, Xu R, Yang Z, Gasevic D, Ye T, Guo Y, Li S (2023) Long-term impacts of non-occupational wildfire exposure on human health: a systematic review. Environ Pollut 320:121041. https://doi.org/10.1016/j.envpol.2023.121041
Garcês A, Pires I (2023) The Hell of wildfires: the impact on wildlife and its conservation and the role of the veterinarian. Conservation 3:96–108. https://doi.org/10.3390/conservation3010009
Global Forest Watch (n.d.) Dashboard. https://www.globalforestwatch.org/dashboards/. Accessed 11 Sept 2023
Godoy MM, Martinuzzi S, Masera P, Defossé GE (2022) Forty years of wildland urban interface growth and its relation with wildfires in central-western Chubut, Argentina. Front For Glob Change 5:850543. https://doi.org/10.3389/ffgc.2022.850543
Gude P, Rasker R, Van Den NJ (2008) Potential for future development on fireprone lands. J For:198–205. http://headwaterseconomics.org/wphw/wp-content/uploads/PGude_2008_Forestry.pdf
Halofsky JE, Peterson DL, Harvey BJ (2020) Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA. Fire Ecol 16:4. https://doi.org/10.1186/s42408-019-0062-8
Hazra D, Gallagher P (2022) Role of insurance in wildfire risk mitigation. Econ Model 108:105768. https://doi.org/10.1016/j.econmod.2022.105768
Headwater Economics (2022) Wildfires destroy thousands of structures each year. https://headwaterseconomics.org/natural-hazards/structures-destroyed-by-wildfire/ Accessed 6 Apr 2023
Hedberg S, Franco GE (2022) More frequent wildfires could drive up the cost of insurance. BRINK News. https://www.brinknews.com/more-frequent-wildfires-could-drive-up-the-cost-of-insurance/ Accessed 25 Apr 2023
Hesseln H (2018) Wildland fire prevention: a review. Curr For Rep 4:178–190. https://doi.org/10.1007/s40725-018-0083-6
Hetzel G, Goldammer JG (2004) The use of prescribed fire on embankments along railway tracks for reducing wildfire ignition in Germany. Int For Fire News 30:65–69
Hitch G (2020) Bushfire royal commission hears that Black Summer smoke killed nearly 450 people. ABC News Australia. https://www.abc.net.au/news/2020-05-26/bushfire-royal-commission-hearings-smoke-killed-445-people/12286094. Accessed 17 Apr 2023
Howard (2014) Flammable planet: wildfires and the social cost of carbon, Sept 2014. Costofcarbon.org. Accessed 29 Apr 2023
ICA (2021) ICA welcomes Victorian bushfire funding but warns insurance taxes 'unsustainable'. 21 May 2021. https://insurancecouncil.com.au/resource/ica-welcomes-victorian-bushfire-funding/. Accessed 15 Oct 2024
Igini M (2022) Top 12 largest wildfires in history. Earth.Org. https://earth.org/largest-wildfires-in-history/. Accessed 26 Mar 2023
Ikeda K, Tanimoto H (2015) Exceedances of air quality standard level of $PM_{2.5}$ in Japan caused by Siberian wildfires. Environ Res Lett 10(10):105001. https://doi.org/10.1088/1748-9326/10/10/105001
Institute for Catastrophic Loss Reduction (2019) Protect your home from wildfire. https://www.iclr.org/wildfire/#1510741678219-44ef3623-da4e. Accessed 25 Sept 2023
Institute of Actuaries of Australia (2022) Home insurance affordability and socioeconomic equity in a changing climate, Aug 2022. https://www.actuaries.asn.au/public-policy-and-media/thought-leadership/green-papers/home-insurance-affordability-and-socioeconomic-equity-in-a-changing-climate. Accessed 26 Apr 2023
Insurance Information Institute (2022) Wildfire: state of the risk. Triple-I Issues Brief. https://www.iii.org/sites/default/files/docs/pdf/triple-i_state_of_the_risk_wildfire_01242022.pdf
Insurance Information Institute (2023a) Facts + statistics: global catastrophes. https://www.iii.org/fact-statistic/facts-statistics-global-catastrophes#Top%2010%20Costliest%20Global%20Wildfires%20by%20Insured%20Losses,%201900-2022%20(1). Accessed 27 Apr 2023
Insurance Information Institute (2023b) Spotlight on: catastrophes – insurance issues. https://www.iii.org/article/spotlight-on-catastrophes-insurance-issues#top. Accessed 27 Sept 2023
IPCC (2020) The concept of risk in the IPCC sixth assessment report: a summary of cross working group discussions. https://www.ipcc.ch/event/guidance-note-concept-of-risk-in-the-6ar-cross-wg-discussions/. Accessed 29 Sept 2023
IPCC (2022) Sixth assessment report – climate change 2022: impacts, adaptation and vulnerability. https://www.ipcc.ch/report/ar6/wg2/
Jacoby A (2022) Burning down the house: analyzing California's inverse condemnation strict liability rule for utility-caused wildfires. South Calif Interdiscip Law J 31(1):107–162
Jolly WM, Cochrane MA, Freeborn PH, Holden ZA, Brown TJ, Williamson GJ, Bowman DMJS (2015) Climate-induced variations in global wildfire danger from 1979 to 2013. Nat Commun 6(1):1–11. https://doi.org/10.1038/ncomms8537
Jones MW, Smith A, Betts R, Canadell JG, Prentice IC, Le Quéré C (2020) Climate change increases the risk of wildfires. ScienceBrief. https://sciencebrief.org/briefs/wildfires. Accessed 17 Apr 2023

JRC (Joint Research Center) (2017) JRC Science for policy report: forest fires in Europe, Middle East and North Africa 2016. https://publications.jrc.ec.europa.eu/repository/handle/JRC107591
JRC (Joint Research Center) (2022) JRC Science for policy report: forest fires in Europe, Middle East and North Africa 2021. https://publications.jrc.ec.europa.eu/repository/handle/JRC130846
Karp P (2020) Australia's volunteer firefighting force declined 10% in past decade. The Guardian, 28 Jan 2020. https://www.theguardian.com/australia-news/2020/jan/29/australias-volunteer-firefighting-force-declined-10-in-past-decade. Accessed 23 Apr 2023
Kelly M, Kleffner AE (2003) Optimal loss mitigation and contract design. J Risk Insur 70(1):53–72. https://doi.org/10.1111/1539-6975.00047
Kelly M, Kleffner AE (2023) Catastrophe insurance solutions with flexible market responses. Report prepared for Finance Canada
Kemper K (2020) How governments can take action to limit extreme wildfires. World Bank Blogs. https://blogs.worldbank.org/climatechange/how-governments-can-take-action-limit-extreme-wildfires. Accessed 12 Apr 2023
Kobziar LN, Godwin D, Taylor L, Watts AC (2015) Perspectives on trends, effectiveness, and impediments to prescribed burning in the southern US. Forests 6(3):561–580. https://doi.org/10.3390/f6030561
Kochi I, Donovan GH, Champ PA, Loomis JB (2010) The economic cost of adverse health effects from wildfire-smoke exposure: a review. Int J Wildfire 19(7):803–817. https://doi.org/10.1071/WF09077
Kootenay Rockies Tourism (2018) Economic impact of the 2017 wildfires on tourism in the Kootenay Rockies Tourism Region. https://www.krtourism.ca/data-insights/. Accessed 31 Mar 2023
Kousky C (2017) Disaster as learning experiences or disasters as policy outcomes? Examining flood insurance purchases after hurricanes. Risk Anal 37(3):517–530. https://doi.org/10.1111/risa.12646
Kousky C, Michel-Kerjan E, Raschky P (2018) Does federal disaster assistance crowd out flood insurance? J Environ Econ Manag 87:150–164. https://doi.org/10.1016/j.jeem.2017.05.010
KPMG (2020) Responding to a changing landscape: general insurance industry review 2020, Dec 2020
Kunreuther H, Michel-Kerjan E (2013) Managing catastrophic risks through redesigned insurance: challenges and opportunities. In: Dionne G (ed) The handbook of insurance, 2nd edn. Springer Books, New York, pp 517–546
Kunreuther H, Pauly M, McMorrow S (2013) Insurance and behavioral economics: improving decisions in the most misunderstood industry. Cambridge University Press, New York
Künzli N, Avol E, Wu J, Gauderman WJ, Rappaport E, Millstein J, Bennion J, McConnell R, Gilliland FD, Berhane K, Lurmann F, Winer A, Peters JM (2006) Health effects of the 2003 Southern California wildfires on 429 children. Am J Respir Crit Care Med 174(11):1221–1228. https://doi.org/10.1164/rccm.200604-519OC
Le Goff H, Flannigan MD, Bergeron Y (2009) Potential changes in monthly fire risk in the eastern Canadian boreal forest under future climate change. Can J For Res 39(12):2369–2380. https://doi.org/10.1139/X09-153
Lee A (2021) Growers, wineries share risk on smoke damage. Farm Industry News. https://www.farmprogress.com/grapes/growers-wineries-share-risk-on-smoke-damage. Accessed 17 Apr 2023
Li H, Su J (2023) Mitigating wildfire losses via insurance-linked securities: modeling and risk management perspectives. J Risk Insur. https://doi.org/10.1111/jori.12449
Liao Y, Walls MA, Wibbenmeyer M, Pesek S (2022) Insurance availability and affordability under increasing wildfire risk in California. Resources for the Future Issue Brief 22-09. https://www.rff.org/publications/issue-briefs/insurance-availability-and-affordability-under-increasing-wildfire-risk-in-california/. Accessed 5Apr 2023
Liberty Mutual Re (2021) Parametric forestry wildfire cover. https://www.libertymutualre.com/static/2021-05/LM+Re+Forestry+Factsheet.pdf. Accessed 23 Apr 2023
Macias-Fauria M, Michaletz ST, Johnson EA (2011) Predicting climate change effects on wildfires requires linking processes across scales. Wiley Interdiscip Rev Clim Chang 2(1):99–112. https://doi.org/10.1002/wcc.92
Madhusoodanan J (2021) Wildfire smoke ruins taste of some wines. The Washington Post. https://www.washingtonpost.com/health/wildfires-smoke-wine-taste/2021/09/10/ce4cd4ec-1007-11ec-bc8a-8d9a5b534194_story.html. Accessed 28 Mar 2023
Maestas J, Pellant M, Okeson L, Tilley D, Havlina D, Cracroft T, Brazee B, Williams M, Messmer D (2016) Fuel breaks to reduce large wildfire impacts in sagebrush ecosystems. Plant Materials Technical Note, (6). Boise
Mattioli W, Ferrara C, Lombardo E, Barbati A, Salvati L, Tomao A (2022). Estimating wildfire suppression costs: a systematic review. Int For Rev 24(1):15–29
Martinez N, Young S, Carroll D, Williams D, Pollard J, Christopher M, Carus F, Jones D, Heard S, Franklin B, Smith E, Porter D (2021) Wildfire resilience insurance: quantifying the risk reduction of ecological forestry with insurance. Nature Conservancy Research Paper. https://www.nature.org/content/dam/tnc/nature/en/documents/FINALwildfireresilienceinsurance6.27.21.pdf
McCarty JL, Smith TE, Turetsky MR (2020) Arctic fires re-emerging. Nat Geosci 13(10):658–660. https://doi.org/10.1038/s41561-020-00645-5

McCormick D (2021) How electric utility companies can stop wildfires before they start. https://sigblog.hexagon.com/how-electric-utility-companies-can-stop-wildfires-before-they-start/. Accessed 24 Apr 2023

Mell WE, Manzello SL, Maranghides A et al (2010) The wildland–urban interface fire problem – current approaches and research needs. Int J Wildland Fire 19:238. https://doi.org/10.1071/WF07131

Mina U, Dimri AP, Farswan S (2023) Forest fires and climate attributes interact in central Himalayas: an overview and assessment. Fire Ecol 19(1):14. https://doi.org/10.1186/s42408-023-00177-4

Mongabay.com (2020) The Asian forest fires of 1997–1998. https://rainforests.mongabay.com/08indo_fires.htm. Accessed 31 Mar 2023

Moody's (2023) Reinsurers defend against rising tide of natural catastrophe losses, for now. https://www.moodys.com/web/en/us/about/insights/data-stories/reinsurers-mitigate-lower-profits.html. Accessed 23 Mar 2023

Moriondo M, Good P, Durao R, Bindi M, Giannakopoulos C, Corte-Real J (2006) Potential impact of climate change on fire risk in the Mediterranean area. Clim Res 31(1):85–95. https://www.int-res.com/abstracts/cr/v31/n1/p85-95/

Moritz M, Batllori E, Bradstock R, Gill AM, Handmer J, Hessburg J, Leonard J, McCaffrey S, Odion DC, Schoennagel T, Syphard AD (2014) Learning to coexist with wildfire. Nature 515:58–66. https://doi.org/10.1038/nature13946

Mueller JM, John BL, González-Cabán A (2009) Do repeated wildfires change homebuyers' demand for homes in high-risk areas? A hedonic analysis of the short-and long-term effects of repeated wildfires on house prices in Southern California. J Real Estate Financ Econ 38(2):155–172. https://doi.org/10.1007/s11146-007-9083-1

Munich Re (2019) New hazard and risk level for wildfires in California and worldwide. https://www.unisdr.org/preventionweb/files/64490_newrisklevelforwildfiresandclimatec.pdf. Accessed 5 Sept 2023

Munich Re (n.d.) Wildfires and bushfires. Climate change increasing wildfire risk. https://www.munichre.com/en/risks/natural-disasters-losses-are-trending-upwards/wildfires-as-the-climate-changes-so-do-the-risks.html. Accessed 3 Apr 2023

National Fire Protection Association (n.d.) Preparing homes for wildfire. https://www.nfpa.org/Public-Education/Fire-causes-and-risks/Wildfire/Preparing-homes-for-wildfire. Accessed 3 Apr 2023

National Interagency Fire Center (2020) National large incident year-to-date report. https://gacc.nifc.gov/sacc/predictive/intelligence/NationalLargeIncidentYTDReport.pdf. Accessed 3 Apr 2023

National Park Service (2022) Wildfire causes and evaluations. https://www.nps.gov/articles/wildfire-causes-and- evaluation.htm. Accessed 3 Apr 2023

National Park Services (n.d.) Indigenous Fire Practices Shape our Land. https://www.nps.gov/subjects/fire/indigenous-fire-practices-shape-our-land.htm. Accessed 28 Sept 2023

Neary DG, Leonard JM (2020) Effect of fire on grassland soils and water: a review. In: Kindomihou VM (ed) Grasses and grassland aspect. IntechOpen. https://doi.org/10.5772/intechopen.90747

NFPA Research (2021) The fifth needs assessment of the US fire service. https://www.nfpa.org/News-and-Research/Data-research-and-tools/Emergency-Responders/Needs-assessment. Accessed 27 Apr 2023

Nix S (2018) When and where do wildfires occur? https://www.thoughtco.com/when-and-where-do-wildfires-occur-3971236. Accessed 29 Mar 2023

Northwest Fire Science Consortium (2019) NWFSC fire facts: what are? Types of fires, 26 Apr 2019. https://nwfirescience.org/biblio/nwfsc-fire-facts-what-are-types-fire. Accessed 25 Aug 2023

O'Hara KC, Ranches J, Roche LM, Schohr TK, Busch RC, Maier GU (2021) Impacts from wildfires on livestock health and production: producer perspectives. Animals 11(11):3230. https://doi.org/10.3390/ani11113230

OCHA (Office for the Coordination of Humanitarian Affairs) (2022) Latin America & The Caribbean weekly situation update (7–13 March 2022) as of 14 March 2022. https://reliefweb.int/report/world/latin-america-caribbean-weekly-situation-update-7-13-march-2022-14-march-2022. Accessed 15 Mar 2023

OECD (2018) The contribution of reinsurance markets to managing catastrophe risk. www.oecd.org/finance/the-contribution-of-reinsurance-markets-to-managing-catastrophe-risk.pdf. Accessed 15 Mar 2023

OECD (2021) Enhancing financial protection against catastrophe risks: the role of catastrophe risk insurance programmes. https://www.oecd.org/daf/fin/insurance/Enhancing-financial-protection-against-catastrophe-risks.htm. Accessed 15 Mar 2023

OECD (2023) Taming wildfires in the context of climate change. https://www.oecd.org/environment/taming-wildfires-in-the-context-of-climate-change-dd00c367-en.htm. Accessed 27 Sept 2023

Page WG, Wagenbrenner NS, Butler BW, Blunck DL (2018) An analysis of spotting distances during the 2017 fire season in the Northern Rockies, USA. Can J For Res 49:317–325. https://doi.org/10.1139/cjfr-2018-0094

Palmer J (2022) The devastating mudslides that follow forest fires. Nature 601:184–186. https://doi.org/10.1038/d41586-022-00028-3

Pandey P, Huidobro G et al (2023) A global outlook on increasing wildfire risk: current policy situation and future pathways. Trees For People. https://doi.org/10.1016/j.tfp.2023.100431

Papadopoulos K (2021) 500 million euros allocated for Attica and Evia residents devastated by fires. Greek City Times. https://greekcitytimes.com/2021/08/10/500-million-euros-allocated-for-attica-and-evia-residents/. Accessed 13 Apr 2023

Parisien M-A, Barber QE, Hirsch KG, Stockdale CA, Erni S, Wang X, Arsenault D, Parks SA (2020) Fire deficit increases wildfire risk for many communities in the Canadian boreal forest. Nat Commun 11:2121. https://doi.org/10.1038/s41467-020-15961-y

Parks Canada (2022) Fire and disturbance ecology. https://parks.canada.ca/nature/science/conservation/feu-fire/feuveg-fireveg/dirige-prescribed/role. Accessed 29 Sept 2023

Paxson C, Rouse CE (2008) Returning to New Orleans after Hurricane Katrina. Am Econ Rev 98(2):38–42. https://doi.org/10.1257/aer.98.2.38

Pettenella D, Marchetti M, Marino D, Marucci A, Ottaviano M, Lasserre B (2022) Proposal for a harmonized methodology to assess socioeconomic damages from forest fires in Europe. Europa EU Forest Focus Studies. https://effis.jrc.ec.europa.eu/reports-and-publications/forest-focus-studies

Porter K, Scawthorn C, Sandink D (2021) An impact analysis for the National Guide for Wildland-Urban Interface Fires. Institute for Catastrophic Loss Reduction, May 2021. https://nrc-publications.canada.ca/eng/view/object/?id=b40907a1-0214-4cd3-a01a-a233011d0d93. Accessed 25 Sept 2023

Radeloff VC, Helmers DP, Kramer HA, Mockrin MH, Alexandre PM, Van Butsic AB-M, Hawbaker TJ, Martinuzzi S, Syphard AD, Stewart SI (2018) Rapid growth of the United States Wildland-Urban Interface raises wildfire risk. Proc Natl Acad Sci USA 115(13):3314–3319. https://doi.org/10.1073/pnas.1718850115

Raschky PA, Schwarze R, Schwindt M, Zahn F (2013) Uncertainty of governmental relief and the crowding out of flood insurance. Environ Resour Econ 54(2):179–200. https://doi.org/10.1007/s10640-012-9586-y

Reid CE, Maestas MM (2019) Wildfire smoke exposure under climate change: impact on respiratory health of affected communities. Curr Opin Pulm Med 25(2):179–187. https://doi.org/10.1097/MCP.0000000000000552

Reid CE, Brauer M, Johnston FH, Jerrett M, Balmes JR, Elliott CT (2016) Critical review of health impacts of wildfire smoke exposure. Environ Health Perspect 124(6):1334–1343. https://doi.org/10.1289/ehp.1409277

Richardson L, Loomis JB, Champ PA (2013) Valuing morbidity from wildfire smoke exposure: a comparison of revealed and stated preference techniques. Land Econ 89(1):76–100. https://doi.org/10.3368/le.89.1.76

Ritchie H, Rosado P, Roser M (2022) Natural disasters. https://ourworldindata.org/natural-disasters. Accessed 15 Sept 2023

Rosevear AB (1932) A short history of fire insurance law. Can Bar Rev 10:405–413. https://canlii.ca/t/t6j6

Sastry N (2009) Tracing the effects of Hurricane Katrina on the population of New Orleans: the displaced New Orleans residents pilot study. Sociol Methods Res 38(1):171–196. https://doi.org/10.1177/0049124109339370

Sauter PA, Möllmann TB, Anastassiadis F, Mußhoff O, Möhring B (2016) To insure or not to insure? Analysis of foresters' willingness-to-pay for fire and storm insurance. Forest Policy Econ 73:78–89. https://doi.org/10.1016/j.forpol.2016.08.005

Scholten RC, Jandt R, Miller EA, Rogers BM, Veraverbeke S (2021) Overwintering fires in boreal forests. Nature 593:399–404. https://doi.org/10.1038/s41586-021-03437-y

Sever M (2020) What is left in the air after a wildfire depends on exactly what burned. Eos 101. https://doi.org/10.1029/2020EO138965

Shakesby RA (2011) Post-wildfire soil erosion in the Mediterranean: review and future research directions. Earth Sci Rev 105(3–4):71–100. https://doi.org/10.1016/j.earscirev.2011.01.001

Shaposhnikov D, Revich B, Bellander T, Bedada GB, Bottai M, Kharkova T, Kvasha E, Lezina E, Lind T, Semutnikova E, Pershagen G (2014) Mortality related to air pollution with the Moscow heat wave and wildfire of 2010. Epidemiology 25:359–364. https://doi.org/10.1097/EDE.0000000000000090

Sigmund G, Santín C, Pignitter M, Tepe N, Doerr SH, Hofmann T (2021) Environmentally persistent free radicals are ubiquitous in wildfire charcols and remain stable for years. Commun Earth Environ 2:68. https://doi.org/10.1038/s43247-021-00138-2

Simpson H, Bradstock R, Price O (2019) A temporal framework of large wildfire suppression in practice, a qualitative descriptive study. Forests 10(10):884. https://doi.org/10.3390/f10100884

Sinco Kelleher J, Boone R (2023) Hawaii officials say DNA tests drop Maui fire death count to 97. Associated Press, 18 Sept 2023. https://apnews.com/article/maui-hawaii-wildfire-death-toll-3dc505d4d83b6af5ee01fdaf173c4f01. Accessed 29 Sept 2023

Sjöström J, Granström A (2023) Human activity and demographics drive the fire regime in a highly developed European boreal region. Fire Saf J 136:103743. https://doi.org/10.1016/j.firesaf.2023.103743

Snape G (2023) Major firms join forces to back NYC parametric flood pilot. Insurance Business Magazine, 6 Apr 2023. https://www.insurancebusinessmag.com/us/news/catastrophe/major-firms-join-forces-to-back-nyc-parametric-flood-pilot-442062.aspx. Accessed 28 Sept 2023

St. Denis LA, Short KC, McConnell K, Cook MC, Mietkiewicz NP, Buckland M, Balch JK (2023) All-hazards dataset mined from the US National Incident Management System 1999–2020. Sci Data 10:112. https://doi.org/10.1038/s41597-023-01955-0

Staley DM, Negri JA, Kean JW, Laber JL, Tillery AC, Youberg AM (2016) Updated logistic regression equations for the calculation of post-fire debris-flow likelihood in the Western United States. United States Geological Survey Open-File Report 2016-1106. https://doi.org/10.3133/ofr20161106

Storey MA, Price OF, Bradstock RA, Sharples JJ (2020) Analysis of variation in distance, number, and distribution of spotting in Southeast Australian wildfires. Fire 3(2):10. https://doi.org/10.3390/fire3020010

Swiss Re (2017) Commercial insurance: innovating to expand the scope of insurability. Sigma 5/2017. https://www.swissre.com/institute/research/sigma-research/sigma-2017-05.html. Accessed 25 Aug 2023

Swiss Re (2021) Natural catastrophes and inflation in 2020: secondary perils in the spotlight, but don't forget primary-peril risks. Sigma 1/2021. https://www.swissre.com/institute/research/sigma-research/sigma-2021-01.html. Accessed 20 Apr 2023

Swiss Re (2023) Natural catastrophes and inflation in 2022: a perfect storm. Sigma 1/2023. https://www.swissre.com/institute/research/sigma-research/sigma-2023-01.html. Accessed 15 Mar 2023

Syed Z (2021) How the Schultz Fire reveals long-term costs of big wildfires, from floods to habitat loss. The Arizona Republic. https://www.azcentral.com/story/news/local/arizona-environment/2021/11/29/schultz-fire-reveals-hidden-costs-big-wildfires/6040015001/. Accessed 6 Apr 2023

Taylor CA Jr (2006) Chapter 12: targeted grazing to manage fire risk. In: Targeted grazing: a natural approach to vegetation management and landscape enhancement. American Sheep Industry Association, pp 107–114. https://www.webpages.uidaho.edu/rx-grazing/Handbook/ASITargetGrazingBook2006.pdf. Accessed 22 Sept 2023

The Weather Network (2023) Farmzone: forecasts and observations. https://www.farmzone.com/index.php?product=forestfire&pagecontent=firebasics. Accessed 30 Aug 2023

Thiessen T (2020) Australia bushfires burn tourism industry: $4.5 billion as holidayers cancel. Forbes News. https://www.forbes.com/sites/tamarathiessen/2020/01/20/australia-bushfires-hit-tourism-industry-as-holidayers-cancel/. Accessed 17 Apr 2023

Thomas D, Butry D, Gilbert S, Webb D, Fung J (2017) The costs and losses of wildfires: a literature review. NIST Special Publication 1215. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1215.pdf

Tolhurst KG, McCarthy G (2016) Effect of prescribe burning on wildfire severity: a landscape-scale study from the 2003 fires in Victoria. Aust For 79(1):1–14. https://doi.org/10.1080/00049158.2015.1127197

Tomshin O, Solovyev V (2022) Spatio-temporal patterns of wildfires in Siberia during 2001–2020. Geocarto Int 37(25):7339–7357. https://doi.org/10.1080/10106049.2021.1973581

Touma D, Stevenson S, Swain DL, Singh D, Kalashnikov DA, Huang X (2022) Climate change increases risk of extreme rainfall following wildfire in the western United States. Sci Adv 8(13):eabm0320

Tubridy M (2022) Wildfires in Russia 101, 29 Aug 2022. https://russiapost.info/regions/wildfires. Accessed 26 Apr 2023

Tyukavina A, Potapov P, Hansen MC et al (2022) Global trends of forest loss due to fire from 2001 to 2019. Front Remote Sens 3. https://doi.org/10.3389/frsen.2022.825190

United Nations Convention to Combat Desertification (UNCCD) (2017) Global land outlook, 1st edn. https://www.unccd.int/resources/publications/global-land-outlook-1st-edition. Accessed 30 Sept 2023

United Nations Environment Programme (UNEP) (2022) A rapid response report – spreading like wildfire: the rising threat of extraordinary landscape fires. UNEP Research Report. https://www.unep.org/resources/report/spreading-wildfire-rising-threat-extraordinary-landscape-fires. Accessed 15 Apr 2023

United Nations Office for Disaster Risk Reduction (UNDRR) (n.d.) Build Back Better. https://www.undrr.org/terminology/build-back-better. Accessed 23 Sept 2023

United States, Department of the Interior (2021) Wildland fire management. Fiscal year 2020 annual report and large fire review. https://www.doi.gov/wildlandfire. Accessed 27 Apr 2023

United States, Department of the Interior (2022a) Fuels management. https://www.doi.gov/wildlandfire/fuels. Accessed 27 Apr 2023

United States, Department of the Interior (2022b) Suppression. https://www.doi.gov/wildlandfire/suppression. Accessed 27 Apr 2023

United States, Fire Administration (2022) What is the WUI? https://www.usfa.fema.gov/wui/what-is-the-wui.html. Accessed 30 Aug 2023

United States, Forest Service (n.d.) Mechanical treatment. https://www.fs.usda.gov/managing-land/fire/mechanical-treatment. Accessed 6 Sept 2023

Urbanski S (2014) Wildland fire emissions, carbon, and climate: emission factors. For Ecol Manag 317:51–60. https://doi.org/10.1016/j.foreco.2013.05.045

Valente F, Laurini M (2021) Spatio-temporal analysis of fire occurrence in Australia. Stoch Env Res Risk A 35(9):1759–1770. https://doi.org/10.1007/s00477-021-02043-8

Van Leer K (2018) Southern California: Thomas Fire Loss Estimate and Mudslide Commentary, RMS, 22 Jan 2018. https://www.rms.com/blog/2018/01/22/southern-california-thomas-fire-loss-estimate-and-mudslide-commentary. Accessed 29 Sept 2023

Vaughan G, Draude AP, Ricketts HMA, Schultz DM, Adam M, Sugier J, Wareing DP (2018) Transport of Canadian forest fire smoke over the UK as observed by lidar. Atmos Chem Phys 18:11375–11388. https://doi.org/10.5194/acp-18-11375-2018

Vicente FB, Carbajal N, Pineda Martinez LF (2014) Estimation of total yearly CO_2 emissions by wildfires in Mexico during the period 1999–2010. Adv Meterol 2014:958457. https://doi.org/10.1155/2014/958457

Wang X, Thompson DK, Marshall GA, Tymstra C, Carr R, Flannigan MD (2015) Increasing frequency of extreme fire weather in Canada with climate change. Clim Chang 130(4):573–586. https://doi.org/10.1007/s10584-015-1375-5

Wang G, Wang P, Wang T, Li Z, Yu J, Liu C, Bolgov MV (2018) The variation of Lake Baikal water level and causal analysis since 1900. Resour Sci (in Chinese) 11:2177–2186. https://doi.org/10.18402/resci.2018.11.04

Wang B, Spessa AC, Feng P, Hou X, Yue C, Luo J-J, Ciais P, Waters C, Cowie A, Nolan RH, Nikonovas T, Jin H, Walshaw H, Wei J, Guo X, Liu DL, Yu Q (2022) Extreme fire weather is the major driver of severe bushfires in southeast Australia. Sci Bull 67(6):655–664. https://doi.org/10.1016/j.scib.2021.10.001

Weir JR, Bidwell TG, Stevens R, Mustain J (2012) Firebreaks for prescribed burning. Oklahoma Cooperative Extension Service, Oklahoma State University, Research Paper NREM-2890. https://extension.okstate.edu/fact-sheets/firebreaks-for-prescribed-burning.html. Accessed 21 Sept 2023

Westerling AL, Bryant BP (2008) Climate change and wildfire in California. Clim Change 87(1):231–249. https://doi.org/10.1007/s10584-007-9363-z

Westhaver A (2017) Why some homes survived: learning from the Fort McMurray wildland/urban interface fire disaster. Institute for Catastrophic Loss Reduction research paper series – number 56, Mar 2017. https://www.iclr.org/researchpapers/

Wibbenmeyer M, Anderson SE, Plantinga AJ (2019) Salience and the government provision of public goods. Econ Inq 57(3):1547–1567. https://doi.org/10.1111/ecin.12781

Wills M (2018) How eucalyptus trees stoke wildfire. JSTOR Daily. https://daily.jstor.org/how-eucalyptus-trees-stoke-wildfires/. Accessed 28 Mar 2023

Wittwer G, Waschik R (2021) Estimating the economic impacts of the 2017–2019 drought and the 2019–2020 bushfires on regional NSW and the rest of Australia. Aust J Agric Resour Econ 65:918–936. https://doi.org/10.1111/1467-8489.12441

World Bank (2016) The cost of fire: an economic analysis of Indonesia 2015 fire crisis. https://openknowledge.worldbank.org/entities/publication/391e19f4-18f7-56c2-8c28-bb89cc9e7d6f. Accessed 12 Apr 2023

World Bank (2020) Managing wildfires in a changing climate. World Bank Policy Note. https://www.profor.info/content/managing-wildfires-changing-climate. Accessed 12 Apr 2023

Wyland S (2020) Fire remains burned into New Mexico's memory 20 years later. Washington Times, 17 May 2020. https://www.washingtontimes.com/news/2020/may/17/fire-remains-burned-into-new-mexicos-memory-20-yea/. Accessed 27 Apr 2023

Xanthopoulos G, Athanasiou M, Kaoukis K (2022) Suppression versus prevention. The disastrous forest fire season of 2021 in Greece. Wildfire 2022(Q2):18–24

Yang X, Jin X, Zhou Y (2021) Wildfire risk assessment and zoning by integrating maxent and GIS in Hunan Province, China. Forests 12(10):1299. https://doi.org/10.3390/f12101299

Zhang T (2019) 30 Firefighters killed in huge Sichuan forest fire. That's Mags, 1 Apr 2019. https://www.thatsmags.com/china/post/27449/30-firefighters-killed-in-sichuan-forest-fire. Accessed 29 Apr 2023

Zhang C, Xing L, Chen J (2020) Expanding insurance protection for China's urban residents. Swiss Re Research Paper. https://www.swissre.com/institute/research/topics-and-risk-dialogues/china/expertise-publication-insurance -protection-china-urban-residents.html. Accessed 23 Apr 2023

Catastrophe Bonds

Pauline Barrieu, Alexander Braun, and Despoina Makariou

1 Introduction

Catastrophe bonds (cat bonds) are a securitized form of reinsurance through which disaster risk, often related to natural catastrophes, can be transferred to the capital markets. They were developed in the mid-1990s in response to natural disasters that caused significant insured losses, most notably Hurricane Andrew in 1992 and the Northridge earthquake in 1994 (Polacek 2018). Such events are characterized by a low frequency but high severity, meaning that they are rare but when they occur, they can cause significant damage, ranging from physical destruction of property to loss of life and economic disruption. From an insurance perspective, large-scale natural disasters are problematic because they give rise to stochastic dependence of claims on a regional level (Cummins 2006). This hampers diversification and generates heavy-tailed aggregate loss distributions (Beer et al. 2019).

Accordingly, nondiversification traps and capitalization frictions are considered to be primary reasons for the failure of the private market for catastrophe insurance (see Jaffee and Russell 1997; Froot 2001; Zanjani 2002; Harrington and Niehaus 2003; Froot and O'Connell 2008; Ibragimov et al. 2009).[1] Insurers underwriting disaster risk need to hold considerable amounts of capital to maintain their ability to settle losses even in extreme scenarios. They are reluctant to do so due to their high cost of equity capital.

Cat bonds are a possible remedy because most uninsurable catastrophe risks are still globally diversifiable through securities markets (Cummins 2006). The objective of this chapter is to provide an overview and assessment of cat bonds as a dominant instrument in the area of alternative capital

This chapter succeeds the chapter "Innovations in Insurance Markets: Hybrid and Securitized Risk-Transfer Solutions" by J. David Cummins and Pauline Barrieu in the Handbook of Insurance, 2nd edition.

[1] See Niehaus (2002) for an overview of the early literature on catastrophe risk allocation.

P. Barrieu
London School of Economics and Political Science, London, United Kingdom
e-mail: p.m.barrieu@lse.ac.uk

A. Braun · D. Makariou (✉)
University of St. Gallen, St. Gallen, Switzerland
e-mail: alexander.braun@unisg.ch; despoina.makariou@unisg.ch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_6

and insurance-linked securities. We explain the relevant terminology, introduce the reader to the mechanics of cat bonds and their various characteristics, showcase the evolution of the market, evaluate the risk-return profile, and discuss the state of the art in cat bond pricing. Throughout the whole chapter, we highlight relevant academic research combined with current evidence from industry practice. The chapter exhibits the following structure. In Sect. 2, we provide a high-level overview of alternative risk transfer and alternative capital and discuss the prominence of cat bonds among other instruments in these sectors. In Sect. 3, we introduce the characteristics of cat bonds, including their general mechanics, the most important trigger types, the main perils, and the types of embedded reinsurance contracts. In Sect. 4, we describe the cat bond market and its evolution, the investor base, trading and liquidity, as well as the risk-return profile of the asset class. In Sect. 5, we review the different approaches for cat bond pricing. Finally, in Sect. 6, we conclude the chapter.

2 Categorization

2.1 Alternative Risk Transfer

Cat bonds are one particular form of alternative risk transfer (ART). ART is a set of solutions for the transfer of risk that go beyond the traditional insurance and reinsurance sector (Cummins and Weiss 2009).[2] Figure 1 contains an ART taxonomy. The categories differ from left to right by the degree to which they utilize the capital markets as a risk bearer. Self-insurance means that a corporation sets aside funds to cover potential losses rather than purchasing coverage from an insurance company. This is often used for risks that are predictable and measurable. An institutionalized form of self-insurance are captives, i.e., insurance companies that are wholly owned and controlled by the corporation that is looking to self-insure. Self-insurance does not involve the capital markets.

Furthermore, structured reinsurance solutions, also known as hybrid covers, are customized reinsurance contracts, designed to meet specific needs of the primary insurer that cannot be addressed by traditional reinsurance, such as multiple types of risk, multiple years, and multiple triggers. They are typically used to manage complex or large risks, provide capital relief, enhance capital efficiency, mitigate earnings volatility, and improve the cedent's overall financial performance. Structured reinsurance solutions may incorporate features of both insurance and capital markets.

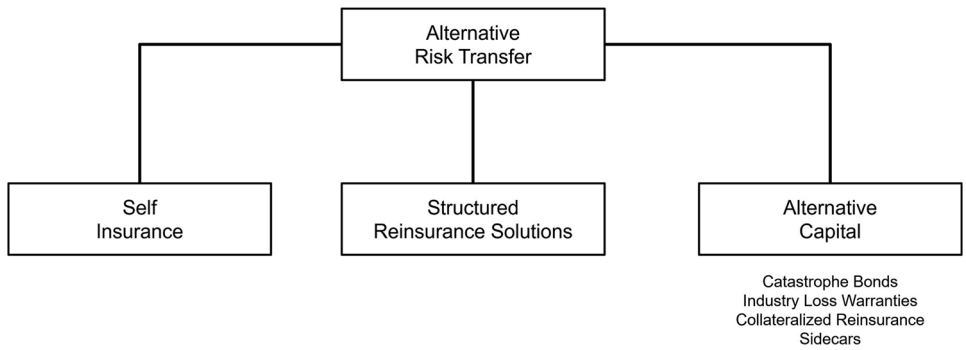


Fig. 1 A Taxonomy of alternative risk transfer

[2] A more detailed overview of ART is available in Cummins (2008) and Cummins and Weiss (2009). Note that we exclusively focus on the nonlife side of ART. More information on the much smaller market of ART instruments for life insurance risk can, e.g., be found in Barrieu et al. (2012).

The third distinct category within ART is alternative capital. The corresponding solutions are explicitly designed for a direct risk transfer from the cedent into the capital markets. Alternative capital comes into play in the presence of low-frequency high-severity risks that are hard to digest for insurers and reinsurers. Apart from cat bonds, alternative capital comprises industry loss warranties (ILWs), collaterized reinsurance, and sidecars. In the following section, we briefly discuss the latter three instruments, before moving on to cat bonds in Sect. 3.

2.2 Alternative Capital

Figure 2 illustrates the evolution of alternative capital over the last two decades. With a volume of around $100 bn, it amounts to between 15 and 20% of today's total reinsurance capital and is thus an important pillar in the risk transfer landscape. Figure 3 breaks the aggregate amount of alternative capital down into the individual instruments listed in Fig. 1. Cat bonds and collaterized reinsurance represent the two largest segments, while sidecars and ILWs are of lesser importance in terms of overall volume.

ILWs are highly standardized reinsurance contracts whose payout depends on two variables that need to exceed predefined theshold levels: the total loss to the insurance industry from a catastrophe as captured by an industry loss index and the individual losses incurred by the insurer purchasing the ILW cover (Beer et al. 2019). There is also a pure derivative version of the ILW, called a catastrophe swap, in which the payoff exclusively depends on industry losses (Braun 2011). Catastrophe swaps can be used by noninsurance entities such as hedge funds to gain exposure to catastrophe risk. Benefits of ILWs include their simplicity and the speed at which claims can be settled. However, because ILW payoffs are primarily driven by the index trigger, they involve basis risk (Gatzert and Kellner 2011).

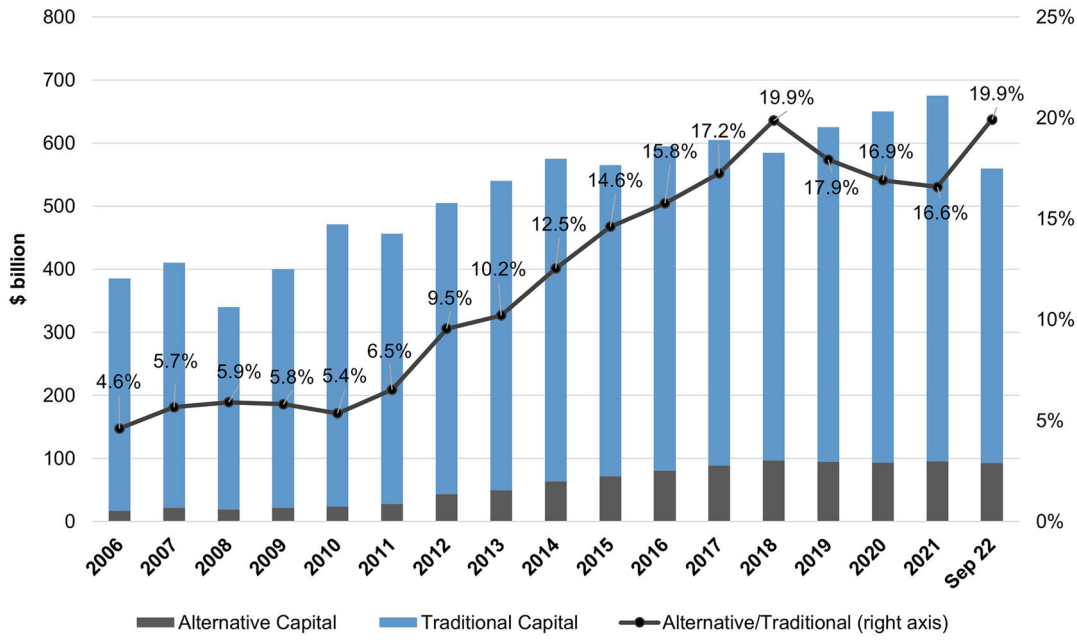


Fig. 2 Global reinsurance capital. (Source: AON 2023)

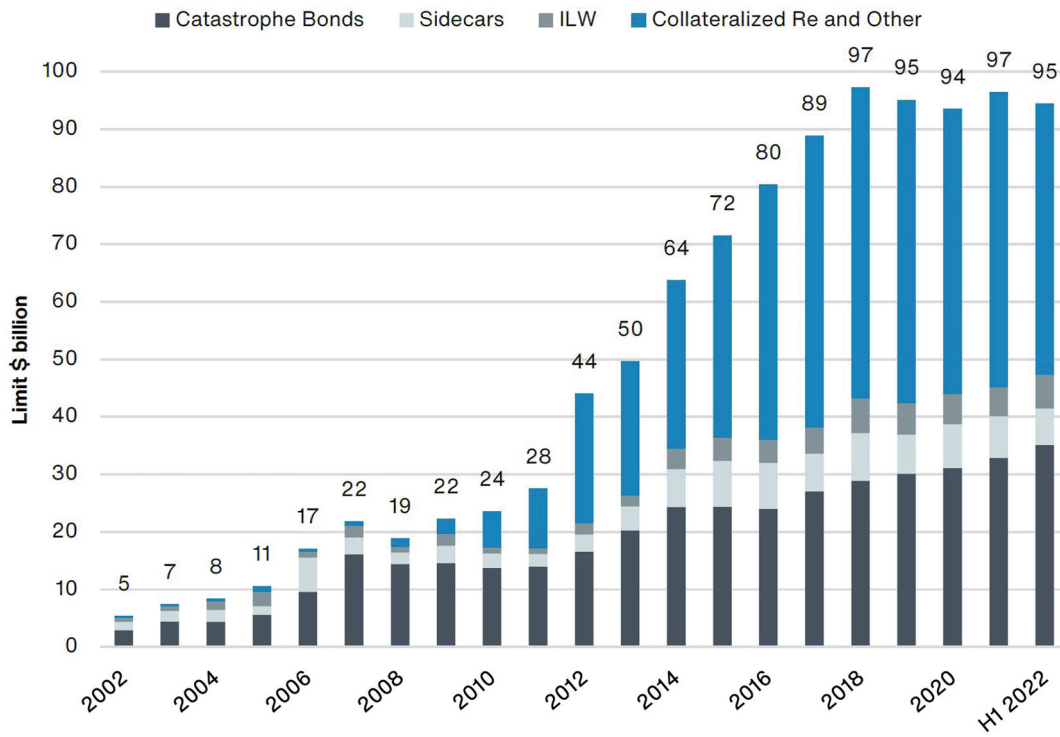


Fig. 3 Breakdown of alternative capital. (Source: AON 2022)

This means that the cedent may face uncovered losses when the aggregate industry loss is too low to trigger a payout.

Collateralized reinsurance allows capital market investors to sell reinsurance coverage and thus gain exposure to catastrophe risk. This is possible through an interposed transformer vehicle with a reinsurance license. In contrast to classical reinsurance, the full layer is collateralized (hence the instrument's name) by cash or highly rated assets. In contrast to ILWs, collateralized reinsurance contracts are bespoke instead of standardized, and in contrast to cat bonds, they are not securitized. Thus, there is no secondary market, and investors need to hold on to their position until maturity. Another difference to cat bonds is that collateralized reinsurance transactions tend to cover lower layers of the loss tower and thus exhibit a more aggressive risk-return profile. Academic literature on collateralized reinsurance is scarce. One study has been contributed by Bockius and Gatzert (2022) who analyze the impact of counterparty risk on the basis risk of ILWs as well as collaterized reinsurance under different dependence structures.

Sidecars are financial structures that enable investors to take exposure to catastrophe risk alongside reinsurance companies. The reinsurer sponsoring the sidecar transfers the underlying risk to a special purpose vehicle through a quota share reinsurance contract (Cummins and Weiss 2009).[3] The special purpose vehicle (SPV), in turn, is used to raise capital for the sidecar by issuing equity and debt to

[3] Quota share reinsurance is a form of proportional reinsurance under which two counterparties share claims and premiums according to a fixed percentage. As we will discuss below, the typical cat bond structure is similar to a sidecar but contains excess-of-loss reinsurance instead of proportional reinsurance. Sidecars and cat bonds can therefore function in tandem as complementary tools, similar to how quota share and excess-of-loss reinsurance contracts complement each other in a conventional reinsurance program.

investors. The returns on these investments depend on the premiums and claims that materialize on the underlying insurance portfolio. A tiered structure, akin to an asset-backed security, can be employed for the debt issued by the SPV. This approach caters to lenders with varying levels of risk tolerance. To date, sidecars have been primarily established in Bermuda and employed for property-catastrophe risk (Ramella and Madeiros 2007). However, Bugler et al. (2021) provide some first insights adapting the structure to the longevity sector.

3 Structural Characteristics

3.1 General Mechanics

We now begin our in-depth discussion of cat bonds. Where not indicated otherwise, the content of Sects. 3 and 4 is based on the recent cat bond primer by Braun and Kousky (2021). Cat bonds are insurance-linked securities (ILS), because their values are driven by insurance loss events. The standard structure of a cat bond is illustrated in Fig. 4. The sponsor or cedent is the entity aiming to offload catastrophic risk from its balance sheet. Sponsors are typically insurers or reinsurers, but there have been instances where corporations (e.g., Disney), public institutions (e.g., the New York City Metropolitan Transportation Authority), and sovereign nations (e.g., the Republic of Chile) relied on cat bonds to manage their risk exposure. To issue a cat bond, the sponsor establishes an SPV. The function of the SPV is to insulate both parties in the transaction from each other's counterparty

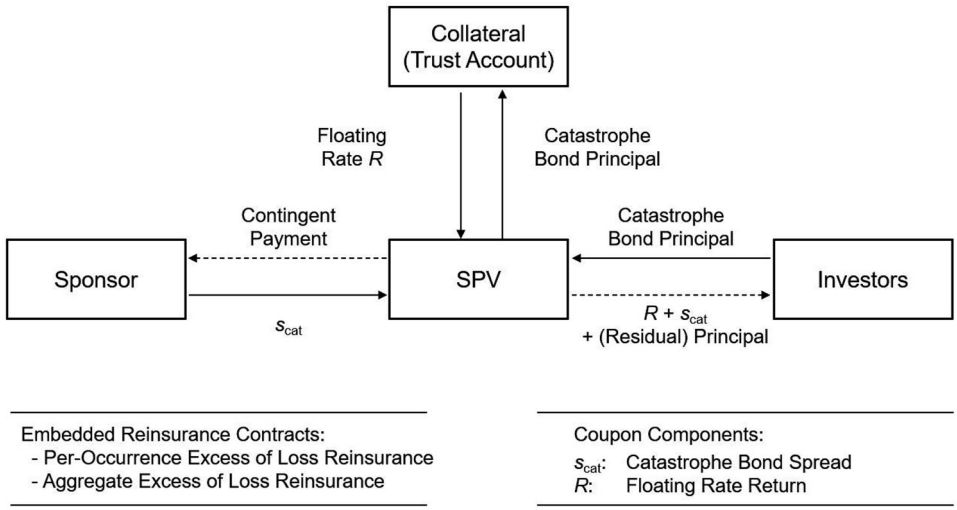


Fig. 4 Typical structure of a cat bond. The sponsor transfers the property-catastrophe risk in its books to the SPV by means of an excess-of-loss reinsurance contract. To back the risk with capital, the SPV issues fixed-income securities (the cat bonds) to investors. The principal paid by investors is held in safe collateral in a trust account. If a predetermined trigger event occurs during the bond's term, the collateral is liquidated and all or part of the corresponding proceeds are transferred to the sponsor as an indemnification under the embedded reinsurance contract. Consequently, for investors, a trigger event implies a complete or partial loss of the principal. If, in contrast, the bond reaches maturity without the trigger event taking place, the principal is refunded to the investors. For bearing the catastrophe risk, investors are compensated with regular coupons, consisting of a floating interest rate plus the cat bond spread. (Source: Braun 2016)

default risk.[4] In securitization markets, this is called bankruptcy remoteness. After its creation, the SPV grants coverage to the sponsor under an excess-of-loss reinsurance contract. To capitalize itself, the SPV issues the cat bond to investors. It then uses the proceeds from the investors to purchase high-quality collateral, which it holds in a trust account. Typical collateral securities for cat bonds are U.S. Treasury Bills, accessed through treasury money market funds. They exhibit a low default probability, a high liquidity, and, due to their short terms, little interest rate risk. An alternative collateral that has increased in use since the mid-2010s are Structured Notes provided by the International Bank for Reconstruction and Development (IBRD) (Swiss Re 2018).

The collateral produces a consistent flow of variable rate returns. Together with the rate on line or premium of the reinsurance contract, which is paid by the sponsor, these floating rate returns form the coupon received by the cat bond investors. In the cat bond context, the rate on line is called spread. Cat bond coupons thus comprise a term premium (the floating rate) and a risk premium (the spread). The latter compensates investors for shouldering the disaster risk. Due to their high-quality collateral, cat bonds themselves are floating rate notes with little interest rate and credit risk. In other words, cat bond investors benefit from an almost pure investment in property-catastrophe risk (Braun et al. 2022). This has important implications for the risk-return and correlation profile of the asset class, which we discuss in Sect. 4.

If a predetermined trigger event occurs during the bond's term, the collateral is liquidated and all or part of the corresponding proceeds are transferred to the sponsor as an indemnification under the embedded reinsurance contract. Consequently, for investors, a trigger event implies a complete or partial loss of the principal. If, in contrast, the bond reaches maturity without the trigger event taking place, the principal is repaid to the investors. Cat bonds have an average maturity of three years, with longer maturities of up to five years being less common (Braun 2016).

There are two types of coverage, determined by the reinsurance contract embedded within the cat bond: per occurrence and annual aggregate. In the per-occurrence case, the cat bond features an excess-of-loss per event (XL/E) reinsurance contract, which means it will only trigger if losses resulting from a single catastrophe event surpass a specified threshold (attachment point). On the other hand, the annual-aggregate case involves a stop-loss reinsurance contract, allowing for the accumulation of insured losses caused by all catastrophe events in a single year. At the time of writing, the outstanding market volume was nearly evenly divided between per-occurrence (51.3%) and annual-aggregate (48.7%) cat bonds (Artemis.bm 2023).

Cat bonds can exhibit a proportional or a binary payoff. In a proportional payout scenario, the percentage loss of principal for investors increases as the underlying insurance losses surpass a preset threshold, called the attachment point. Investors suffer a full write-down once the underlying insurance losses reach a second, higher threshold, called the exhaustion point. On the other hand, a binary payout transfers a predetermined amount of principal (usually 100%) to the sponsor as soon as the underlying insurance losses hit the attachment point.

Finally, as fixed-income securities, cat bonds are sometimes rated by credit rating agencies. For example, Moody's rating methodology for cat bonds is based on the expected loss to investors, which includes the likelihood that one or more catastrophe events will happen and the potential losses to investors arising from their occurrence (Artemis.bm 2016). However, since the expected loss is provided by catastrophe risk modelling firms and not the rating agencies themselves (see Sect. 5), more and more new issuances do not have any credit rating attached to them. This increase in nonrated cat bonds in recent years reflects the maturation of the market (Makariou et al. 2021). Today there

[4] If the bonds were issued directly by the sponsor instead of the SPV, investors would lose their principal in case the sponsor defaulted. Similarly, if the investors did not post the collateral upfront, their default would leave the sponsor without its hedge against property-catastrophe losses.

are a lot of specialized investors that purely rely on the risk assessment of the catastrophe modelling firms and do not derive any value added from an additional credit rating.

3.2 Triggers

Basic Types

In order to determine whether a trigger event has occurred, cat bonds can use a variety of trigger types. The trigger type defines the conditions of the catastrophic event that would lead to the release of the principal to the sponsor (Braun 2016). Figure 5 shows a breakdown of trigger types in the cat bond market at the time of writing.

The dominant mechanism today, accounting for a little more than 60% of the outstanding cat bond risk capital, is the indemnity trigger. A cat bond with an indemnity trigger works exactly like a standard reinsurance contract. It references the actual insured losses incurred by the sponsor and is thus a perfect hedge against disaster risk.

The second most common trigger type, the industry loss index trigger, represents 25% of the outstanding risk capital. Under this trigger, a payout to the sponsor is due if an index of aggregate insured losses incurred by all insurers in the covered territory surpasses a threshold value.[5] The index that determines the trigger event is compiled by an independent third party, the so-called calculation agent, from loss reports of the insurance industry. Typical calculation agents are Property Claims Services (PCS), a subsidiary of Verisk, in the U.S. and PERILS AG in Europe.

The least common trigger type today is the parametric trigger, which constitutes a mere 5% of the outstanding risk capital. Under this mechanism, a trigger event occurs, if a quantifiable characteristic of the covered catastrophe, such as earthquake strength on the moment magnitude scale or hurricane peak wind speed in miles per hour, exceeds a threshold at a given range of measurement stations (Hagedorn et al. 2009). To obtain these measurements, cat bonds may rely on government agencies,

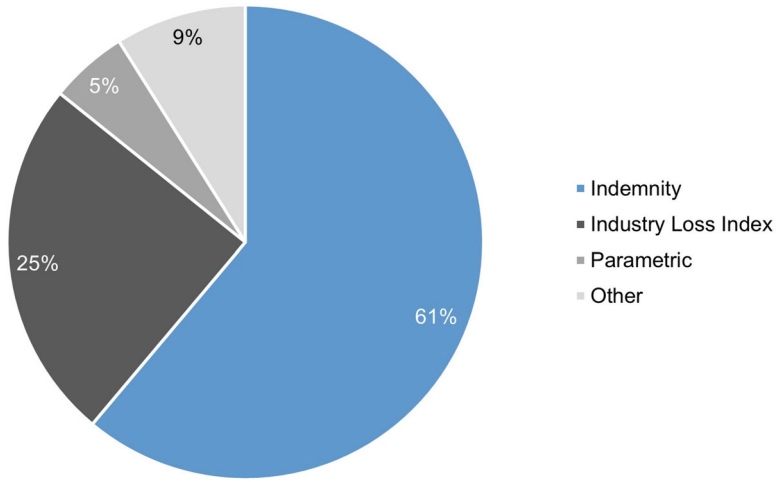


Fig. 5 Cat bond capital outstanding by triggers. (Source: Artemis.bm 2023)

[5] Note that the industry loss trigger explicitly references industry-wide *insured* losses, i.e., the losses covered by insurance policies. Due to prevailing protection gaps, these are usually much lower than the overall *economic* losses caused by natural disasters.

such as the National Oceanic and Atmospheric Association (NOAA) for meteorological data and the U.S. Geological Survey for seismological data (Braun 2023).

The vast majority of cat bonds today relies on one of these three trigger types. In the early years of the cat bond market, however, a fourth trigger type, the modelled loss trigger, was also prevalent. Under a modelled loss trigger, a catastrophe risk model is used to estimate the losses of the sponsor. These triggers have recently become rather rare. Finally, cat bonds may also employ multiple triggers at once. This is similar to double-trigger reinsurance contracts or ILWs, which comprise an industry loss index trigger and an indemnity trigger (Gatzert and Schmeiser 2012).[6]

Criteria for the Trigger Choice

Hagedorn et al. (2009) discusses the choice of triggers for cat bond transactions and highlights four key criteria: (i) transparency, (ii) basis risk, (iii) settlement time, and (iv) accounting and regulatory treatment. These criteria need to be traded off against each other, as sponsors and investors have different target functions and no single trigger type allows for the optimal outcome in all dimensions. This is most obvious for (i) and (ii): reducing basis risk leads to more intransparency and vice versa.

The concern with intransparency is a classical moral hazard problem. If the information asymmetry between sponsor and investor is large, the former may be inclined to take actions that increase the loss frequency or severity. Specific forms of moral hazard in the context of cat bonds are the relaxation of underwriting standards (pre-event) or claims handling standards (post-event) by the sponsor to the detriment of investors (Braun 2016). Basis risk, on the other hand, describes the possibility that the payoff from the cat bond is less than perfectly correlated with the sponsor's loss (Doherty 1997). In traditional reinsurance, there is no basis risk, since the payoff is determined by the actual losses in the sponsor's portfolio and hence perfectly matches his exposure.

Moral hazard, basis risk, and the trade-off of these two phenomena in the catastrophe risk transfer market have long been a subject of the academic literature (see Doherty 1997; Harrington and Niehaus 1999; Doherty and Richter 2002; Cummins et al. 2004). In recent years, Zhang and Tsai (2018) developed a model to quantify the hedge effectiveness of cat bonds with industry loss index triggers and analyzed its impact on pricing.

Further work has considered the issue of moral hazard from an empirical perspective. Götze and Gürtler (2020b) suggest an approach to measure moral hazard in the cat bond market. According to their analyses, sponsors of indemnity-trigger cat bonds are prone to ex ante but not ex post moral hazard. Braun (2016) finds no evidence for a spread markup on indemnity-trigger cat bonds but documents that long-standing sponsors with a strong track record achieve a tighter execution pricing. He concludes that the reputation of the sponsor seems to be more important to investors than moral hazard concerns associated with the trigger type. Braun et al. (2022) confirm the results of Braun (2016) for realized cat bond returns.

Evaluation of the Trigger Types

Indemnity triggers exhibit no basis risk for the sponsor but are intransparent for investors, since the latter do not have direct access to information about the sponsor's insurance portfolio. This raises the aforementioned moral hazard concern. If the sponsor is an insurer or reinsurer, it may loosen underwriting and claims handling standards once the cat bond coverage is in place. In addition to a potential moral hazard problem, indemnity triggers require a prolonged settlement process, i.e., loss reporting and verification for the sponsor's insurance portfolio must be completed before funds can be released. Since the indemnity trigger works just as classical reinsurance, it usually qualifies for

[6] The indemnity trigger in these instruments is included to achieve regulatory acceptance as a reinsurance substitute, and the industry loss index trigger mitigates moral hazard. We discuss the criteria for trigger choice in the next paragraph.

Table 1 Evaluation of the trigger types (Hagedorn et al. 2009)

Trigger/Criterion	Transparency	Basis risk	Settlement	Statutory acceptance
Indemnity	Low	Low	Slow	High
Industry index	High	Moderate	Slow	Moderate
Parametric	Very high	Very high	Fast	Low

reinsurance accounting[7] and the sponsor achieves full regulatory capital relief under modern risk-based capital standards, such as Solvency II (Braun and Weber 2017).

The moral hazard issue associated with indemnity triggers can be mitigated with an industry loss index trigger, because the sponsor can neither influence the losses of all other insurers nor the behavior of the calculation agent. Yet, industry loss index triggers introduce basis risk, because the underlying index will not be perfectly correlated with the sponsor's own losses. As the reporting of claims for a whole range of insurance portfolios is time-consuming, the industry loss index trigger, just as the indemnity trigger, requires an extended settlement time during which investor capital will be trapped in the cat bond's collateral.[8] Another disadvantage for the sponsor is that industry loss index triggers do not generally qualify for reinsurance accounting and regulatory capital relief (Smack 2016; Braun and Weber 2017).

Finally, the parametric trigger offers a high transparency and enables a rapid payout, as the necessary measurements come from independent third parties such as national weather services and are available immediately after the catastrophe has occurred. However, since the underlying physical parameters will exhibit a relatively low correlation with insurance losses, this trigger type presents a high basis risk for cedents and does not qualify for reinsurance accounting and regulatory relief. Table 1 summarizes these considerations for the three main trigger types.

3.3 Perils

Cat bonds are designed to offer coverage against the most expensive natural disasters globally. The underlying natural disaster risk is characterized by two dimensions: peril and reference territory. The former refers to the type of hazard insured (e.g., windstorms or earthquakes), whereas the latter specifies the location where the hazard event must occur to be covered by the bond. Figure 6 shows the prevailing peril/territory combinations in terms of outstanding risk capital at the time of writing. The coverage of windstorm and earthquake risks in the U.S., Japan, and Europe is a reflection of the scale and significance of these primary insurance markets relative to the rest of world. In recent years, however, the available range of peril/territory combinations has expanded through a growing number of cat bonds that cover natural catastrophe risks in emerging markets such as China (Panda Re 2015-1), Turkey (Bosphorus Re 2015-1), Chile (IBRD CAR 116), Colombia (IBRD CAR 117), Peru (IBRD CAR 120), Mexico (FONDEN 2020), and the Philippines (IBRD CAR 123-124).

[7] In the U.S., e.g., "credit for reinsurance can be claimed if the collateral for paying the reinsured claims is held in an U.S. trust" (Smack 2016).

[8] To reduce uncertainty surrounding the trigger probability, catastrophe risk modelling firms usually offer an initial estimate of aggregate insured losses immediately after the natural disaster. This estimate is then updated as new loss information becomes available.

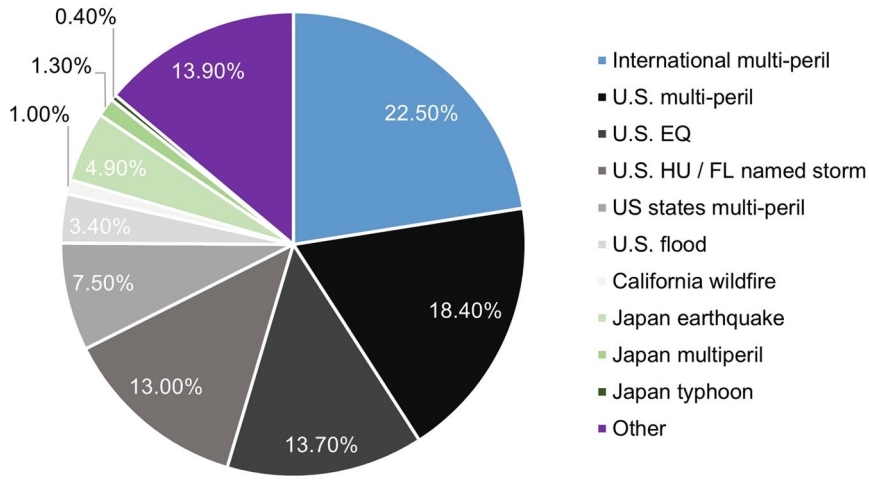


Fig. 6 Cat bond capital outstanding by Perils. (Source: Artemis.bm 2023)

4 Market Characteristics

4.1 Evolution of the Cat Bond Market

Figure 7 shows both the new issuance volume and the outstanding risk capital in the cat bond market from 1997 to 2022. The first cat bond was issued by Hannover Re in 1994 (Cummins and Weiss 2009). Since then, the market has expanded considerably. Issuance volumes grew from less than $1 billion in 1997 to more than $8 billion in 2007, followed by a sharp decline during the 2008 economic downturn. A correction of structural deficiencies[9] led to a rebound and prolonged upswing of the primary cat bond market, ending with a record issuance volume of $14 billion and a total outstanding capital of almost $36 billion in 2021. To put these numbers in perspective, we can draw on a comparison with Fig. 2. In 2021, cat bonds accounted for 5 to 6% of the total global property reinsurance policy limits. Since cat bonds are mainly used as coverage against the most extreme insurance events, they represent much larger proportions of the total limits for high-layer property reinsurance.

A trend that has added to the strong development of the cat bond market is the advent of cat bond lite structures. The cat bond lite allows for a maximally smooth and low-cost issuance process (Artemis.bm 2022b). In contrast to the regular 144A cat bond, it comes with a notably reduced documentation and does not provide a risk analysis by a catastrophe model vendor. It therefore primarily targets highly sophisticated ILS investors with own catastrophe risk modelling capabilities. Cat bond lite structures initially appeared in the early 2010s and surpassed $1 billion in issuance volume in 2020 for the first time (Artemis.bm 2023). At first, the issuance size of individual cat bond lite transactions was very small, but meanwhile it amounts to between $30 million and $50 million on average, with some outliers reaching the $100 million mark.

[9] Prior to late 2008, it was typical to use lower-grade collateral, such as structured finance securities, and safeguard it against interest rate risk and impairment using a total return swap (TRS). In exchange for fixed coupons and value gains from the assets in the trust account, the swap counterparty would offer a floating rate payment minus the TRS spread and cover any potential value losses (Towers Watson 2010). However, the collapse of investment bank Lehman Brothers, which served as a swap counterparty in four cat bond transactions, exposed that the combination of TRS and inadequate collateral, in terms of both quality and maturity, introduced a non-negligible degree of interest rate and credit risk (Cummins and Weiss 2009; Braun 2016).

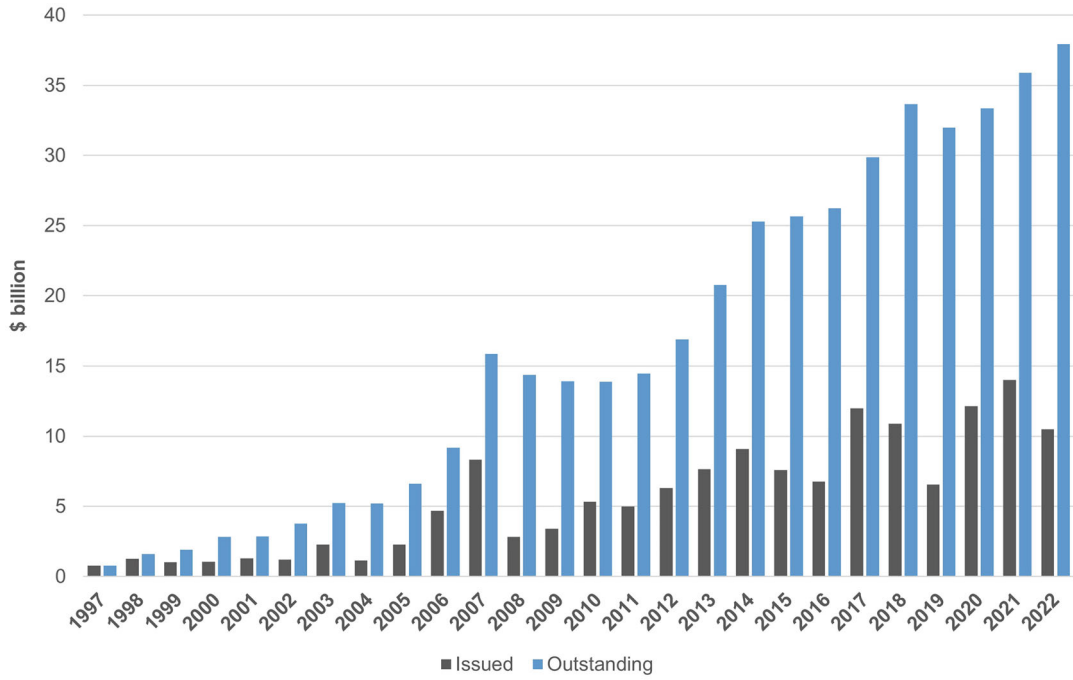


Fig. 7 Cat bond issuance and capital outstanding. (Source: Artemis.bm 2023)

Despite this remarkable evolution, there are no signs of slowing growth. In 2022, the cat bond market exhibited the largest overall size in history, with an outstanding volume of almost $38 billion (see Fig. 7). Cat bonds have long become an important part of the strategic arsenal of risk-hedging tools regularly used by insurers and reinsurers. With new perils and rapid technological innovation, the market's expansion can be expected to continue (Braun and Kousky 2021).

4.2 Investor Base

Throughout the 2010s, investor demand for cat bonds as an alternative investment was particularly high, likely driven by the persistent low-interest rate environment. Direct investments in cat bonds demand considerable expertise and can thus be challenging for investors who are not familiar with extreme event risk. A cat bond portfolio may consistently grow in value over an extended period, displaying very low return volatility. Yet, in the event of a significant catastrophe loss year like 2017, the investment could suddenly face a substantial drawdown. Moreover, some cat bond structures can be quite complex, encompassing multiple perils, territories, and triggers.

Apart from investing directly, investors can attain exposure to cat bonds through open-end funds. While these funds are occasionally categorized alongside mutual funds or hedge funds within the fixed-income domain, the returns generated by funds that hold cat bonds exhibit a rather distinct pattern (Braun et al. 2019). This unique behavior is characterized by low volatilities and low correlations to the broader capital markets. Most cat bond funds pursue buy and hold strategies. In some cases, however, sophisticated trading strategies such as live and dead cat trading are attempted (Jaeger et al. 2010). The former implies speculating on the pricing movement of a cat bond, while a

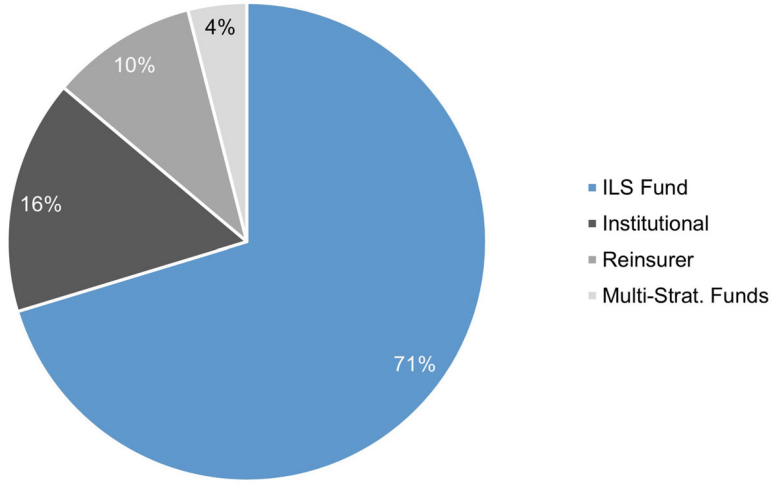


Fig. 8 ILS investor base by investor category. (Source: AON 2021a)

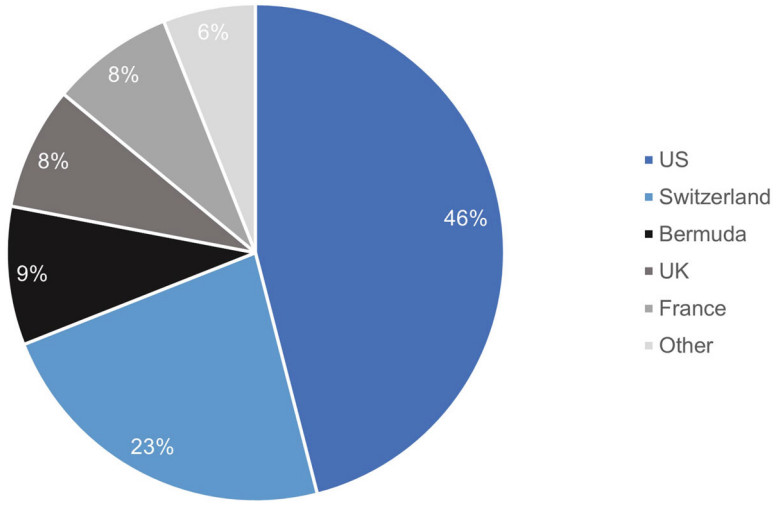


Fig. 9 ILS investor base by geography. (Source: AON 2021a)

natural disaster is evolving (e.g., a hurricane before landfall). The latter implies buying a cat bond at a major discount during the reporting process for the underlying claims and speculating on a price recovery if insured claims fall short of the attachment point.

Figures 8 and 9 illustrate the investor base of the cat bond market at the time of writing. Dedicated ILS fund managers currently absorb nearly three quarters of the newly issued volume. These experts possess extensive experience in reinsurance risk, making them well prepared to create and manage diversified cat bond portfolios for other institutional investors like pension funds.[10] The latter appreciate the low correlation of cat bond returns with the rest of their asset portfolios and their liquidity advantage compared to other alternative capital instruments. Apart from specialized ILS funds, the cat bond investor base comprises institutional investors (16%), reinsurers (10%), and

[10] 14 cat bond funds are UCITS-compliant and thus open to retail investors (Plenum Investments 2022).

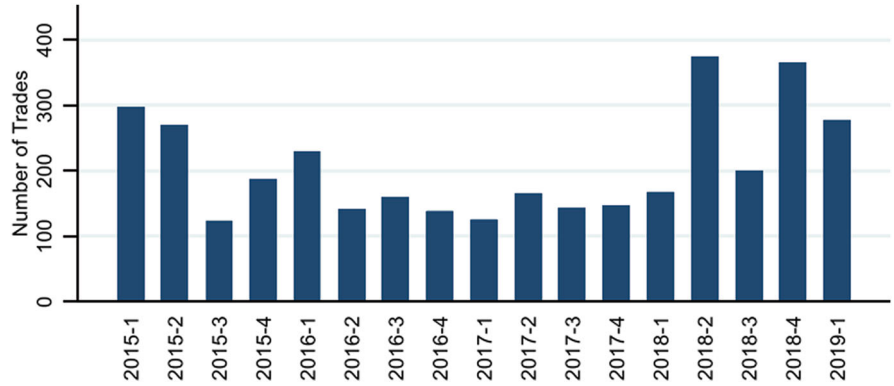


Fig. 10 Number of cat bond trades per quarter (Q1/2015−Q1/2019). (Source: Herrmann and Hibbeln 2023)

multi-strategy funds (4%). In terms of geography, the U.S. (46%) and Switzerland (23%) are the leading cat bond investor hubs, followed by Bermuda (9%), France (8%), and the UK (8%). The "Other" category comprises countries such as Canada, Japan, Germany, and Sweden.

4.3 Trading and Liquidity

A major advantage of cat bonds is the existence of an active secondary market. This means that investors can trade out of positions before maturity if necessary. The most impactful academic analysis of liquidity in the secondary cat bond market to date is Herrmann and Hibbeln (2023).[11] Figure 10, taken from their paper, shows the quarterly number of trades in the secondary market for cat bonds in the period from Q1/2015 to Q1/2019. Utilizing transaction data reported on TRACE (Trade Reporting and Compliance Engine), they observe that the seasonality of certain perils influences intrayear trading patterns. U.S. wind cat bonds, for example, are less frequently traded during the hurricane season and are more commonly traded as they approach maturity. Moreover, according to Herrmann and Hibbeln (2023), the prevailing trading patterns suggest that the secondary market is primarily controlled by brokers who do not act as market makers by maintaining a proprietary inventory. They also find that liquidity for individual bonds depends on the overall activity in the market and that cat bonds with a lower probability of first loss are more liquid than those with higher probabilities of first loss. Finally, using realized bid-ask spreads as a measure of liquidity, they estimate that 21% of the observable yield spread in the secondary market are a liquidity premium.

Building on Herrmann and Hibbeln (2023), Braun et al. (2022) investigate whether liquidity is a significant driver of realized returns of cat bonds. Their analysis also relies on TRACE data (from July 2014 to December 2020), and bid-ask spreads as a measure for liquidity. Sorting the whole cross section of cat bonds in a given month into quintile portfolios according to their liquidity and taking the difference between the return series of the least and the most liquid portfolio, Braun et al. (2022) create a liquidity factor. In doing so, they find that wider bid-ask spreads, which indicate less liquidity, are associated with higher expected excess returns. Their model, including the liquidity factor, can be applied to measure the liquidity premium in cat bond returns.

[11] While Zhao and Yu (2019) also measure the liquidity of cat bonds, they use liquidity proxies from the primary market and assess their impact on yield spreads.

4.4 Risk-Return and Correlation Profile

Owing to the existence of the secondary market for cat bonds, monthly return data is available and can be used for an empirical assessment of the asset class' risk-return profile.[12] Figure 11 shows the historical performance of cat bonds compared to the S&P 500 stock market index, the Barclays U.S. Treasury Index, and the Barclays U.S High Yield Corporate Bond Index. During the period under consideration, cat bonds delivered equity-like annual returns of between 6.8% p.a., paired with a low volatility (3.7%) and a negligible correlation of less than 0.2 to other asset classes.

Historical cat bond return time series have been analyzed by several empirical studies. Cummins and Weiss (2009) draw on the Swiss Re Cat Bond Index suite as well as the return time series of a small set of dedicated ILS funds to assess the risk-return profile of cat bonds. A similar effort is made by Kish (2016), using the Eurekahedge ILS Advisors Index.[13] Moreover, Mariani and Amoruso (2016) and Trottier et al. (2019) also use Swiss Re data. The latter go beyond a pure descriptive analysis and develop a regime-switching model for the returns of cat bonds. In addition to the risk-return profile, a number of studies have considered the correlation of cat bond returns with the broader capital markets (see Carayannopoulos and Perez 2015; Sterge and van der Stichele 2016; Mouelhi 2021; Haffar and Le Fur 2022) and the diversification effects of cat bonds in multi-asset class portfolios (see Clark et al. 2016; Drobetz et al. 2020).

However, knowledge about the drivers of expected excess returns on cat bonds is still scarce. So far, the only empirical studies in this area are Braun et al. (2019) and Braun et al. (2022). The former

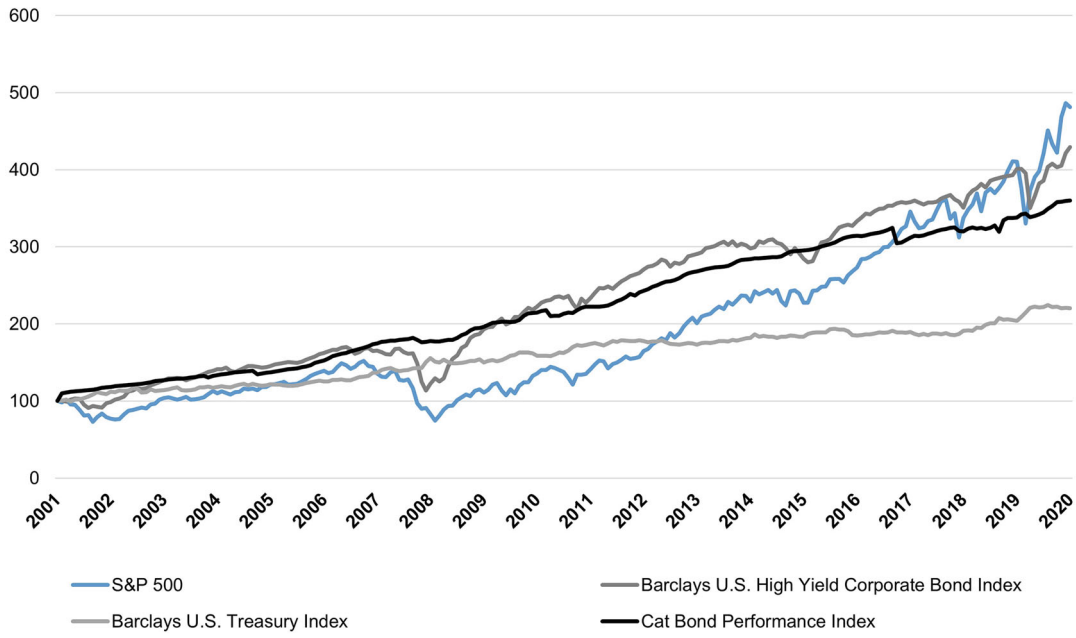


Fig. 11 Historical cat bond performance. (Sources: Bloomberg and Braun et al. (2022))

[12] Since historical return data was not available in the early years of the market, first studies of the risk-return profile, such as Litzenberger et al. (1996), relied on hypothetical returns.

[13] The Eurekahedge ILS Advisors Index is an equally weighted portfolio of 27 ILS funds that have at least 70% of their assets under management allocated to nonlife insurance risks. For more information see the Eurekahedge website.

develop novel factor models in the spirit of arbitrage pricing theory to explain the expected excess return of ILS funds based on the cat bond market and various additional factors. In doing so, they are able to show that most ILS funds do not generate alpha, but earn excess returns associated with hitherto unknown ILS-specific risk factors. Braun et al. (2022) analyze the drivers of realized returns on the bond level, using univariate sorting and multivariate Fama-MacBeth regressions. They then develop a factor model for the expected excess returns of cat bonds that includes a factor, reflecting a cat bond's time-varying probability of first loss as in Herrmann and Hibbeln (2021), and a liquidity factor based on Herrmann and Hibbeln (2023) (see description in Sect. 4.3). Finally, both Braun et al. (2019) and Braun et al. (2022) document a (weak) link between the cat bond market and the corporate bond market, which is attributed to the fact that the vast majority of rated cat bonds falls into the high-yield risk category.

5 Pricing

5.1 Preliminaries

As catastrophe bonds connect insurance and capital markets, they can be priced by means of (i) actuarial methods, (ii) contingent claims approaches, (iii) utility-based approaches, and (iv) statistical approaches. All methods have one thing in common: they derive the spread (or price) of a cat bond based on an estimate of the underlying distribution of insured losses.[14]

Below, we illustrate how this underlying distribution of insurance losses relates to key metrics relevant for cat bond pricing. Let $\tilde{C}$ denote the stochastic claims in an insurance portfolio covered by the cat bond. The cat bond layer is defined by the attachment point A and the exhaustion point E. Based on these three variables, the stochastic percentage loss $\tilde{L}$ to investors on a cat bond with proportional payoff is defined as follows:[15]

$$\tilde{L}(\tilde{C}) = \begin{cases} 0, & \text{for } \tilde{C} \leq A \\ \frac{\tilde{C}-A}{E-A}, & \text{for } A < \tilde{C} \leq E \\ 1, & \text{for } E < \tilde{C}. \end{cases} \tag{1}$$

Intuitively, if the underlying insured claims $\tilde{C}$ remain below A, there is no trigger event on the cat bond. If the claims $\tilde{C}$ materialize between A and E, investors suffer a percentage write-down proportional to the amount by which A is exceeded. Finally, if the claims exceed E, the cat bond is a total loss to investors.

Given the payoff function of the cat bond in (1) and the probability density function (pdf) of the underlying insurance claims, $f_{\tilde{C}}(x)$, one can derive the standard cat bond risk metrics.[16] The probability of first loss (PFL) and the probability of exhaustion (POE) are formally defined as follows:

$$PFL = \Pr(A < \tilde{C}) = \int_A^\infty f_{\tilde{C}}(x)dx \qquad POE = \Pr(E < \tilde{C}) = \int_E^\infty f_{\tilde{C}}(x)dx. \tag{2}$$

[14] In case of a parametric trigger, the distribution of physical parameter values is used instead.

[15] If the cat bond exhibits a binary payoff, then this expression simplifies to two cases: i) 0 for $\tilde{C} < A$ and ii) a lump sum N for $A \leq \tilde{C}$. We use N for the lump sum, as the binary payoff is usually the full bond notional.

[16] Due to the scarcity of historical data, $f_{\tilde{C}}(x)$ cannot be estimated empirically. Instead, it is the output of a catastrophe risk model. We will discuss the role of these models in the next section.

Moreover, the conditional expected loss (CEL) is[17]

$$\begin{aligned} CEL &= \mathbb{E}(\tilde{L}(\tilde{C}) \mid A < \tilde{C}) = \tfrac{\Pr(A<\tilde{C}\leq E)}{\Pr(A<\tilde{C})}\, \mathbb{E}(\tilde{L}(\tilde{C}) \mid A < \tilde{C} \leq E) + \tfrac{\Pr(E<\tilde{C})}{\Pr(A<\tilde{C})} \\ &= \int_A^\infty f_{\tilde{C}}(x)\, \tilde{L}(x)\, dx, \end{aligned} \tag{3}$$

and the expected loss, $EL = PFL \cdot CEL$, equals[18]

$$\begin{aligned} EL &= \mathbb{E}(\tilde{L}(\tilde{C})) = \Pr(A < \tilde{C})\, \mathbb{E}(\tilde{L}(\tilde{C}) \mid A < \tilde{C}) \\ &= \int_0^\infty f_{\tilde{C}}(x)\, \tilde{L}(x)\, dx. \end{aligned} \tag{4}$$

Instead of the pdf, industry practitioners often use the exceedance probability curve $EP(x)$. The two concepts are linked through the cumulative distribution function (cdf) of the underlying insured claims, $F_{\tilde{C}}(x)$, as follows: $EP(x) = \Pr(x < \tilde{C}) = 1 - F_{\tilde{C}}(x)$,[19] with

$$F_{\tilde{C}}(x) = \Pr(\tilde{C} \leq x) = \int_0^x f_{\tilde{C}}(u)\, du. \tag{5}$$

5.2 Catastrophe Risk Models

Historical data on the most extreme natural disasters is too limited to estimate the tail of the loss distribution (Poliquin and Lalonde 2012). Hence, risk assessment for cat bonds critically depends on catastrophe risk models, maintained by specialized firms, such as AIR, RMS, and CoreLogic. These model vendors use the latest scientific insights to create a bottom-up estimate of the insured loss distribution $f_{\tilde{C}}(x)$. A typical catastrophe risk model comprises three modules: (i) the hazard module, (ii) the vulnerability module, and (iii) the financial module (NAIC 2023).

The hazard module generates stochastic event sets of hypothetical disaster scenarios, characterized by key physical parameters, such as wind speeds or earthquake magnitudes. Within the vulnerability module, these scenarios are then overlaid on extensive databases containing all properties exposed to the respective catastrophe risk. Along with information about each structure's vulnerability, determined by factors like construction materials, age, and size, it becomes possible to estimate the physical damage sustained by buildings and household possessions in every scenario. Finally, the financial module converts the physical damage into financial losses based on property values and active insurance policy terms and aggregates them at the portfolio level.

In combination with the layer (attachment point A and exhaustion point E) and payoff profile (proportional vs. binary) of a risk transfer instrument, the resulting distribution of insured losses allows for a derivation of PFL, POE, CEL, and EL as described in Sect. 5.1. EL in particular has become a central factor in the pricing of cat bonds (Braun 2016). We will discuss this in further detail below.

A major question with regard to catastrophe risk models is their accuracy in estimating the risk metrics that serve as a basis for cat bond pricing. Lane (2022) provides a first empirical analysis in this regard, albeit on a relatively short time series of twenty years of data. He documents that

[17] The expression in the first line is already simplified, using the fact that $\mathbb{E}(\tilde{L}(\tilde{C}) \mid E \leq \tilde{C})$ is 1.

[18] Note from Equation (1) that $\Pr(\tilde{C} \leq A) \cdot \mathbb{E}(\tilde{L}(\tilde{C}) \mid \tilde{C} \leq A)$ is zero.

[19] It is straightforward to see that $EP(A) = PFL$ and $EP(E) = POE$.

actual losses on ILS were close to model-expected losses, if one adjusts for the fact that ILS cover the highest layers of the loss tower. This indicates that the major catastrophe models provide a solid representation the risks inherent in cat bonds.

5.3 Actuarial Pricing

Actuarial premium principles are the standard pricing method in insurance. They are used to derive the premium of a contract based on the expected loss plus a risk loading factor. There is a broad variety of premium principles. An overview can be found in Young (2014). For illustrative purposes, we briefly revisit the variance principle, which, in a cat bond context, formally looks as follows:

$$s_{cat} = \mathbb{E}(\tilde{L}) + \gamma \text{Var}(\tilde{L}), \tag{6}$$

where s_{cat} is the spread on the cat bond, $\text{Var}(\tilde{L})$ is the variance of the losses on the cat bond, and γ is a parameter that reflects the risk aversion of the investor.

Actuarial pricing approaches for cat bonds have been particularly suggested in the earlier years of the market (see, e.g., Major and Kreps 2002; Lane 2004). More recently, Stupfler and Yang (2018) suggest a new financial loss premium principle for cat bonds, which links risk aversion to current developments in the financial market.

5.4 Contingent Claims Pricing

Contingent claims pricing models for cat bonds are rooted in option pricing theory as constituted by Black and Scholes (1973) and Merton (1974). They rely on stochastic processes that describe the occurrence of insurance claims from natural disasters in a probability space $(\Omega, \mathcal{F}, \mathbb{Q})$. This implies a change of measure, from the real-world or physical pricing measure $\mathbb{P}$ to a risk-neutral (or equivalent martingale) measure $\mathbb{Q}$. The cat bond price then equals the expected discounted future cash flows (coupon and principal payments), where the expectation $\mathbb{E}^{\mathbb{Q}}$ is computed under the risk-neutral measure $\mathbb{Q}$ and discounting relies on the risk-free interest rate. For an equivalent martingale measure $\mathbb{Q}$ to exist, markets need to be arbitrage-free. For it to be unique, markets need to be complete, meaning that the payoff of any contingent claim can be replicated using traded securities (Cochrane 2009). Particularly the second condition causes problems in the context of cat bonds. The market for catastrophe risk instruments is generally incomplete, because their underlying are nontraded insurance losses, physical parameters, or industry indices (Cox and Pedersen 2000).

The contingent claims cat bond pricing literature has circumvented this problem in a variety of ways. Some authors simply derive a closed-form solution of their model under $\mathbb{Q}$ and assume the parameters to be identical as under $\mathbb{P}$ or do not further discuss the change of measure (see, e.g., Baryshnikov et al. 2001, Burnecki and Kukla 2003, Pérez-Fructuoso 2008, Wu and Chung 2010, Jarrow 2010, Burnecki et al. 2011, and Hainaut 2012). Others suggest that cat bond payoffs can be spanned with existing catastrophe risk instruments such as reinsurance or ILWs (see, e.g., Balbás et al. 1999, and Gatzert et al. 2019). Yet others tackle the nonuniqueness of the state price system by selecting a particular measure through a distortion operator or transform (see, e.g., Wang 2000, Wang 2004, Godin et al. 2019, and Tang and Yuan 2019).

Most studies, however, follow Merton (1976), who argued that jump risk is unsystematic and therefore fully diversifiable (see, e.g., Lee and Yu 2002, Vaugirard 2003a,b, 2004, Lee and Yu 2007, Ma and Ma 2013, Shao et al. 2017, and Chang et al. 2022). This is supported by the fact that cat bonds

have repeatedly been classified as zero-beta assets (Litzenberger et al. 1996; Canter et al. 1997) and empirical work found only minimal correlations of cat bond returns with other asset classes (see Sect. 4.4). However, according to capital market theory, zero-beta assets should return no more than the risk-free rate, a prediction that is irreconcilable with the large realized excess returns of cat bonds over the last two decades (Braun et al. 2019, 2022).[20]

Against the background of the arbitrariness and the disadvantages associated with some of the above solutions to the problem of incomplete catastrophe risk markets, a promising direction is to follow Breeden and Litzenberger (1978) and isolate the pricing kernel used by the market. In this vein, some authors derive risk-neutral probabilities from the prices of reinsurance contracts (see Geman 1999; Muermann 2008; Härdle and López Cabrera 2010). Once the market for cat bonds became more mature and liquid, however, the logic turned around. Haslip and Kaishev (2010) develop a framework that relies on the observed prices of cat bonds to price reinsurance in a market-consistent way.[21] Beer and Braun (2022) follow a similar logic. They utilize a closed-form solution of the reduced form model by Jarrow (2010) under the forward pricing measure to extract implied Poisson intensities under $\mathbb{Q}$ from secondary market cat bond quotes. They then estimate implied parameter surfaces for peak perils along the time to maturity and the modelled PFL. These surfaces enable a market-consistent valuation of instruments on the same underlying catastrophe risk, for which no prices are observed.

5.5 Utility-Based Pricing

In an incomplete market, where perfect replication is not possible, it is common to base pricing on individual preferences. The highest price an agent is willing to pay for a given risky asset is the one that makes her indifferent between entering into the transaction and not buying the asset, from her individual risk preferences' point of view (see, e.g., Hodges and Neuberger 1989). It is therefore referred to as the buyer's indifference price. To illustrate this approach in a simple setting, let us consider an uncertain universe characterized by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an agent, with a utility function u_B describing their preferences. This agent, with a random wealth W_0^B, has an opportunity to buy a given risky asset with random payout F. In a simple static framework, ignoring the impact of interest rates for the sake of simplicity and setting them equal to 0 without any loss of generality, the buyer's indifference price P^B for the asset F is determined by the following equation:

$$\mathbb{E}^{\mathbb{P}}\left[(u_B(W_0^B + F - P^B)\right] = \mathbb{E}^{\mathbb{P}}\left[u_B(W_0^B)\right]. \tag{7}$$

Intuitively, the agent's expected utility after buying the asset must equal his initial expected utility (without the asset). An important point to note is that the buyer's indifference price is not necessarily the price at which the transaction will take place, but an upper bound to the price the agent is ready to pay to buy the asset. Similarly, the indifference price of the seller is determined by his utility function u_S and initial wealth W_0^S:

$$\mathbb{E}^{\mathbb{P}}\left[u_S(W_0^S - F + P^S)\right] = \mathbb{E}^{\mathbb{P}}\left[u_S(W_0^S)\right]. \tag{8}$$

[20] Whether or not these high excess returns are risk premiums remains an unresolved cat bond pricing puzzle. They may, in fact, represent frictions, unless the natural catastrophe risks are large enough to directly affect marginal consumption of the representative investor (Bauer et al. 2013).

[21] At the time of writing of their paper, the secondary market for cat bonds was still much less developed than today. However, they correctly anticipated that the liquidity and the variety of regions and perils would further develop.

The indifference pricing rules in (7) and (8) are typically nonlinear and depend on the existing portfolio of buyer and seller.[22] Note that a transaction can only take place when $P^B \geq P^S$. The interval $[P^S; P^B]$ provides an acceptable price range for the transaction price rather than a single transaction price, leaving room for negotiation. The indifference pricing approach can used in a dynamic setting, incorporating hedging strategies, etc., as it is extremely flexible. Obtaining closed-form formulae for the indifference prices can be challenging depending on the class of utility functions, or risk measures used to characterize the preferences of the individuals.

The cat bond pricing literature is rich in works which acknowledge that the market is inherently incomplete and resort to utility-based pricing. Key articles in this strand include Cox and Pedersen (2000), Young (2004), Barrieu and El Karoui (2005), Zimbidis et al. (2007), Egami and Young (2008), Zhu (2011), Zheng (2015), Zhu (2017), Trottier et al. (2018), and Dieckmann (2019).

5.6 Statistical Approaches

We distinguish between two types of statistical approaches used in the catastrophe bond literature for pricing purposes. The first one is based on econometrics and the second one on machine learning. Details on how both methodologies are applied in the cat bond pricing context follow.

Econometric Pricing

Econometric pricing can be viewed as an extension of actuarial pricing, which seeks to explain the markup above the expected loss by means of characteristics of the transaction and the prevailing market environment. Due to the availability of primary market data through Lane Financial LLC and the Artemis.bm Deal Directory, most published studies focus on the explanation of the cat bond spread at issuance (see, e.g., Lei et al. 2008; Bodoff and Gan 2009; Papachristou 2011; Galeotti et al. 2013; Ciumaş and Coca 2015; Braun 2016; Mariani et al. 2018). As a typical representative of the class of econometric cat bond pricing models, we briefly revisit the approach by Braun (2016), which has become a benchmark and basis for subsequent work in this area (see, e.g., Makariou et al. 2021; Carayannopoulos et al. 2022).

Using a battery of cross-sectional OLS regressions on all cat bond tranches issued between 1997 and 2012, Braun (2016) shows that the following specification outperforms earlier econometric models in terms of in-sample and out-of-sample pricing accuracy:

$$s_{cat,i} = \beta_{EL} EL_i + \beta_{PEAK} PEAK_i + \beta_{SR} SR_i + \beta_{ROLX} ROLX_i + \beta_{IG} IG_i + \beta_{BBSPR} BBSPR_i + \epsilon. \quad (9)$$

This regression model[23] with error term ϵ predicts the spread $s_{cat,i}$ of cat bond i based on its expected loss EL_i, the current state of pricing in the reinsurance market as reflected by a rate on line index $ROLX_i$, and the spread on comparably rated corporate bonds $BBSPR_i$. Both $ROLX_i$ and $BBSPR_i$ are measured at the time issuance of cat bond i. Furthermore, the model contains three

[22] The dependency of the price on the original portfolio is a reasonable feature for the application to cat bonds, because it offers a theoretical explanation for the spread discount on nonpeak risk observed in empirical work (see, e.g., Braun 2016). Specifically, cat bond investors are prepared to accept a lower spread on an otherwise identical cat bond if it covers a non-US peril and thus offers diversification benefits in an ILS-only portfolio (further details follow in the next section).

[23] Note that the model does not contain an intercept. Braun (2016) argues that the spread should be zero if all of the factors are zero.

dummy variables, reflecting the identity of the sponsor (SR_i), investment grade versus high yield ratings (IG_i), and the distinction between peak and off-peak territories ($PEAK_i$). The dummy SR_i reveals the spread discount achieved by well-respected sponsors that serve the market repeatedly.[24] Moreover, the dummy $PEAK_i$ captures a markup in the spreads for cat bonds involving U.S. disaster risk (peak perils) relative to those covering risks from other parts of the world. The reason is the extremely U.S.-centered distribution of perils in the cat bond market (see Fig. 6), which makes nonpeak perils a sought-after diversifier for ILS-only investors.

In contrast to the analyses of cat bond issuance prices, econometric studies of yield spreads are rather limited. This is due to the challenges involved in securing secondary market data. Lane and Mahul (2008) were among the first to run a regression analysis on secondary market yield spreads. All subsequent insights are concentrated among the three major studies Gürtler et al. (2016), Herrmann and Hibbeln (2021), and Herrmann and Hibbeln (2023). The latter two are particularly innovative in that they are the first to provide empirical evidence of seasonality and liquidity effects in cat bond yields.

Machine Learning

Most recently a new area emerged in the cat bond pricing literature emerged, which employs machine learning models for spread prediction both in the primary market and in the secondary market. Makariou et al. (2021) introduce the random forest method to generate accurate spread predictions for new cat bond issues on both temporal and nontemporal bases and compared their results to benchmark models. In the absence of causal theory, they assess how spread predictors rank in terms of importance using two different methods, namely, permutation importance and minimal depth (see Ishwaran et al. 2010). The latter is random forest specific and has been applied by Makariou et al. (2021) in a financial context for the first time. The authors find their random forest prediction accuracy to be stable subject to multiple iterations of random subsampling and relatively robust to simultaneous missingness of more than one predictor. They also show how random forest can speed up decisions in the catastrophe bond industry both for would-be issuers and investors.

Götze et al. (2020) compare different machine learning methods for cat bond pricing at issuance, providing evidence that the random forest approach outperforms neural networks and linear regression, which they combine with variable selection via Lasso and Ridge penalizations. Götze et al. (2023) extend their earlier work using secondary market data. Given the current fast-paced developments in machine learning further research in this area will likely appear.

5.7 Historical Price Trends

Figure 12 shows the evolution of average issuance spreads and expected losses in the cat bond market between 1997 and 2022. The graph also includes the corresponding multiple, defined as spread divided by expected loss. The numbers underline that cat bonds are used to cover rare events with expected losses from one to three percent. Turning to the coupon spreads, two spikes stand out: one in 2006, following the extreme 2005 hurricane season with Katrina, Rita and Wilma, and another one in 2009, right in the middle of the financial crisis. This indicates a link of cat bond spreads to reinsurance premiums and corporate bond spreads, which has been empirically documented based on primary and secondary market data (see Braun 2016; Gürtler et al. 2016). Accordingly, the

[24] The sponsor variable first appeared in Braun (2016). Subsequent studies have either included it as a pricing factor (see, e.g., Chang et al. 2020b) or investigated the sponsor effect in more detail (see Götze and Gürtler 2020a; Chatoro et al. 2023).

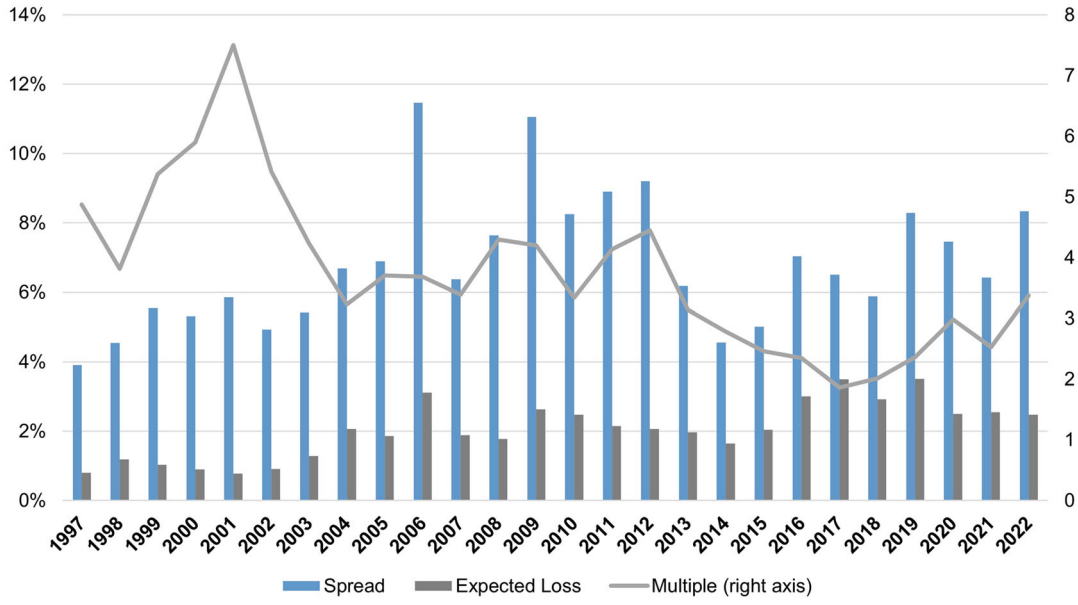


Fig. 12 Evolution of pricing in the cat bond market. (Source: Artemis.bm 2023)

reinsurance underwriting cycle with its hard and soft market periods has become a key factor in econometric cat bond pricing models (see Sect. 5.6).

The high multiples around the turn of the millennium may have been attributable to a novelty premium or behavioral aspects (Bantwal and Kunreuther 2000).[25] Subsequently, multiples have declined for almost two decades, potentially reflecting a learning process among investors associated with the maturation of the market (Braun and Weber 2017).[26] Following an extended period of benign natural disaster losses, multiples reached their lows around 2017. In the subsequent years, they began to rise again, driven by a high catastrophe activity and severe wildfire, hurricane and extreme weather losses (AON 2021b). For the better part of the last two decades, cat bond multiples hovered in the range of three to five, which is in line with reports about reinsurance pricing for higher layers of coverage (see, e.g., Froot and O'Connell 2008).

6 Concluding Remarks and Outlook

Today, cat bonds are an established instrument in the risk management toolkit of insurance and reinsurance companies where they coexist with traditional reinsurance contracts.[27] A key question

[25] Froot (2001) provides eight theoretical arguments for high multiples in catastrophe risk markets and deems capital market imperfections to be the most convincing one. Froot and Posner (2002) conclude that parameter uncertainty is unlikely to be the reason for the high multiples in the cat bond market.

[26] Empirical evidence for this logical argument is scarce. Some results are provided by Braun et al. (2013), who, based on a survey of European insurance firms, document that experience and expertise is a driver of investor demand for cat bonds.

[27] Determinants of the mix between traditional reinsurance and cat bonds brought forward by the theoretical literature are regulation, basis risk, information asymmetries, the covered risk layer, and the default risk of reinsurers (Nell and Richter 2000, 2004; Mutenga and Staikouras 2007; Cummins and Trainar 2009; Finken and Laux 2009; Klein and Wang 2009; Lakdawalla and Zanjani 2012; Gibson et al. 2014; Braun and Weber 2017; Trottier and Lai 2017; Subramanian

going forward is in which directions the market for cat bonds and ILS in general will continue to evolve. Three focus areas in this regard are: (i) new perils such as pandemic risk and cyber risk, (ii) sovereign risk transfer, and (iii) sustainability and climate change.

According to Swiss Re, pandemic risk could be an important driver of further ILS market growth (Artemis.bm 2021). The World Bank took an initial step in this regard by issuing the first pandemic bond for developing countries in 2017, which covered various infectious diseases (World Bank 2017). While some aspects of this transaction, especially the long time until the trigger threshold was reached, have been criticized (see, e.g., Zheng and Mamon 2023; Erikson and Johnson 2020), it ultimately paid out approximately $195 million after the onset of COVID-19 (World Bank 2020).

Another ILS market prospect is cyber risk. The recent growth in the cyber insurance market creates an urgent need for the transfer of tail risks. The academic literature has begun to cover the topic of cyber risk securitization (see, e.g., Xu and Zhang 2021; Liu et al. 2022; Braun et al. 2023). Braun et al. 2023, in particular, assess the viability of transferring cyber risk via ILS. Their findings suggest that cyber ILS can be beneficial for both cedents and investors, provided the cyber risk is adequately comprehended. Therefore, hurdles associated with cyber risk modelling must be addressed before a substantial cyber ILS market can develop. In practice, the speciality insurer Beazley sponsored the first cyber cat bond in January 2023. This bond will pay out if aggregate claims from a cyberattack on its customers exceed $300 million (Beazley 2023).

Promising perspectives for cat bonds also persist in the area of sovereign risk transfer. Particularly emerging market governments can increase the resilience of their public finances against natural catastrophes. To date, only a small proportion of cat bonds have been sponsored by governments, but more and more countries come forward to use this form of ex ante risk financing (Maran 2023). One example is Mexico, which to date, has recovered about $200 million of disaster losses from cat bonds of its sovereign risk pool FONDEN (Guy Carpenter 2023). Another example are the Philippines which have sponsored the first sovereign cat bond in Southeast Asia, comprising coverage against earthquake and tropical cyclone risk (World Bank 2019). Due to the forceful effects of climate change, know-how from the cat bond market is already diffusing into the general sovereign bond markets. Specifically, the government of Barbados has issued government bonds with parametric natural catastrophe clauses (Artemis.bm 2022a).

Finally, climate change and sustainability considerations impact the cat bond market in at least three ways. First, cat bonds are an important solution when it comes to the reduction of global disaster risk protection gaps, particularly in emerging markets. They are thus an inherently sustainable asset class. Second, global warming impacts the atmospheric perils (hurricanes, floods, wildfires etc.) that are being securitized in many cat bonds. It will therefore have a profound impact on the viability and pricing of the respective transactions (Michel-Kerjan and Morlaye 2008; Morana and Sbrana 2019). Third, cat bonds can support the transition to a low-carbon economy by investing exclusively in green collateral. The Italian insurer Generali has recently created the first ever green cat bond called Lion Re III. According to a 2021 press release, the latter comprises EUR 200 million worth of reinsurance coverage against European windstorms and Italian earthquakes over four years. The collateral of Lion Re III has been invested into green bonds of the European Bank for Reconstruction and Development (EBRD).

and Wang 2018; Faias and Guedes 2020; Chang et al. 2020a). Götze and Gürtler (2022) provide empirical evidence on the relationship between cat bonds and traditional reinsurance.

References

AON (2021a) ILS annual report 2021. Technical report
AON (2021b) Reinsurance market outlook. Technical report
AON (2022) ILS annual report 2022. Technical report
AON (2023) Reinsurance market dynamics. Technical report
Artemisbm (2016) Moody's updates catastrophe bond rating methodology
Artemisbm (2021) Cat bond market can grow to USD 50bn, pandemic risk & ESG Are Drivers: Swiss Re
Artemisbm (2022a) Barbados embeds parametric Nat Cat clause in sovereign bond issue.
Artemisbm (2022b) Cat bond lites are growing in number and importance
Artemisbm (2023) ARTEMIS catastrophe bond & ILS market dashboard
Balbás A, Longarela IR, Lucia JJ (1999) How financial theory applies to catastrophe-linked derivatives. An empirical test of several pricing models. J Risk Insur 66(4):551–581
Bantwal VJ, Kunreuther HC (2000) A CAT bond premium puzzle? J Psychol Financ Market 1(1):76–91
Barrieu P, El Karoui N (2005) Inf-convolution of risk measures and optimal risk transfer. Finance Stochast 9(2):269–298
Barrieu P, Bensusan H, Karoui NE, Hillairet C, Loisel S, Ravanelli C, Salhi Y (2012) Understanding, modelling and managing longevity risk: key issues and main challenges. Scand Actuar J 2012(3):203–231
Baryshnikov Y, Mayo A, Taylor DR (2001) Pricing of CAT bonds. Working paper
Bauer D, Phillips RD, Zanjani GH (2013) Financial pricing of insurance. In: Handbook of insurance, chap 22, pp 627–644
Beazley (2023) Beazley launches market's first cyber catastrophe bond.
Beer S, Braun A (2022) Market-consistent valuation of natural catastrophe risk. J Bank Finance 134:106350
Beer S, Braun A, Marugg A (2019) Pricing industry loss warranties in a Lévy–Frailty framework. Insurance Math Econ 89:171–181
Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81(3):637–654
Bockius H, Gatzert N (2022) The impact of counterparty risk on the basis risk of industry loss warranties and on (collateralized) reinsurance under (non-)linear dependence structures. J Risk Finance 23(3):245–263
Bodoff NM, Gan Y (2009) An analysis of the market price of cat bonds. Casualty actuarial society E-forum, Spring 2009
Braun A (2011) Pricing catastrophe swaps: A contingent claims approach. Insurance Math Econ 49(3):520–536
Braun A (2016) Pricing in the primary market for cat bonds: new empirical evidence. J Risk Insurance 83(4):811–847
Braun A (2023) Catastrophe bonds for sovereign disaster risk transfer in Asia: advantages, challenges and policy recommendations. Tech. rep., forthcoming
Braun A, Kousky C (2021) Catastrophe bonds. Wharton risk center primer
Braun A, Weber J (2017) Evolution or revolution? How solvency II will change the balance between reinsurance and ILS. J Insurance Regulat 36(4):1–26
Braun A, Mueller K, Schmeiser H (2013) What drives insurers' demand for cat bond investments? Evidence from a Pan-European survey. Geneva Papers Risk Insurance Issues Pract 38(3):580–611
Braun A, Ben Ammar S, Eling M (2019) Asset pricing and extreme event risk: common factors in ILS fund returns. J Bank Finance 102:59–78
Braun A, Herrmann M, Hibbeln M (2022) Common risk factors in the cross section of catastrophe bond returns. Working paper
Braun A, Eling M, Jaenicke C (2023) Cyber insurance-linked securities. ASTIN Bull, 1–22
Breeden DT, Litzenberger RH (1978) Prices of state-contingent claims implicit in option prices. J Bus 51(4):621–651
Bugler N, Maclean K, Nicenko V, Tedesco P (2021) Reinsurance sidecars: The next stage in the development of the longevity risk transfer market. North Am Actuar J 25(sup1):S25–S39
Burnecki K, Kukla G (2003) Pricing of zero-coupon and coupon CAT bonds. Appl Math 30(3):315–324
Burnecki K, Kukla G, Taylor D (2011) Pricing of catastrophe bonds. In: Cizek P, Härdle WK, Weron R (eds) Statistical tools for finance and insurance. Springer, Berlin, Heidelberg, pp 371–391
Canter MS, Cole JB, Sandor RL (1997) Insurance derivatives: a new asset class for the capital markets and a new hedging tool for the insurance industry. J Appl Corp Finance 10(3):69–81
Carayannopoulos P, Perez MF (2015) Diversification through catastrophe bonds: lessons from the subprime financial crisis. Geneva Papers Risk Insurance Issues Pract 40(1):1–28
Carayannopoulos P, Kanj O, Perez MF (2022) Pricing dynamics in the market for catastrophe bonds. Geneva Papers Risk Insurance Issues Pract 47(1):172–202
Chang C, Chang J, Yu MT, Zhao Y (2020a) Portfolio optimization in the catastrophe space. Eur Financ Manag 26(5):1414–1448
Chang C, Chang J, Yu MT (2022) Pricing hurricane bonds using a physically based option pricing approach. North Am Actuar J 26(1):27–42

Chang CW, Wang YJ, Yu MT (2020b) Catastrophe bond spread and hurricane arrival frequency. North Am J Econ Finance 54:100906
Chatoro M, Mitra S, Pantelous AA, Shao J (2023) Catastrophe bond pricing in the primary market: the issuer effect and pricing factors. Int Rev Financ Anal 85:102431
Ciumaş C, Coca RA (2015) Analysis of risk premium determinants on cat bonds. Proc Econ Finance 32:1487–1493
Clark SP, Dickson M, Neale FR (2016) Portfolio diversification effects of catastrophe bonds. SSRN Electron J. https://doi.org/10.2139/ssrn.2806432
Cochrane J (2009) Asset pricing: revised edition. Princeton University Press
Cox SH, Pedersen HW (2000) Catastrophe risk bonds. North Am Actuarl J 4(4):56–82
Cummins JD (2006) Should the government provide insurance for catastrophes? Federal Reserve Bank St Louis Rev 88(4):337–379
Cummins JD (2008) CAT bonds and other risk-linked securities: state of the market and recent developments. Risk Manag Insurance Rev 11(1):23–47,
Cummins JD, Trainar P (2009) Securitization, insurance, and reinsurance. J Risk Insurance 76(3):463–492
Cummins JD, Weiss MA (2009) Convergence of insurance and financial markets: hybrid and securitized risk-transfer solutions. J Risk Insurance 76(3):493–545
Cummins JD, Lalonde D, Phillips R (2004) The basis risk of catastrophic-loss index securities. J Financ Econ 71(1):77–111
Dieckmann S (2019) A consumption-based evaluation of the cat bond market. In: Advances in pacific basin business, economics and finance, pp 1–26
Doherty NA (1997) Innovations in managing catastrophe risk. J Risk Insurance 64(4):713–718
Doherty NA, Richter A (2002) Moral hazard, basis risk, and gap insurance. J Risk Insurance 69(1):9–24
Drobetz W, Schröder H, Tegtmeier L (2020) The role of catastrophe bonds in an international multi-asset portfolio: diversifier, hedge, or safe haven? Finance Res Lett 33:101198
Egami M, Young VR (2008) Indifference prices of structured catastrophe (CAT) bonds. Insurance Math Econ 42(2):771–778
Erikson SL, Johnson L (2020) Will financial innovation transform pandemic response? Lancet 20(5):529–530
Faias JA, Guedes J (2020) The diffusion of complex securities: the case of CAT bonds. Insurance Math Econ 90:46–57
Finken SN, Laux C (2009) Catastrophe bonds and reinsurance: the competitive effect of information-insensitive triggers. J Risk Insurance 76(3):579–605
Froot KA (2001) The market for catastrophe risk: a clinical examination. J Financ Econ 60(2-3):529–571
Froot KA, O'Connell PG (2008) On the pricing of intermediated risks: theory and application to catastrophe reinsurance. J Bank Finance 32(1):69–85
Froot KA, Posner SE (2002) The pricing of event risks with parameter uncertainty. Geneva Papers Risk Insurance Theory 27(2):153–165
Galeotti M, Gürtler M, Winkelvos C (2013) Accuracy of premium calculation models for CAT bonds-an empirical analysis. J Risk Insurance 80(2):401–421
Gatzert N, Kellner R (2011) The influence of non-linear dependencies on the basis risk of industry loss warranties. Insurance Math Econ 49(1):132–144
Gatzert N, Schmeiser H (2012) Industry loss warranties: contract features, pricing, and central demand factors. J Risk Finance 13(1):13–31
Gatzert N, Pokutta S, Vogl N (2019) Convergence of capital and insurance markets: consistent pricing of index-linked catastrophe loss instruments. J Risk Insurance 86(1):39–72
Geman H (1999) Learning about risk: some lessons from insurance. Rev Finance 2(2):113–124
Gibson R, Habib MA, Ziegler A (2014) Reinsurance or securitization: the case of natural catastrophe risk. J Math Econ 53:79–100
Godin F, Lai VS, Trottier DA (2019) A general class of distortion operators for pricing contingent claims with applications to CAT bonds. Scand Actuar J 2019(7):558–584
Götze T, Gürtler M (2020a) Hard markets, hard times: on the inefficiency of the CAT bond market. J Corp Finance 62:101553
Götze T, Gürtler M (2020b) Risk transfer and moral hazard: an examination on the market for insurance-linked securities. J Econ Behav Organization 180:758–777
Götze T, Gürtler M (2022) Risk transfer beyond reinsurance: the added value of CAT bonds. Geneva Papers Risk Insurance Issues Pract 47(1):125–171
Götze T, Gürtler M, Witowski E (2020) Improving CAT bond pricing models via machine learning. J Asset Manag 21(5):428–446
Götze T, Gürtler M, Witowski E (2023) Forecasting accuracy of machine learning and linear regression: evidence from the secondary CAT bond market. J Bus Econ 93:1629–1660
Gürtler M, Hibbeln M, Winkelvos C (2016) The impact of the financial crisis and natural catastrophes on CAT bonds. J Risk Insurance 83(3):579–612

Guy Carpenter (2023) Catastrophe bonds, disaster risk insurance facilities and natural disaster funds: protecting our planet and the public purse. https://www.guycarp.com/insights/2020/05/catastrophe-bonds-disaster-risk-insurance-facilities-and-natural-disaster-funds-protecting-our-planet-and-the-public-purse.html
Haffar A, Le Fur E (2022) Dependence structure of CAT bonds and portfolio diversification: a copula-GARCH approach. J Asset Manag 23(4):297–309
Hagedorn D, Heigl C, Mueller A, Seidler G (2009) Choice of triggers. In: Barrieu P, Albertini L (eds) The handbook of insurance-linked securities, 1st edn. Wiley, West Sussex, pp 37–48
Hainaut D (2012) Seasonality modeling for catastrophe bond pricing. Bull Français d'Actuariat 12(23):129–150
Härdle WK, López Cabrera B (2010) Calibrating CAT bonds for mexican earthquakes. J Risk Insurance 77(3):625–650
Harrington S, Niehaus G (1999) Basis risk with PCS catastrophe insurance derivative contracts. J Risk Insurance 66(1):49–82
Harrington SE, Niehaus G (2003) Capital, corporate income taxes, and catastrophe insurance. J Financ Intermediat 12(4):365–389
Haslip GG, Kaishev VK (2010) Pricing of reinsurance contracts in the presence of catastrophe bonds. ASTIN Bull 40(1):307–329
Herrmann M, Hibbeln M (2021) Seasonality in catastrophe bonds and market-implied catastrophe arrival frequencies. J Risk Insurance 88(3):785–818
Herrmann M, Hibbeln M (2023) Trading and liquidity in the catastrophe bond market. J Risk Insurance 90(2):283–328
Hodges S, Neuberger A (1989) Optimal replication of contingent claims under transaction costs. Rev Futur Mark 8:222–239
Ibragimov R, Jaffee D, Walden J (2009) Nondiversification traps in catastrophe insurance markets. Rev Financ Stud 22(3):959–993
Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS (2010) High-dimensional variable selection for survival data. J Am Stat Assoc 105(489):205–217
Jaeger L, Mueller S, Scherling S (2010) Insurance-linked securities: what drives their returns? J Alternat Invest 13(2):9–34
Jaffee DM, Russell T (1997) Catastrophe insurance, capital markets, and uninsurable risks. J Risk Insurance 64(2):205–230
Jarrow RA (2010) A simple robust model for Cat bond valuation. Finance Res Lett 7(2):72–79
Kish RJ (2016) Catastrophe (CAT) bonds: risk offsets with diversification and high returns. Financ Serv Rev 25(3):303–329
Klein RW, Wang S (2009) Catastrophe risk financing in the United States and the European Union: A comparative analysis of alternative regulatory approaches. J Risk Insurance 76(3):607–637
Lakdawalla D, Zanjani G (2012) Catastrophe bonds, reinsurance, and the optimal collateralization of risk transfer. J Risk Insurance 79(2):449–476
Lane M (2022) The ILS loss experience: natural catastrophe issues 2001–2020. Geneva Papers Risk Insurance Issues Pract 49:97–137
Lane MN (2004) Arbitrage algebra and the price of multi-peril ILS. J Risk Finance 5(2):45–51
Lane MN, Mahul O (2008) Catastrophe risk pricing: an empirical analysis. http://ideas.repec.org/p/wbk/wbrwps/4765.html
Lee JP, Yu MT (2002) Pricing default-risky CAT bonds with moral hazard and basis risk. J Risk Insurance 69(1):25–44
Lee JP, Yu MT (2007) Valuation of catastrophe reinsurance with catastrophe bonds. Insurance Math Econ 41(2):264–278
Lei DT, Wang JH, Tzeng LY (2008) Explaining the spread premiums on catastrophe bonds. Working paper
Litzenberger RH, Beaglehole DR, Reynolds CE (1996) Assessing catastrophe reinsurance-linked securities as a new asset class. J Portfolio Manag 23(5):76–86
Liu Z, Wei W, Wang L (2022) An extreme value theory-based catastrophe bond design for cyber risk management of power systems. IEEE Trans Smart Grid 13(2):1516–1528
Ma ZG, Ma CQ (2013) Pricing catastrophe risk bonds: a mixed approximation method. Insurance Math Econ 52(2):243–254
Major JA, Kreps RE (2002) Catastrophe risk pricing in the traditional market. In: Lane MN (ed) Alternative risk strategies, chap 10. Risk Books, London, UK
Makariou D, Barrieu P, Chen Y (2021) A random forest based approach for predicting spreads in the primary catastrophe bond market. Insurance Math Econ 101:140–162
Maran R (2023) Drivers of sovereign catastrophe bond issuance: an empirical analysis. SN Bus Econ 3(6):103
Mariani M, Amoruso P (2016) The effectiveness of catastrophe bonds in portfolio diversification. Int J Econ Financ Issues 6(4):1760–1767
Mariani M, Amoruso P, Didonato R, Caragnano A (2018) The drivers of Cat bond spread in the primary market. Int J Bus Manag 13(2):65
Merton RC (1974) On the pricing of corporate debt: the risk structure of interest rates. J Finance 29(2):449–470

Merton RC (1976) Option pricing when underlying stock returns are discontinuous. J Financ Econ 3(1-2):125–144
Michel-Kerjan E, Morlaye F (2008) Extreme events, global warming, and insurance-linked securities: how to trigger the "tipping point". Geneva Papers Risk Insurance Issues Pract 33(1):153–176
Morana C, Sbrana G (2019) Climate change implications for the catastrophe bonds market: an empirical analysis. Econ Model 81:274–294
Mouelhi C (2021) The relationship between Cat bond market and other financial asset markets: evidence from cointegration tests. Eur J Bus Manag Res 6(2):78–85
Muermann A (2008) Market price of insurance risk implied by catastrophe derivatives. North Am Actuar J 12(3):221–227
Mutenga S, Staikouras SK (2007) The theory of catastrophe risk financing: a look at the instruments that might transform the insurance industry. Geneva Papers Risk Insurance Issues Pract 32(2):222–245
NAIC (2023) Catastrophe models
Nell M, Richter A (2000) Catastrophe index-linked securities and reinsurance as substitutes. Working paper
Nell M, Richter A (2004) Improving risk allocation through indexed Cat bonds. Geneva Papers Risk Insurance Issues Pract 29(2):183–201
Niehaus G (2002) The allocation of catastrophe risk. J Bank Finance 26(2-3):585–596
Papachristou D (2011) Statistical analysis of the spreads of catastrophe bonds at the time of issue. ASTIN Bull 41(1):251–277
Pérez-Fructuoso MJ (2008) Modeling loss index triggers for CAT bonds: A continuous approach. Variance 2(2):253–265
Plenum Investments (2022) Market survey CAT bond UCITS funds. Technical report
Polacek A (2018) Catastrophe bonds: A primer and retrospective. Chicago Fed Lett 405:1–7
Poliquin B, Lalonde D (2012) Role of catastrophe risk modelling in ILS. In: Lane M (ed) Alternative (re-)insurance strategies, risk books, pp 195–218
Ramella M, Madeiros L (2007) Bermuda sidecars: supervising reinsurance companies in innovative global markets. Geneva Papers Risk Insurance Issues Pract 32(3):345–363
Shao J, Papaioannou AD, Pantelous AA (2017) Pricing and simulating catastrophe risk bonds in a Markov-dependent environment. Appl Math Comput 309:68–84
Smack L (2016) Catastrophe bonds—regulating a growing asset class. Risk Manag Insurance Rev 19(1):105–125
Sterge AJ, van der Stichele B (2016) Understanding Cat bonds. J Alternat Invest 19(1):27–40
Stupfler G, Yang F (2018) Analyzing and predicting Cat bond premiums: A financial loss premium principle and extreme value modeling. ASTIN Bull 48(1):375–411
Subramanian A, Wang J (2018) Reinsurance versus securitization of catastrophe risk. Insurance Math Econ 82:55–72
Swiss Re (2018) Insurance-linked securities market update. Industry report XXIX
Tang Q, Yuan Z (2019) Cat bond pricing under a product probability measure with POT risk characterization. ASTIN Bull 49(2):457–490
Towers Watson (2010) Catastrophe bonds evolve to address credit risk issues. Report (May)
Trottier DA, Lai VS (2017) Reinsurance or CAT bond? How to optimally combine both. J Fixed Income 27(2):65–87
Trottier DA, Lai VS, Charest AS (2018) CAT bond spreads via HARA utility and nonparametric tests. J Fixed Income 28(1):75
Trottier DA, Lai VS, Godin F (2019) A characterization of CAT bond performance indices. Finance Res Lett 28:431–437
Vaugirard VE (2003a) Pricing catastrophe bonds by an arbitrage approach. Q Rev Econ Finance 43(1):119–132
Vaugirard VE (2003b) Valuing catastrophe bonds by Monte Carlo simulations. Appl Math Finance 10(1):75–90
Vaugirard VE (2004) A canonical first passage time model to pricing nature-linked bonds. Econ Bull 7(2):1–7
Wang SS (2000) A class of distortion operators for pricing financial and insurance risks. J Risk Insurance 67(1):15–36
Wang SS (2004) CAT bond pricing using probability transforms. Geneva Papers Special Issue Insurance State Art Cat Bond Pric (278):19–29
World Bank (2017) World Bank launches first-ever pandemic bonds to support $500 million pandemic emergency financing facility
World Bank (2019) World Bank catastrophe bond transaction insures the Republic of Philippines against natural disaster-related losses up to USD 225 million
World Bank (2020) Fact sheet: Pandemic emergency financing facility
Wu YC, Chung SL (2010) Catastrophe risk management with counterparty risk using alternative instruments. Insurance Math Econ 47(2):234–245
Xu M, Zhang Y (2021) Data breach CAT bonds: modeling and pricing. North Am Actuar J 25(4):543–561
Young VR (2004) Pricing in an incomplete market with an affine term structure. Math Finance 14(3):359–381
Young VR (2014) Premium principles. In: Wiley StatsRef: statistics reference online
Zanjani G (2002) Pricing and capital allocation in catastrophe insurance. J Financ Econ 65(2):283–305

Zhang X, Tsai CCL (2018) The optimal write-down coefficients in a percentage for a catastrophe bond. North Am Actuar J 22(1):1–21
Zhao Y, Yu MT (2019) Measuring the liquidity impact on catastrophe bond spreads. Pacific Basin Finance J 56:197–210
Zheng M (2015) Heterogeneous expectations and speculative behavior in insurance-linked securities. Discrete Dynam Nature Soc 2015:574091
Zheng X, Mamon R (2023) Assessment of a pandemic emergency financing facility. Progress Disaster Sci 18:100281
Zhu W (2011) Ambiguity aversion and an intertemporal equilibrium model of catastrophe-linked securities pricing. Insurance Math Econ 49(1):38–46
Zhu W (2017) Wanting robustness in insurance: a model of catastrophe risk pricing and its empirical test. Insurance Math Econ 77:14–23
Zimbidis AA, Frangos NE, Pantelous AA (2007) Modeling earthquake risk via extreme value theory and pricing the respective catastrophe bonds. ASTIN Bull 37(1):163–183

Part III

New Challenges

Cyber Risk and Cyber Insurance

Martin Eling

1 Introduction: Motivation and Relevance

Most industry commentators agree that a large portion of the future risks come from the digital world (Centre for the Study of Financial Innovation 2019). Considering artificial intelligence, big data, cloud computing, the internet of things, and self-driving cars, the digitalization of the economy and society certainly creates a lot of new opportunities which will result in economic welfare in the next years. However, the increasing dependence on the IT sector and the integration of digital technologies into almost any sphere of life also create enormous cyber risks for individuals and organizations. Recently, increasing amounts of personal information have been breached, and substantial financial losses occurred because of malicious attacks and other events. For example, in August 2013 all registered accounts (three billion personal records) of Yahoo were breached (Fung 2017). In May 2019 First American had 885 million records exposed online including bank transactions and social security numbers (Liyanaarachchi 2020). The trend toward home office during and after the coronavirus lockdowns creates numerous new opportunities for criminal activities on the internet (Pranggono and Arabo 2021). Cases of ransomware attacks (e.g., Wannacry, NotPetya; see August et al. 2022) or other cyber extortions (see, e.g., Lallie et al. 2021) resulting in huge direct and indirect costs show that cyber risks vary substantially.

Many corporations protect themselves against cyber risks especially by investing in their security systems, meaning from a risk management point of view that the current focus of cyber risk management is on prevention. Cyber insurance as an alternative risk management strategy has seen a market growing rapidly compared to other insurance lines in the last few years (Romanosky et al. 2019). However, the cyber insurance market is still in its infancy and does not have the size some industry commentators expected a few years ago (Eling and Schnell 2016). One of the reasons for this underdevelopment might be that for both insurers and insureds, it is still unclear what the driving forces and mechanisms of cyber risk events are and what value by cyber insurance policies is created. Furthermore, asymmetric information and restrictive cover limits make it difficult to match supply and demand in the cyber insurance market (Biener et al. 2015).

M. Eling (✉)
University of St. Gallen, St. Gallen, Switzerland
e-mail: martin.eling@unisg.ch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_7

Over the last years, we have seen an increasing interest in cyber risk and insurance against cyber risk both in industry and academia. We document a steady growth of research on cyber risk and cyber risk insurance (see Fig. 1 in the Appendix), not only in computer science but also increasingly in business and economics (see Fig. 2 in the Appendix). There also have been top publications in finance, economics, and management journals focusing on market reactions to cyber risk events (Kamiya et al. 2021; Foerderer and Schuetz 2022; Florackis et al. 2023) and potential systemic risks arising from such events (August et al. 2022; Eisenbach et al. 2022; Crosignani et al. 2023). Yet, insurance is not a major component of this research.

In this chapter, we review both the academic literature and industry studies to better understand the insurability of cyber risks. For this, we analyze the status quo of the cyber insurance market and discuss current insurability limitations that hamper the growth of this new insurance market. We also seek alternative risk transfer solutions in the conventional and unconventional insurance market, especially the use of "cyber cat bonds" which have been discussed recently in industry working groups. The results of the analysis illustrate that specific properties of cyber risk explain the underdevelopment of the cyber insurance market. Cyber risks are characterized by (1) heavy tail distributions, (2) non-linear dependencies leading to potential accumulation risk, (3) model and parameter risk, and (4) asymmetric information. All these properties make cyber risk highly unattractive from an insurance company's point of view. There are also numerous demand side factors that limit market growth, among which are a lack of understanding and a partly fatalistic behavior with respect to these new risks.

In the following, we analyze five main research questions in the context of cyber risk, building upon recent results from academic research and industry studies. After providing a short definition, the first question we consider constitutes an important prerequisite of insurability: cyber risks must to some extent be measured in terms of frequency and severity (Berliner 1982; Biener et al. 2015). Second, we consider the few economic impact assessments of cyber risk scenarios to understand the economic impact of this risk class (building upon results presented by Eling et al. 2022). Third, we consider how far cyber risks differ from other types of risk, especially compared with other operational risks (Eling and Wirfs 2019). Fourth, the status quo of the traditional insurance market for cyber risk is considered (Xie et al. 2020) and based on the discussions in the previous questions we discuss the insurability of cyber risk in a fifth step (Eling and Schnell 2020). Here we also analyze how far an alternative risk transfer might take place involving capital market participants (Braun et al. 2023). We conclude in Sect. 3, also with an outlook on potential future research questions.

2 Five Main Research Questions for Cyber Risk

Cyber risks are operational risks to information and technology assets that have consequences for the confidentiality, availability, and integrity of information and information systems (Cebula and Young 2010). Notable in this definition are the references to the three main goals of information security (confidentiality, availability, and integrity) and to the class of operational risk class which is well known in the insurance industry and for which regulatory frameworks and insurance solutions already exist. A review of academic literature documents that cyber risk is an increasingly important research topic in many academic disciplines.[1] However, the attention it has received in the business

[1] While a number of papers have been published in the computer science domain, little research has been done in the fields of business, economics, and actuarial science. Research on cyber risk and cyber insurance was scarce until 2010, but since then it has grown exponentially, emphasizing the increasing practical and academic relevance of the topic (See Eling 2020).

and economics literature to date is limited (Eling 2020). In the field of risk management and insurance, leading journals (such as the Journal of Risk and Insurance or Insurance: Mathematics and Economics) have published a few articles on aspects of risk modeling, while in general interest journals (such as American Economic Review or the Journal of Finance) the topic has not yet been mentioned.[2] The literature in business and economics is, to a large extent, policy-oriented and a lot of work has been published by institutions such as Lloyd's or the Cambridge Risk Center to increase the awareness of the economic and societal relevance of the topic. Academic articles emphasize the lack of data and modeling challenges (Maillart and Sornette 2010; Biener et al. 2015), the complex and dependent risk structure (Hofmann and Ramaj 2011), and adverse selection and moral hazard (Gordon et al. 2003). Applied research is especially concerned with the potentially huge losses from worst-case scenarios such as the breakdown of critical infrastructure (World Economic Forum 2010; Lloyd's 2015a; Long Finance 2015). The literature thus highlights challenges in the risk management and insurability of cyber risks.

2.1 How Can We Measure Cyber Risk?

An important prerequisite for the measurement of cyber risk is the availability of data, which is rather scarce and problematic. One main reason for this is that institutions that have been compromised tend to not disclose incidences. However, from 2002 onwards U.S. entities that experienced security breaches are required to report the incidence to their customers and other parties, which has enhanced data availability over the recent years; another more recent example is the introduction of the General Data Protection Regulation (GDPR) in the European Union in 2017. However, even if historical data are available, the fast-changing environment for cyber risk might render these data only of limited use; one example is a comparison of the academic papers on the time properties of cyber risk, which come to quite inconclusive and ambiguous results. Maillart and Sornette (2010) document an explosive growth of data breaches until July 2006, followed by a more stationary phase; in contrast, Jung (2021) shows that there has been a structural time break in data breaches in 2014, with the majority of the large breaches happening after that time.

Some aggregated data on cyber risks collected in industry studies have already been discussed in the academic literature (Eling and Schnell 2016). Among these are the yearly cost of data breach studies (Ponemon Institute 2015, 2020), the cyber claims study (NetDiligence 2014), the yearly data breach investigations report (Verizon 2021), or the cyber crime reports by Symantec Corporation (2013, 2016) and McAfee (2014, 2020). Since these studies only report aggregated statistics, their use for modeling cyber risk is limited. The most useful data for modeling would be the claims an insurer actually faces, such as the data recorded by NetDiligence (2014). The three most widely used sources of raw data potentially suitable for actuarial modeling that have been used in the academic literature are:

- SAS OpRisk Global Data: Biener et al. (2015) and Eling and Wirfs (2019) extract cyber losses from a worldwide data set of publicly reported operational losses using text mining via a predefined search algorithm, resulting in a set of 1579 losses.
- Chronology of Data Breaches: Privacy Rights Clearinghouse (PRC) compiles publicly reported data breaches from 2005 until the present (Edwards et al. 2016; Eling and Jung 2018). It contains information about the type of data breach, involved organizations, and number of breached items per incidence. By the end of 2019, the data set contained 9012 entries in total.

[2] There is one recent exception (Kamiya et al. 2021) which we discuss below.

- Advisen Cyber Loss Data Set: This is a proprietary database on cyber risk-related events such as data and system breaches and security violations (used, e.g., by Romanosky 2016; or Aldasoro et al. 2022).

There are several other data sources on attack information, but here the link to actual loss information is rather loose and unclear. For example, the Internet Storm Center (ISC, by SANS Institute) collects the daily number of attacks per origin and target country where sensors register the criminal activity. Often used in IT research are also the so-called honeypots (e.g., by Honeynet), an international nonprofit organization. The honeypot (software or server) simulates a potential target, registers the attacks, and thereby measures malicious activity on the internet (Böhme and Kataria 2006). These data sources should also be closely studied because if correlations between attack activities or other measures of incidence and actual losses could be identified, this might constitute another fruitful area for measuring actual cyber losses and thus actuarial modeling.

Cyber-related loss data that can be used for actuarial modeling is thus still limited, but more and more data are becoming available. If we consider the existing analysis of the frequency and severity of cyber loss information, the standard actuarial toolbox shows that especially tools from extreme value theory are needed to describe cyber losses. For example, Eling and Wirfs (2019) use the Poisson or the negative binomial distribution for frequency and the peaks over threshold approach from extreme value theory for the severity of cyber losses. Wheatley et al. (2016) find breach sizes to be well modeled by an extremely heavy-tailed truncated Pareto distribution, with the tail exponent parameter decreasing linearly from 0.57 in 2007 to 0.37 in 2015. They also document a size effect where both the frequency and severity of breaches scale with organization size like $s^{0.6}$. Several authors (including Wheatley et al. (2016, 2021)) show that the distributions of losses due to cyber-attacks are heavy-tailed with exponents smaller than 1. This implies that the mean and variance of losses do not exist mathematically. In other words, the empirical mean and variance grow with the sample size and never converge. It is then a well-known mathematical result that diversification is not possible in such a situation, so we have fundamental impossibility for diversification—an aspect that has been made for example in the context of diversification of R&D in Pharma by Sornette (2002).

More recent papers try to improve the modeling of cyber risk frequency and severity going beyond standard actuarial models. For example, for the analysis of risk frequency, Bessy-Roland et al. (2021) propose a multivariate Hawkes model which captures the clustering and autocorrelation of times of cyber events depending on certain characteristics for predicting cyber risk frequency. They apply the model to the data set of PRC, which yields a better fit than traditional models. One example with respect to the severity is Farkas et al. (2021) who combined a Generalized Pareto modeling and a regression tree approach. The use of regression trees allows the identification of criteria for claim classification and evaluation. By applying such a method to the PRC data set, they find that some sectors (such as healthcare, education, and nonprofit organizations) have lighter tails than the others and that it is important to separate typical and extreme claims (denoted as cyber risk of daily life and extreme cyber risks by Eling and Wirfs (2019)).

Additionally, several recent papers study both the frequency and severity of cyber risk, especially in the context of network modeling. This is important because the cyber risk is a textbook example of systemic risk. For example, if all companies are using the same software or cloud, it is likely that one event will not only affect one company, but many companies at the same time; furthermore, in the digital world, the interconnectedness of information systems enables cyber incidents to spread quickly and wide (Zukis 2021; European Systemic Risk Board 2020). Fahrenwaldt et al. (2018) analyze the pricing of cyber insurance in the context of network models. Here, the spread of the cyber

infection is modeled by an interacting Markov chain, while the frequency and severity of claims are described by a marked point process. Xu and Hua (2019) propose a framework consisting of an epidemic model, loss function, and premium strategy. This model captures the general infection and recovery processes and also the related losses. Jevtić and Lanchier (2020) make a theoretical contribution by proposing a structural model of aggregate loss distribution for cyber risk under the assumption of a tree-based Local Area Network (LAN) topology. More specifically, they impose a contagion process on a network and introduce the topology of costs by focusing on the physical layer of the LAN topology.

Sun et al. (2020) propose a frequency-severity model for data breaches where the frequency is modeled by a hurdle Poisson model and the severity is captured by a nonparametric-generalized Pareto distribution model. This combination entails satisfactory performance for both in-sample and out-of-sample studies when incorporating a positive non-linear dependence between frequency and severity. Fang et al. (2021) tackle the problem of predicting data breach risk by applying a mixed D-vine copula to measure dependence structure. Also, this chapter considers the use of covariates when modeling loss frequency and the POT model when modeling loss severity. Zhang Wu et al. (2023) developed a new approach that combines deep learning with extreme value theory. They first model the time series of cyber risk via deep learning and then model the residuals through extreme value theory. Although this method provides both accurate point prediction and tail fitting, it suffers the problem of lacking explanation of dependence structure due to the "black box" property of deep learning. Zeller and Scherer (2022) propose the use of a marked point process by incorporating underlying risk factors in the modeling. Another area of research is the evolution of cyber risk over time. Wheatley et al. (2021) explore the frequency of data breaches from 2005 to 2017 with 1713 events from the PRC data set (combined with the former Open Security Foundation Dataloss Data Base). The results show a constant rate of breaches over the years. However, for the type of hacks, the frequency and severity are increasing.

Overall, this list of research papers emphasizes that cyber risk is a very active field of research especially in the mathematical/actuarial domain. Research is still more limited in the business and economics domain, for example when it comes to understanding the demand and supply side limitations in this new insurance market.

2.2 What Is the Economic Impact of Cyber Risk?

Estimating the costs and impact caused by cyber risk is difficult as there is high uncertainty and no accepted source of information. Some types of cybercrime might even cause no costs at all or cannot be quantified (such as spreading racism, mobbing, and trading illegal drugs). However, several industry contributions estimate overall costs, costs per incidence, and cost per record of a data breach. The costs of cyber risk are generally estimated to be in an area of 1–2% of global GDP. However, the estimates vary quite substantially which to some extent can be explained with the definition applied. For example, while Symantec Corporation (2013) considers only direct costs, McAfee (2014) also incorporates indirect costs, such as the reputational costs for the hacked company. The costs per data breach a hacked company faces show less variation and are estimated between 2.1 and 4.2 million US$.[3] The loss of each record (e.g., credit card number) causes costs from 217 to 956 US$. McAfee

[3] The costs per incidence estimates are 2.1 million US$ (KPMG 2013), 2.4 million US$ (Kaspersky Lab 2013), 2.6 million US$ (Kaspersky Lab 2019, for malware attacks), 3.8 million US$ (Ponemon Institute 2015), and 4.2 million US$ (Ponemon Institute 2020). NetDiligence (2019) compare small and large firms and estimate ranges from 0.18 million US$ (SMEs) to 5.6 million US$ (large firms). SMEs are defined as organizations with less than 2 billion US$ in

(2014, 2020) also provides estimates of the cyber risk costs for different countries, showing strong variations between different countries.[4]

Overall, it seems that the existing cost estimates are far from perfect. These numbers also have to be interpreted with caution as most of them have been estimated by security firms and consulting companies that might be incentives to present rather high-cost estimates. Anderson et al. (2013) discuss the methodological flaws of such estimates and suggest an alternative, which in the aggregate, however, also yields a number in the hundreds of billions of US dollars. Anderson et al. (2013), however, explicitly say that aggregation is not really meaningful. The authors also argue that the major part of cyber costs are indirect losses (loss of trust; not attributable to an individual victim) and defense costs (e.g., antivirus software, insurance) and not direct losses (e.g., theft of money).

From a micro perspective, cyber risk can have severe consequences for companies. In the field of business, the manifold detrimental effects of cyber risks have been analyzed, typically using event studies and scenario analyses. Several studies investigate the effects cyber risk incidences have on companies' stock prices. For example, Cavusoglu et al. (2004) show in an event study that a security breach negatively affects a company's stock price. They estimate the loss to be as high as 2.1% of the market volume or 1.65 billion US$ per security breach. A major part of the discount is explained by the reputational damage (Sinanaj and Muntermann 2013). Cavusoglu et al. (2004) also show that stock prices of information security providers increase on average in value by 1.36% or 1.06 billion US$ after the announcement of another company's security breach. On the contrary, Campbell et al. (2003) as well as Hovav and D'Arcy (2003) find only limited evidence that data breaches or denial of service attacks negatively influence the company's stock price. However, Campbell et al. (2003) provide evidence that a breach of confidential data has a larger negative effect on the stock price than nonclassified information would have; Hovav and D'Arcy (2003) show a negative price effect for companies with a business model that is heavily based on the internet. Thus, the markets seem to behave rationally as the discount is proportional to the expected loss associated with different data. Besides the sensitive customer data of a company, intellectual property such as copyrights, trademarks, industrial designs, and patents also might be at stake.

In the event study literature, Kamiya et al. (2021) developed a model where a firm has optimal exposure to cyber risk. With rational, fully informed agents and no hysteresis, a successful cyberattack should then not affect the target's reputation or its post-attack policies. However, Kamiya et al. (2021) documented that when a successful attack involves the loss of personal financial information, there is also a significant loss of shareholder wealth, which is much larger than the attack's out-of-pocket costs. The authors also show that an attack decreases a firm's risk appetite and decreases the risk-taking incentives of management. Successful cyberattacks adversely affect the stock price of firms in the target industry. The results imply that successful attacks with a loss of personal financial information provide adverse information about cyber risk to target firms, their stakeholders, and their competitors.

annual revenue, while large companies are defined as organizations with 2 billion US$ or more in annual revenue. The numbers give the impression that losses are increasing over time.

[4] The average total cost of a data breach (in million US$) is 9.05 in the United States, 5.40 in Canada, 4.89 in Germany, and 4.69 in Japan (Ponemon Institute 2020). It is not yet considered in the academic literature why such cross-country differences exist, which might come from differences in economic activities (e.g., more IT-affine industries) criminal activities, culture, or prevention. Another aspect is that the value of breached data can also vary by its type. One may expect that the value of a lost credit card number can be higher than that of an email address; biometric information could be more sensitive than other privacy information. It may be a fruitful avenue for future research to develop a global model for the costs of data breaches by type, also given that many cyber insurers primarily cover the costs of data breaches.

As an alternative to empirical studies and to raise awareness among policymakers, media, the public, and executives, a variety of scenarios have been proposed in the applied business literature and industry studies (Kelly et al. 2016; Risk Management Solutions Inc. 2016; Ruffle et al. 2014). These worst-case scenarios include various incidents that lead to a disruption of critical infrastructure and thus to economic losses. One often-discussed scenario is the monocultures in software and hardware markets that result in potential loss accumulation in the wake of a cyber incident (Eling and Schnell 2016).[5] The projected economic effects of the scenarios show some extreme variations, ranging from 0.2% to 2% of the GDP in the year of the event. However, as the studies lack common approaches, objectives, and thematic areas, the comparability of the scenarios and a comparative economic impact analysis are questionable. This leaves managers and policymakers with a vague idea of the potential frequency and severity of extreme cyber risks, resulting in ambiguous risk management strategies.

To overcome this limitation, Eling et al. (2022) implemented a dynamic inoperability input–output model that allows a consistent analysis and comparison of the economic impacts resulting from seven widely discussed cyber risk scenarios. Their model accounts for the frequently omitted qualitative context of the scenarios to be considered as part of the economic projection. Overall, the loss estimations remain in an insurable range from 0.7 to 35 billion US$. This chapter provides the first effort to develop a standardized evaluation framework that allows for a consistent assessment of cyber risk scenarios, thereby enabling comparability. The methodological foundation of the chapter is the dynamic input–output inoperability model which is based on the Leontief (1974) input–output model and has been developed and used in other contexts (Haimes and Jiang 2001). The inoperability input–output model allows analyzing how the inoperability of a given economic sector affects the overall economy and how the inoperability is amplified and spread through interdependencies (Panzieri and Setola 2008). Through its ability to model ripple effects, the inoperability input–output model has aroused the interest of both researchers and policy and is particularly promising in analyzing extreme cyber scenarios.

2.3 Are Cyber Risks Different from Other Types of Risk?

To understand the potential insurability limitations of cyber risk, one approach might be to compare cyber risk with other types of risks. Such comparisons can also be helpful in developing analogies that can identify potential future developments in the cyber insurance market. For example, analogies might be taken from the development of environmental liability law and respective insurance offerings in the United States in the 1980s where at the beginning data were also scarce. In this context, especially the comparison with other more extreme risks and operational risk seems relevant. Table 1 (which is adapted and expanded from Eling and Wirfs 2016) shows a comparison of terror risk, catastrophic (cat) risk, operational risk, and cyber risk.

Reflecting on the main results from the table, the following five differences between cyber risk and other types of risk become obvious: First, for cyber risk, there is not enough data to predict loss frequency and loss severity, which is a fundamental prerequisite of insurability. Without reliable indications on frequency and severity, only rough rules of thumb can be applied and consequently, only policies with very narrow cover limits can be offered.

Second, as indicated above, cyber risk is a textbook example of systemic risk—in contrast to other types of risks, including operational risk. This aspect points to another fundamental insurability prerequisite which is the independence of loss exposures. Risk pooling is only a meaningful concept

[5] Paté-Cornell et al. (2018) provide a deeper discussion of economic scenarios causing disruptions of critical infrastructure.

Table 1 Comparison of terror, catastrophic, operational, and cyber risks

#	Terror risk	Cat risk	Operational risk	Cyber risk
Data and measurability	Data exist; daily terror is measurable, e.g., political risk indices; extreme events are not well measurable (e.g., 9/11)	Data exist for decades; time trend with climate change; measurability given	Data exist, e.g., SAS OpRisk data; risk is also measurable (see Basel III, Solvency II)	Too little data exists especially on larger events Small losses rather measurable (hacker attacks and data breaches) Blackout scenario not measurable
Independence and maximum possible loss	Risk can be closely correlated; losses are geographically bounded	Can be closely correlated (e.g., hail damages); losses are geographically bounded	In general rather uncorrelated and losses are limited to one company	Small losses rather independent (hacker attack) Blackout scenario closely correlated (all stakeholders are affected), geographically not bounded
Time Horizon	Short- and long-tail	Typically short tail	Short- and long-tail	Can be both, but usually short tail
Risk nature	In general 1st party, but also 3rd party	1st party	Both	Both, but focus more on property than on liability
Frequency	Low	Low	Small losses are rather high-frequency business Extreme losses are rather low frequency	Small losses are rather high-frequency business (e.g., data fraud or theft) Blackout scenario is rather low frequency
Severity	High	High	Small losses are common (theft if not cyber) Blackout scenarios, which then would be high severity, conceivable (e.g., nuclear catastrophe in Chernobyl, product failure)	Small losses are common (hacker attack) Blackout scenario rather high severity (critical information infrastructure breakdown is Top 7 by severity, see WEF 2015)
Standardization	Definition yes; coverage no	Definition yes; coverage relatively standardized (e.g., cat bonds)	Definition yes; coverage no	Definition no; coverage no

Adapted from Eling and Wirfs (2016)

if losses are not highly correlated, but for cyber risk, it remains vague and unclear in how far accumulation risks exist. The insurer must consider the sum of all cover limits across policies as a potential worst-case loss scenario which substantially limits the risk capacity and the number of policies offered.

Third, for cyber risk, there is still a lack of standardized languages as well as an unavailability of standardized coverages. This creates a lot of uncertainty both on the demand and supply side of the potential insurance market. On the supply side, there is a lack of specialist product developers,

Table 2 Cyber and noncyber risk losses (in million US$)

					Quantiles			VaR	
Category	N	Mean	Std. dev.	Min.	25%	50%	75%	(95%)	Max.
Cyber risk	2794	39.88	365.55	0.1	0.49	1.83	9.41	129.90	16314.65
Noncyber risk	34,710	98.48	932.46	0.1	1.66	6.10	28.18	284.62	91066.96

underwriters, agents, and brokers who can offer cyber insurance. On the supply side, for most clients the potential risk and the set of adequate risk management responses available at affordable prices are unclear.

Fourth, another aspect that might cause legal disputes is that in many cases it cannot be verified how far a loss is due to a cyber event and who is liable for a claim. This creates another source of uncertainty that for other types of risk typically does not exist to a comparable extent.

Finally, another aspect that is also in contrast to other types of risk is that cyber risk can affect insurance companies in two fundamentally different ways, that is they do not only underwrite the risk, but are themselves exposed to it. Since insurance companies rely on their IT infrastructure, they are themselves highly vulnerable to cyber risk, an exposure that is treated in regulatory frameworks as part of operational risk (operational cyber risk). The underwriting of cyber risk also seems to be a business opportunity for insurance companies (underwriting cyber risk). This interaction between the underwriting and operational risk is another fundamental difference between cyber risk and other types of risk that need to be considered in the risk management and regulation of insurance companies. This connection creates a correlation and thus might further reduce the willingness of insurers to cover cyber risks on a large scale.

In the following, we compare cyber risk and other operational risks using an updated version of the SAS OpRisk data set that has been considered by Biener et al. (2015) and Eling and Wirfs (2019). Both papers extract cyber losses from the world's largest data set of publicly reported operational losses using text mining, resulting in a set of 996 and 1579 losses, respectively, in the period from January 1995 to September 2009 (March 2014). The same algorithm with new data until February 2021 yields 2794 cyber losses. In Table 2, we show a summary of the cyber risk sample and compare its characteristics with other operational risks. Mean and median cyber losses are significantly smaller than those for operational risk and in the range of what was previously published.[6] The maximal loss in our sample is 16 billion US$ compared to 91 billion US$ for noncyber risk.[7] Thus, both on average and in extreme cases, the loss amounts for cyber risk are much smaller than for other operational risks.[8]

Table 3 separates the cyber and noncyber risk loss data into several subcategories. The geographic separation (Panel A) shows that Northern American companies experience more than twice as many

[6] For cyber risks, Biener et al. (2015) report a mean (median) of US$ 40.53 (1.87) million, while Eling and Wirfs (2019) report a mean (median) of US$ 43.49 (1.53) million. For noncyber, the respective numbers are 99.65 (6.20) in Biener et al. (2015) and 98.92 (5.09) in Eling and Wirfs (2019). More recently, Aldasoro et al. (2022) report a mean loss of US$ 2.62 million for cyber risk, considering the Advisen data set.

[7] The largest cyber risk case occurred at the Bank of China in February 2005 when US$ 13,313.51 million were laundered through one of its branches, which was possible because the bank's internal money laundering controls were manipulated by employees. The largest noncyber risk case was Toyota with a $58B lawsuit filed by the US Justice Department for faulty pollution control systems 1999.

[8] Notable in this context is that cyber risk policies typically cover a maximum such as, e.g., US$ 50 million. Actual cover limits vary. If US$ 50 million is the limit, then 90% of the cases in our sample would be covered completely by the policy. Most policies in practice however have a cover limit of US$ 1 or five million (then 37% or 67% of all cases would be covered).

(53.8%) cyber risk incidents than European firms (23.2%). For loss severity, we find that Northern America has some of the lowest mean cyber risk and noncyber risk losses, whereas Europe has higher losses. Both these results may be due to North American firms being more capable of and willing to invest in risk-mitigating measures for extreme losses, which results from a longer tradition of recognizing and managing cyber risks as compared to Europe or Asia. The higher frequency of cyber events in North America might be explained by the fact that reporting requirements for cyber incidents in North America have been in place for more than a decade, while only recently have such requirements been implemented in other parts of the world (e.g., in the European Union in 2018 with the implementation of the new General Data Protection Regulation).

Panel B of Table 3 provides a separation into financial services and nonfinancial services industries. For cyber risk, 78.6% of all incidents occur in the financial services industry. This is not surprising since financial services firms, such as banks and insurance firms, store a significant amount of critical personal data. However, the average loss resulting from cyber risk for firms in nonfinancial services industries is about twice as high as for financial services firms. This finding might be explained by the fact that financial services firms have a higher awareness regarding critical data and better protect against severe losses from cyber risk. For noncyber risks, firms in the nonfinancial services industries face higher average losses than firms in the financial services sector; however, the difference is not as substantial as it is for cyber risk.

An important aspect of cyber risk is contagion, and thus our next separation of the data is between incidents affecting only one single firm and those affecting multiple firms (Panel C of Table 3). If just one firm is involved (83.2% of the cyber risk cases), the average loss per firm per case is more than twice as high as if more than one firm is involved. This result may appear counterintuitive; however, if more than one firm is affected, cyber-attacks are identified earlier and thus losses can be limited. Also, there may be economies of scale in solving the problems created by cyber incidents when multiple firms are involved (e.g., forensic investigation costs).

Panel D of Table 3 separates the sample based on firm size. With increasing size, the number of incidents increases, i.e., firms with more than 250 employees have more cyber losses. Interestingly, we observe a U-shaped pattern in the mean losses both for cyber and noncyber risk. It may be that smaller firms do not have the awareness and resources to protect against cyber risk, while large firms have diseconomies of scale due to complexity.

To more closely analyze the distributional characteristics of cyber risk compared to other operational risks, Biener et al. (2015) and Eling and Wirfs (2019) also estimate the loss severity distribution. The estimation is conducted by means of a spliced distribution, where a generalized Pareto distribution (GPD) models the tail. The results show that the distribution of cyber risk differs considerably from the distribution of other operational risks. For example, the distribution of the noncyber risk sample is much heavier tailed than that of the cyber risk sample, explaining in part the much higher maximal losses in these categories. This finding implies that when modeling operational risk, cyber risk needs to be considered separately.

Another aspect is the concept of "risk information deficit," namely, the fact that, when investigating some cyber-attacks, it soon becomes clear that before disaster struck, there were employees in the organization involved who were aware of dangerous conditions that had the potential to escalate to a critical level. However, for a variety of reasons, information about these risks did not reach decision-makers in time. So the organization continued to move toward catastrophe, effectively unaware of the possible threat—despite the fact that some of its employees could see the possibility of the disaster coming. Although this lack of internal risk communication is one of the principal causes of many disasters, weak internal feedback and organizational silence remain underestimated in many studies on disaster management, risk management, and organizational behavior (see Chernov and Sornette 2016; Chernov et al. 2022), including those on cyber risk.

Table 3 Cyber and noncyber risk losses by different categories (in million US$)

	Cyber risks				Noncyber risks			
	N	Share of cyber risk incidents	Mean	Median	*N*	Share of noncyber risk incidents	Mean	Median
All								
	2794	100%	39.88	1.83	34,710	100%	98.48	6.10
Panel A: Region of domicile								
Africa	85	3.04%	15.28	1.02	487	1.40%	85.63	2.89
Asia	579	20.72%	59.24	1.71	4332	12.48%	126.34	3.90
Europe	483	17.29%	42.71	2.07	4736	13.64%	122.59	5.32
North America	1504	53.83%	35.41	1.78	23,707	68.30%	89.69	6.75
Mid/South America	44	1.57%	17.04	2.46	662	1.91%	140.78	5.49
Other	99	3.54%	12.08	2.20	786	2.26%	37.14	4.39
Panel B: Industry								
Nonfinancial	658	23.55%	68.35	4.94	18,130	52.23%	106.33	8.63
Financial	2136	76.45%	31.11	1.35	16,580	47.76%	89.90	3.90
Panel C: Relation to losses in other firms								
One firm affected	2311	82.71%	39.79	1.75	24,587	70.84%	90.90	5.97
Multiple firms affected	483	17.28%	40.33	2.22	10,123	29.16%	116.87	6.53
Panel D: Company size by number of employees[a]								
Small	290	10.38%	13.91	1.14	2476	7.13%	29.10	2.18
Medium	559	20.07%	35.72	1.71	6223	17.93%	57.24	4.73
Large	1826	65.35%	45.23	1.83	20,373	58.69%	129.39	7.17

[a] Small firms have less than 250 employees, medium firms employ between 250 and 4999 people, and large firms have over 5000 employees. In Panel D, some events are dropped because of the lack of data on the number of employees

In conclusion, cyber risks are different from other types of risk, both theoretically as well as empirically. Cyber risks are smaller than other operational risks and less heavy tailed, but still need to be modeled using extreme value theory. There are various features that make cyber risk different from other types of risk, such as difficulties in measuring frequency, severity, and dependence structure as well as uncertainty in cyber risk management creating a lot of ambiguity in the supply and demand side of a potential cyber risk insurance market.

2.4 How Does the Market for Cyber Insurance Look Like Today?

The cyber insurance market is mainly a commercial insurance market, while cyber insurance solutions in the retail market are still at the very beginning. Regular policies for commercial property and liability insurance are available in most insurance markets worldwide. However, most property and liability policies only cover damage to physical assets such as production facilities and exclude cyber risk. In some older policies, the contract terms were also unclear on whether cyber events would be included or excluded (Lloyd's 2015b). Some customers might think that cyber incidents are covered, while the insurer assumes that they are not. Such an ambiguous situation might cause legal disputes and judge's decision discretion with legal costs increasing the insurers' financial risk (comparable to recent discussions for business interruption insurance and pandemics). Another problem mentioned above which might cause legal disputes is that in many cases it cannot be verified in how far a loss is due to a cyber event and who is liable for a claim. As a consequence of all these problems, insurers seek to state more explicit contract terms in two ways. Either the insurance companies adapt their policies by explicitly excluding cyber risks in traditional policies and providing dedicated "standalone cyber policies." Or alternatively, they could explicitly include cyber policies and adjust the premiums accordingly, leading to so-called "affirmative cyber policies" or "packed policies."

Over the last 10 years, a specialized market providing coverage for cyber risks has emerged especially in the United States. As yet, however, market coverage is small. Moreover, outside the United States, cyber insurance coverage is not well known and little used. In Europe, many corporations are not even aware that this type of insurance exists and only very few have purchased coverage. In the U.S. market, premium volumes of all types of cyber insurance have increased from $1.4 billion in 2015 to $6.5 billion in 2021. Also, there are 570 individual companies from 152 insurance groups that provide cyber insurance, with the top 20 insurers accounting for 83% of premiums written (National Association of Insurance Commissioners 2022). The insurers had a difficult year in 2020, as the direct loss plus defense and cost containment ratio for standalone policies rose sharply to 72.8% from an average of 40.5% in the previous 4 years; for packaged policies, the loss ratio increased to 58.6% from an average of 40.3% in the previous 4 years (Aon 2021).

Recent studies (Johansmeyer 2021; Gallagher 2021) indicate that the market has reached a possibly temporary plateau globally in 2020, especially under the COVID-19 situation, although there have just been cyber insurance policies exceeding $1 billion in coverage. As of 2020, the global cyber insurance market is estimated to have $5 billion in premiums (Johansmeyer 2021). Companies with coverage in the range of $100–199 million only represent 25% of the global market ($1.44 billion in premiums from around 500 companies), those with more than $200 million in insurance coverage account for 20% of the global market ($1.1 billion in premiums from 250 companies). Therefore, more than half of the market is made of policies with coverage lower than 100$ million (Johansmeyer 2021). Munich Re expects the global cyber insurance market to reach a value of approximately 20 billion US$ by the year 2025, which corresponds to a yearly growth rate of more than 25% (Munich Re 2021). Munich Re expects that especially small- and mid-sized companies will disproportionally suffer from cyber incidents and therefore drive the demand (Munich Re 2021). More recently, IMARC (2023) estimates

the global cyber insurance market at 9.8 billion US$ in 2022; they expect the market to reach 31.7 billion US$ by 2028, exhibiting a 22% growth rate from 2023 to 2028.

On the supply side, insurers are estimated to have ceded around half of their cyber premiums to reinsurers. The concentration for reinsurers is also staggering, with only four reinsurers taking up more than 60% of premiums and more than 75% of reinsurers having less than $100 million in premiums (Johansmeyer 2021). As for the price of cyber insurance, Chen (2021) shows that the average cost in the United States is $1485 per year. More specifically, the premiums range from $650 to $2357 for a policy with a liability of $one million and a deductible of $10 k for a typical company with $one million in revenue. Overall, the cyber insurance markets became challenging in 2020 with increasing loss ratios both in stand-alone and packaged lines. This challenge seems to last for the subsequent years and has led to a significant increase in the premiums (e.g., Howden 2022, documents a 185% price increase from January 2020 to April 2022). This circumstance, together with restrictive minimum standards and conservative cover limits keep agents exposed to cyber risks and hamper market growth that otherwise could be bigger. Higher prices might though attract new capital and together with improved risk management practices help to stabilize premiums and make cyber insurance economically more viable.

The U.S. market is much more developed than its European counterpart, partly because the United States have reporting requirements for cyberattacks in place for several years with relatively heavy fines for violations. The new regulations have considerably increased the awareness of cyber risk and increased the demand, especially for liability (third-party) cyber coverage. The U.S. market is thus mainly dominated by third-party coverage, whereas the few policies that already exist in Europe focus on first-party coverage. In 2018 reporting obligations for data breaches were also introduced in the European Union, which might be an important driver in the development of the European cyber insurance market.

Little information is available on the Asian insurance market for cyber risk, but it is likely that the market is even smaller than the European market. According to Iyer and Mishra (2021), compared to Europe and the United States, many Asian countries still lag in cyber security policy and strategies. This is astonishing given that Asia accounts for 25% of global cyberattacks, and Asia Pacific is home to 40% of the world's data centers and is expected to overtake the United States by 2024 (Iyer and Mishra 2021).

2.5 Are Cyber Risks Insurable?

In the Conventional Insurance Market

Biener et al. (2015) discuss cyber risk beyond the background of the insurability criteria by Berliner (1982). The first critical aspect which is also discussed above is that independence of losses is not given for cyber risk; risk pooling might thus not always work appropriately. Another problem is the unpredictability of the loss exposure, also because of a lack of data. Even if there is data available, it is questionable whether the historical data is a meaningful indicator for the future due to the dynamic nature of cyber risks and the risk of change. Another significant problem in cyber insurance is information asymmetry. Companies that have experienced a serious cyber-attack are more likely to buy insurance (Shackelford 2012), thus resulting in adverse selection. Insurance companies alleviate adverse selection by screening (e.g., audits up-front), self-selection (e.g., questionnaires in the underwriting process), and signaling (e.g., IT-compliance certificates). In addition, there is a moral hazard (i.e., the change of behavior after buying insurance). One example is the insured's lack of incentive to invest in self-protection measures following the purchase of insurance if full coverage is

offered. Insurers use instruments such as screening (e.g., audit) and risk sharing (e.g., deductibles and cover limits) to reduce moral hazard.

The cover limits are a particular essential problem for the development of the cyber insurance market. Existing policies tend to cover only small losses (ten million US$ to 50 million US$, depending on the product and provider; only for large companies more than 100 million US$ coverage is offered; Finkle 2015; Johansmeyer 2021) and contain several exclusions (e.g., self-inflicted losses, accessing unsecured websites, or terrorism). Potential extreme scenarios can thus not be well covered by existing insurance policies. Clients are, however, seeking coverage, especially for such extreme scenarios so supply and demand do not fit together very well. Another problematic aspect of cover limits is the policy complexity. Given the large number of exclusions, minimum standards for prevention, and the dynamic nature of cyber risk, there is uncertainty about what the cyber policy actually covers. Making this situation worse is that the policies in the market have no agreed-upon terminology, which makes the offerings very difficult to compare. In the end, it turns out that the policies available in the market do not really cover what the supply side is looking for: an efficient protection not for the small risks, but for the extreme cases that might make big losses.

There is no established model for cyber losses and only recently actuarial research on risk modelling has started. Eling and Wirfs (2019) show that the lognormal distribution frequently used for risk modeling is not the optimal choice when modeling cyber losses. As its probability mass in the tail is too low, the lognormal distribution would underestimate risk measures such as value at risk (VaR) or tail value at risk (TVaR). The best fit is identified for the peaks over threshold (POT) model, where the body consists of a lognormal and the tail of a generalized Pareto distribution (GPD). Eling and Wirfs (2019) calculate a Pareto index of around 0.6 depending on the region, industry, and time. These results not only question the applicability of risk measures like TVaR[9] but can, under certain circumstances (especially in underdeveloped markets), affect the optimal risk pool size. Building on the ideas of Ibragimov et al. (2009) on catastrophic risk, Eling and Schnell (2020) shed light on diversification properties when building portfolios of cyber risk. They illustrate the existence of a so-called "nondiversification trap" which indicates that for insurance companies it might be optimal not to diversify. More concretely, this means that insurance companies do not build up portfolios of cyber insurance policies because portfolio diversification does not decrease the risk, but rather increases the risk. These results might help to explain the current state of the cyber insurance market, which is far behind the expected market volume.

To investigate why the market for cyber insurance is underdeveloped, Eling and Schnell (2020) model demand and supply for cyber risk insurance and compare fair prices from the insurer's perspective with the willingness to pay of the potential clients. They calibrate their model with actual cyber loss data to document a potential market failure in the current market environment. The results show that the specific properties of cyber risk explain the underdevelopment of the cyber insurance market. As discussed above cyber risks are characterized by heavy tail distributions (Maillart and Sornette 2010; Edwards et al. 2016) and (tail) dependence (Herath and Herath 2011).[10] While for traditionally expected utility optimization, the risk always decreases and utility increases as a portfolio get more diversified (if the risks are not perfectly correlated and at least elliptically distributed),

[9] A distribution is heavy tailed (infinite second and higher moments) if the Pareto index $\alpha < 2$ and extremely heavy tailed (infinite first and higher moments) if $\alpha < 1$ (see Ibragimov and Prokhorov 2016). An infinite mean implies that sampling for risk measurement is not meaningful (see Chavez-Demoulin et al. 2006).

[10] While the heavy tails of cyber risks are well documented in the empirical literature, there is only little empirical research on the dependence structure. Many experts claim that certain cyber risks are dependent, e.g., because companies are using the same software or because some attacks spread like viruses.

research on catastrophic risks has shown that in the presence of heavy tails (Ibragimov et al. 2009) and strong (tail) dependence (Ibragimov and Prokhorov 2016) it can be optimal not to diversify at all. Even if there is a critical portfolio size where diversification becomes optimal, coordination problems might hinder the development of an insurance market.[11] Exactly these problems are shown in the analyses.

On top of these well-known coordination problems, cyber risks are prone to asymmetric information leading to high transaction costs. Moreover, the lack of loss experience and respective data result in parameter and model uncertainty. All these properties might lead to high insurance premiums far beyond the client's willingness to pay. Eling and Schnell (2020) discuss potential ways for insurers and policymakers to overcome these market failure problems. Among these are risk and data pooling or reinsurance to address dependence, mandatory insurance, or other governmental intervention (such public–private partnerships) to overcome the coordination problem, improving data quality to reduce modeling uncertainty, and IT security standards to reduce information asymmetry. Future researchers should go deeper into the root causes of the market failure. The leading root cause, the heavy-tailed distributions with exponents less than 1 so that mean and variance do not exist, makes the problem practically intractable. Future research needs to address this intrinsic heavy-tailed structure and discuss ways to develop functioning insurance markets. Experiences from other risk areas (e.g., operational risk) might provide avenues to overcome the coordination problem. One aspect might also be the use of alternative risk carriers which is the discussed in the next section.

In the Alternative Risk Transfer Market

Braun et al. (2023) discuss how far the alternative risk transfer market might provide an alternative for insuring cyber risk. The volume of insurance risk transferred to the capital markets has grown substantially over the last decade. In 2018, insurance-linked securities (ILS) funds surpassed 100 billion US$ of assets under management (Artemis 2018a). Alternative capital funneled through these funds covers cedents against the most extreme types of events, in some cases as a complement to and in others as a substitute for traditional reinsurance. Investors might be interested in bearing the risk since it is largely uncorrelated with the broader capital markets and thus enhances portfolio diversification (Cummins 2008).

So far, alternative capital has predominantly covered property-catastrophe insurance risk from hurricanes or earthquakes. As the global risk landscape continues to evolve, however, new events such as global pandemics and extreme cyber incidents are moving to the forefront of the industry. Extreme cyber risks, unlike those of everyday life, can be characterized by their magnitude across borders and industries (Eling and Wirfs 2019). The estimates of possible losses from an extreme cyber event by far exceed the 150 billion US$ losses on Nat Cat in 2019 (Dreyer et al. 2018; Munich Re 2020).

Drawing from the experience with Nat Cat risks, some industry experts consider other low-frequency and high-severity risks like extreme cyber risks as a natural target for ILS (Artemis 2018b). The rationale is that the capital markets should be incorporated into the transfer as they are the natural bearer of 35 of those extreme risks. The rising awareness of cyber threats is stimulating immense interest in finding alternative solutions for the insurability of those risks. First transactions already exist: Credit Suisse has issued an Operational Risk Cat Bond at 223 million US$ that includes cyber

[11] Ibragimov et al. (2009) distinguish a global market failure where diversification is never optimal no matter how large the portfolio and a local market failure where diversification is only optimal above a critical portfolio size. In the second situation coordination problems between several insures bundled in an insurance pool make it difficult to reach the critical portfolio size. As Ibragimov et al. (2009) point out that a global market failure can be caused by heavy-tailed risks (and strong tail dependency); combined with limited liability a local market failure might be observed.

risk (Artemis 2018c). Most recently, Hiscox backed a solution based on a parametric trigger to transfer cloud outage cyber risk (Artemis 2020a), and Hudson Structured together with Aon has issued a product that covers systemic cyber risks such as malware or distributed denial of service attacks (Artemis 2020b). However, given the distinct characteristics of extreme cyber risks (such as heavy tails, non-linear dependence, model and parameter uncertainty, and high costs due to asymmetric information), it remains unclear how far an alternative risk transfer will work on a broader scale for these risks.

Braun et al. (2023) show that a transfer of cyber risks is feasible for both investors and sponsors. In the first step, the paper investigates investors' preferences for insurance-linked securities with extreme cyber risk as the underlying default driver. With a choice-based conjoint analysis among fund managers in the ILS market, they estimate individual-level part-worth utilities and derive relative attribute importance and feasible product combinations from the preference structures. The results show that investors are willing to participate in the risk transfer of extreme cyber risks and that maturity, multiple, and model risk are the most important product attributes. The authors also compare investors' return expectations with insurers' cost of equity to match demand and supply in the risk transfer market and to model insurer portfolios composed of cyber risks and traditional P&C risks. The results of the choice-based conjoint analysis show that insecurity toward the underlying cyber risk modeling impedes cyber ILS market growth. Compared with other ILS characteristics (e.g., parametric vs. indemnity trigger and funded vs. unfunded instrument) especially a reduction in the modeling uncertainty would lead to a large utility increase and improve the willingness of investors to participate in a cyber risk transfer.

3 Conclusion and Future Research

Table 4 highlights the conclusions for the five main research questions studied in this chapter. The cyber insurance market is still in its early stages, but as market development continues, the risk pools might become larger and more data should become available. With the increasing market development, new competitors should be willing to enter the market which will increase insurance capacity, competition, and push prices down. More competition should also be helpful to develop a more uniform terminology and standardization of products. It might be important to establish standards on definitions, coverages, and pre-coverage risk assessment, all of which will help to reduce ambiguity in insuring cyber risk.

The government might also play an important role in fostering the development of a cyber insurance market. As there is a coordination problem, the government could try to push the cyber insurance market above the critical size needed for diversification. The government could incentivize the use of risk pools or reinsurance. This might also include carrying parts of the risk like done also for other extreme risks (e.g., terror), which can be used to increase the possible cover limits. This might be welfare enhancing since the currently used cover limits restrict the market to small risks and do not allow to realize welfare gains from insuring larger cyber risks above the cover limit. Alternatively, some form of sovereign guarantee might lower the probability of default and thus the value of the default option which decreases the clients' willingness to pay. Contrary to this argument, regulators should consider the usual capital discounts for diversification in case of cyber risk since diversification does not work well for cyber risks. We also note that intervention by the government will only be welfare enhancing if there is a local market failure; meaning that above a critical size, there is a market for cyber risk insurance (Ibragimov et al. 2009). It is not clear empirically under which cases an intervention is welfare enhancing or reducing, leading to some ambiguity with respect to potential policy recommendations.

Table 4 Five main conclusions

Question	Conclusion
Measurement of cyber risk	More and more data are becoming available, also for research (e.g., Advisen, PRC, SAS OpRisk) Risk measurement is difficult because of reporting biases and risk of change in underlying data
Economic impact of cyber risk	The economic impact is estimated to be 1–2% of GDP Major part of cyber costs is indirect losses (e.g., loss of trust) and defense costs (e.g., antivirus software) and not the direct losses (e.g., theft of money)
Comparisons with other types of risks	Cyber risk is a textbook example of systemic risk For cyber risk there is still a lack of standardized languages and no standardized coverages
Current cyber insurance market	Mainly a commercial insurance market, while cyber insurance solutions in the retail market are still at the very beginning Policies for commercial property and liability insurance are available in most insurance markets, but prices are relatively high and cover limits relatively narrow; also high minimum standards are necessary for coverage (cyber insurance a textbook example of holistic risk management, where insurance can only be offered together with prevention)
Insurability limitations	Four distinct characteristics of cyber risk contribute to the underdevelopment of the cyber insurance market: Heavy tail distributions, non-linear dependencies that lead to potential accumulation risk, model and parameter risk, and asymmetric information Additionally, several demand-side factors are also hindering the market's growth

Table 5 Open research questions clustered in three categories

Field	Insurance mathematics	Insurance management	Insurance economics
#1	Lack of data/improving data sets	Costs and detrimental effects of cyber risk	Economics of cyber security
#2	Frequency of cyber events	Privacy and data protection	Insurability of cyber risk
#3	Severity of cyber events	Optimal cyber risk management	Welfare effects of cyber security investments
#4	Risk of change/time dynamics of cyber risk	Innovative ways to manage cyber risks (e.g., ART)	Information-sharing and public and/or private cyber risk insurance pools
#5	Development of cyber risk scenarios	Cyber insurance product design	Systemic risk from cyber risks
#6	Dependence structure between cyber events/Systemic risk	Regulation of cyber risk	Empirically explain low demand

It is crucial to comprehend the nature and properties of cyber risk to provide cyber insurance and estimate risk capital accurately. Future research in this area could provide valuable insights into cyber risk management and the insurability of such risks. These findings would be beneficial to risk managers, policymakers, and regulators in developing effective policies for managing this new and dynamic risk category. Although we acknowledge the limitations of empirical research, we recognize the potential for future studies. Insurance research is broadly classified into three categories: insurance mathematics, management, and economics. Opportunities for future research in these three primary categories are discussed in detail (refer to Table 5).

Regarding future research, one prominent aspect is that the quality of data in the cyber risk domain is still limited. Although we can illustrate first empirical insights with different data sets (e.g., SAS, PRC, and Advisen), future research has to analyze the existing results on a broader empirical basis, especially to better understand under which circumstances the market failure for cyber risk is of local or of global nature. One particularly interesting element in this context is also the dynamic nature of cyber risks, which indicates that the statistical properties of cyber risk might change over time. As a first step, it might be helpful to analyze the time dimension in the existing data in more detail in order to identify potential time trends. For example, Maillart and Sornette (2010) illustrate an explosive growth in data breach frequency until 2006 and a stable rate thereafter; severity remained stable over time. These results might be interpreted in that the risk of change underlying cyber risk data might become less of a problem, but overall it remains unclear whether the past cyber risk data can be a reliable indicator for future cyber events.

In this context also information on individual cyber security at the firm level might yield useful insights into the interaction between prevention and losses which is seen as an important risk factor in the insurance industry but is not well explored in academic research. More detailed security information is needed to verify and better understand the impact of cyber security on loss frequency and severity. Also, the link between attack activity and actual losses can be closely studied to improve the understanding of cyber-attacks and their economic consequences. Furthermore, it is not yet considered in the academic literature what drives the large cross-country differences in cyber risks observed in the empirical literature.

With the changing nature of cyber risk, it is also unclear whether the current market failure will become less or more extreme; more data and loss experiences might help to better understand the tails of the cyber loss distribution, but with the increasing connectivity and globalization it might also be that future extreme events will be more severe and happen more often. Overall, the continuous evolution of the cyber risk landscape remains a fruitful and important area for future research not only in the IT domain but also in business, economics, and actuarial science. Our review shows cyber risk is now a very active field of research especially in the mathematical or actuarial domain, while there is a lack of research in the business and economics domain. Such research can be very useful to better understand the demand and supply side limitations in this new insurance market.

Looking at the mathematical and actuarial perspective, it is notable that cyber risk presents uncertain frequency and severity of potential losses. The dynamic nature of this risk category makes it challenging to collect sufficient data, posing a significant obstacle for insurance managers and researchers alike. Additionally, the absence of historical data renders it uncertain whether the limited data we currently possess can be a reliable indicator of future outcomes. Given the limited data and dynamic nature of cyber risk, there is a dearth of models that can accurately capture the characteristics of this emerging risk class. Developing robust cyber risk models would not only facilitate effective cyber risk management for companies but also foster the growth of a cyber insurance market. In the absence of sufficient data, predictive models and scenario analyses could offer valuable starting points for assessing premiums, risk capital, or reserves for cyber risks.

Many discussions center around two main issues: the nature of extreme cyber scenarios and the dependence structure between cyber losses. Some experts worry that cyber losses may have heavy tails and high correlation, as all companies may use the same software. If this is the case, it raises questions about the possibility of diversifying cyber risks or whether a nondiversification trap exists, as seen in other catastrophic risks. Currently, databases primarily focus on U.S. data breaches, leaving little information available for other countries or types of risk. However, the introduction of new data protection rules in the European Union, including mandatory notification requirements for certain cyber incidents, may improve this situation. These requirements will not only provide data for actuaries and researchers but may also encourage firms to invest in cyber self-protection and insurance.

The global cyber insurance market is expected to expand in the coming years, largely due to the new data protection regulations. However, from a business or management perspective, it is noteworthy that the cyber insurance industry is still very small and has not met the expectations of many experts. Misperceptions of cyber risk may be one explanation. A notable behavioral factor in this regard is the sense of latent fatalism expressed by many individuals when discussing cyber risk (such as "it won't happen to me" or "my data isn't interesting enough"). Understanding the drivers of this perception and increasing awareness through de-biasing research could help to boost demand for cyber risk insurance. Moreover, the effectiveness of communication strategies to improve perceptions and awareness of cyber risks is also a subject of inquiry. Additionally, it may be worth comparing risk perception and risk aversion in the realm of cyber risk with other types of insurance or other types of risks.

There is a need to close the gap between Enterprise Risk Management and Cyber Risk Management, as the former has seen significant advancements while the latter is still in its early stages in many organizations. Investing in people, processes, and systems is crucial, particularly in employing individuals who can facilitate communication between business managers and IT specialists. Given the current limitations in insurability, researching cyber insurance product design and innovative ways to transfer cyber risks, such as alternative risk transfer (ART) instruments like cat bonds, may be worthwhile. In the future, blockchain technology may give rise to new and innovative business models that can reduce and transfer cyber risks. However, a lack of data on the cost of cyber risks makes it challenging for companies to determine the right level of investment in self-protective measures and insurance. Additionally, the existence of "silent covers" in traditional policies raises important questions about how cyber risks are named and accounted for, particularly in older policies like business interruption or directors and officers (D&O). From a regulatory perspective, identifying and measuring these risks is an intriguing, unexplored topic. Silent cyber risk (well-described in Cartagena et al. 2020) can particularly be relevant when discussing insurability of cyber risk (ambiguity in the cyber insurance policy), underdevelopment of the cyber insurance market from the supply side, or difficulty in verification of how far a loss is due to a cyber event.

From an economics perspective, it is crucial to delve deeper into the relationship between insurance economics and cyber risk to gain a better understanding of the economic foundation of cyber security investments. What is the optimal combination of prevention, insurance, and other risk management practices that would be economically viable? What risk characteristics, such as correlation, determine whether insurance or self-protection is more efficient for managing cyber risks? In this context, a significant issue is whether minimum standards for cyber self-protection should be established to bolster the development of cyber insurance and improve economic welfare. This discussion is particularly relevant given the incentive problems inherent in cyber security investments. For individual users and companies, their IT systems' interconnectedness and dependencies imply that they remain vulnerable even if they invest considerably in self-protective measures against cyber risks. Consequently, investments in cyber security possess a "public good" attribute with positive externalities. The coordination issue arises because the value of cyber security investment by one firm is reliant on the investment made by all other companies in cyber security. These incentive problems are a direct result of cyberspace being a dense and complex network of information systems and users. What are the strategies that can be used to overcome this inherent incentive problem in cyber security investments? Information sharing and pooling of insurance in public or private partnerships are often discussed as potential solutions for government intervention.

Another crucial factor that needs to be considered in the macroeconomic context is whether widespread adoption of cyber risk insurance could lead to systemic risk, particularly if IT systems become increasingly interconnected. The degree of systemic risk may also depend on the type of cyber risk covered by insurance policies. Certain risks, such as viruses or phishing, may have a high

correlation across different companies, while other risks, such as insider attacks or hardware failures, may have a relatively low correlation. In the latter case, risk pooling is expected to be more effective, while in the former case, high correlation may lead to undesirable accumulation risks, especially in extreme scenarios. It is worth noting that insurance companies are impacted by cyber risk in two distinct ways: first, in underwriting new cyber insurance policies; and second, in the operational cyber risk arising from their own IT systems. In an industry that relies heavily on trust, reputation, and sensitive data, this operational risk also requires careful consideration.

The prevalence of empirical contributions and the lack of an established theoretical framework to analyze cyber risk are noteworthy. One might question whether a separate theoretical framework for cyber risk is necessary, or if it should be analyzed using classical models like any other new type of risk. However, the unique and complex characteristics of cyber risk, such as asymmetric information, potential correlations, and the "public good" nature, are illustrated in the above discussion. Therefore, it may be necessary to have separate papers that analyze cyber risk from a theoretical perspective.

Acknowledgments I am grateful for the comments and suggestions for improvement provided by two anonymous reviewers, Robert Bergien, Georges Dionne, Omid Ghavibazoo, and Dingchen Ning. Parts of this research were conducted with the support of the Korea Institute of Finance.

Appendix A Google Scholar Citations (Table 6)

Table 6 Google Scholar citations as of Feb. 1st, 2023

	"Cyber risk"	"Cyber insurance"
2000	10	6
2001	22	13
2002	19	9
2003	39	32
2004	42	23
2005	68	51
2006	74	54
2007	86	36
2008	94	41
2009	100	59
2010	135	67
2011	181	68
2012	210	82
2013	329	134
2014	509	189
2015	769	301
2016	1100	403
2017	1390	503
2018	1790	665
2019	2120	634
2020	2770	832
2021	3460	903
2022	3500	889

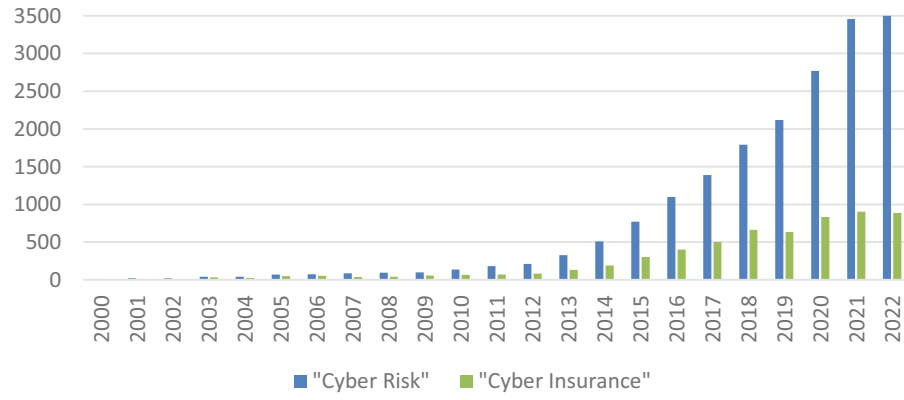


Fig 1 Google Scholar citations as of Feb. 1st, 2023

Appendix B Visualization Treemap for "Cyber Risk" and "Cyber Insurance" (Figs. 2 and 3)



Fig. 2 Visualization Treemap for 663 Web of Science hits on "cyber risk" (Feb 1st, 2023)

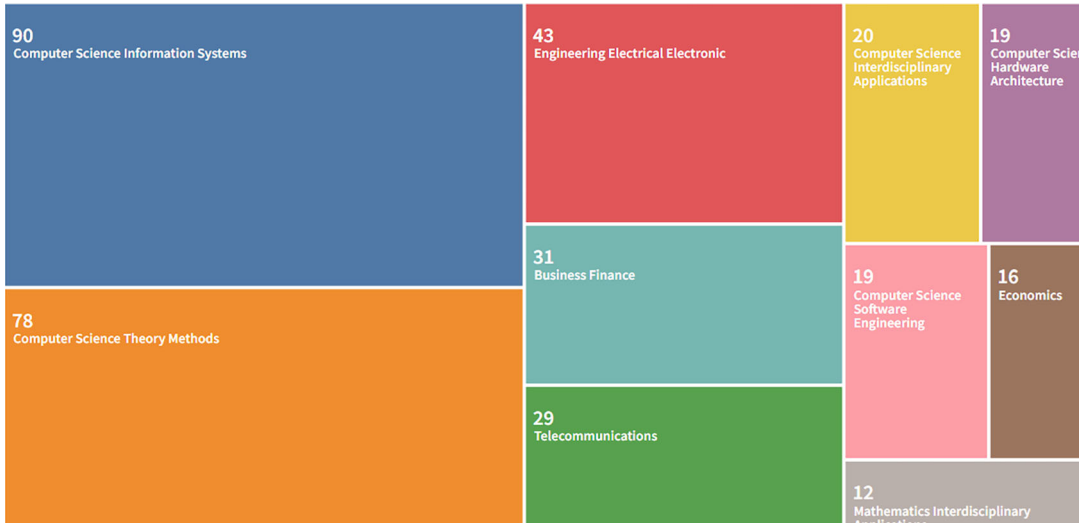


Fig. 3 Visualization Treemap for 226 Web of Science hits on "cyber insurance" (Feb 1st, 2023)

References

Aldasoro I, Gambacorta L, Giudici P, Leach T (2022) The drivers of cyber risk. J Financ Stab 60:100989. https://doi.org/10.1016/j.jfs.2022.100989

Anderson R, Barton C, Böhme R, Clayton R, van Eeten M, Levi M, Moore T, Savage S (2013) Measuring the cost of cybercrime. Paper presented at the 11th workshop on the economics of information, Berlin Brandenburg academy of science, Berlin, 25–26 June 2012

Aon (2021) US Cyber Market Update. http://thoughtleadership.aon.com/Documents/20210609-2021-cyber-market-update.pdf. Accessed 16 Mar 2023

Artemis (2018a) Continued ILS and alternative capital growth expected into 2019. https://www.artemis.bm/news/continued-ils-alternative-capital-growth-expected-into-2019/. Accessed 16 Mar 2023

Artemis (2018b) Munich Re calls for capital markets support on cyber accumulation risks. https://www.artemis.bm/news/munich-re-calls-for-capital-markets-support-on-cyber-accumulation-risks/. Accessed 16 Mar 2023

Artemis (2018c) Operational Re, Credit Suisse's op-risk cat bond, settles at CHF220m. https://www.artemis.bm/news/operational-re-credit-suisses-op-risk-cat-bond-settles-at-chf220m/. Accessed 16 Mar 2023

Artemis (2020a) Parametric cloud outage cyber risk transfer deal backed by Hiscox. https://www.artemis.bm/news/parametric-cloud-outage-cyber-risk-transfer-deal-backed-by-hiscox/. Accessed 16 Mar 2023

Artemis (2020b) Hudson structured and Aon team up for $70m cyber catastrophe product. https://www.artemis.bm/news/hudson-structured-aon-team-up-for-70m-cyber-catastrophe-product/. Accessed 16 Mar 2023

August T, Dao D, Niculescu MF (2022) Economics of ransomware: risk interdependence and large-scale attacks. Manag Sci 68(12):8979–9002. https://doi.org/10.1287/mnsc.2022.4300

Berliner B (1982) Limits of insurability of risks. Prentice Hall, Englewood Cliffs

Bessy-Roland Y, Boumezoued A, Hillairet C (2021) Multivariate Hawkes process for cyber insurance. Ann Actuar Sci 15(1):14–39. https://doi.org/10.1017/S1748499520000093

Biener C, Eling M, Wirfs JH (2015) Insurability of cyber risk: an empirical analysis. Geneva Pap R I-Iss P 40(1):131–158. https://doi.org/10.1057/gpp.2014.19

Böhme R, Kataria G (2006) Models and measures for correlation in cyber-insurance. Paper presented at workshop on the economics of information security (WEIS), University of Cambridge, June 2006

Braun A, Eling M, Jaenicke C (2023) What does it take to transfer cyber risk to the capital market? ASTIN Bull 53:684. https://doi.org/10.1017/asb.2023.22
Campbell K, Gordon LA, Loeb MP, Zhou L (2003) The economic cost of publicly announced information security breaches: empirical evidence from the stock market. J Comput Secur 11(3):431–448. https://doi.org/10.3233/JCS-2003-11308
Cartagena S, Gosrani V, Grewal J, Pikinska J (2020) Silent cyber assessment framework. Br Actuar J 25(2):1–19. https://doi.org/10.1017/S1357321720000021
Cavusoglu H, Mishra B, Raghunathan S (2004) The effect of internet security breach announcements on market value: capital market reactions for breached firms and internet security developers. Int J Electron Commer 9(1):70–104. https://doi.org/10.1080/10864415.2004.11044320
Cebula JJ, Young LR (2010) A taxonomy of operational cyber security risks. http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=9395. Accessed 29 Mar 2023
Centre for the Study of Financial Innovation (2019) Insurance banana skins 2019. https://www.pwc.com/gx/en/financial-services/assets/pdf/insurance-banana-skins-2019.pdf. Accessed 16 Mar 2023
Chavez-Demoulin V, Embrechts P, Nešlehová J (2006) Quantitative models for operational risk: extremes, dependence and aggregation. J Bank Financ 30(10):2635–2658. https://doi.org/10.1016/j.jbankfin.2005.11.008
Chen P (2021) Cyber insurance cost. https://advisorsmith.com/cyber-liability-insurance/cost/. Accessed 16 Mar 2023
Chernov D, Sornette D (2016) Man-made catastrophes and risk information concealment. Springer. https://www.springer.com/gp/book/9783319242996
Chernov D, Sornette D, Sansavini G, Ayoub A (2022) Don't tell the boss! How poor communication on risks within organizations causes major catastrophes. Springer. https://doi.org/10.1007/978-3-031-05206-4
Crosignani M, Macchiavelli M, Silva AF (2023) Pirates without borders: the propagation of cyberattacks through firms' supply chains. J Financ Econ 147(2):432–448. https://doi.org/10.1016/j.jfineco.2022.12.002
Cummins JD (2008) CAT bonds and other risk-linked securities: state of the market and recent developments. Risk Manag Insur Rev 11(1):23–47. https://doi.org/10.1111/j.1540-6296.2008.00127.x
Dreyer P, Jones T, Klima K, Oberholtzer J, Strong A, Welburn JW, Winkelman Z (2018) Estimating the global cost of cyber risk. Rand Corporation. https://doi.org/10.7249/RR2299
Edwards B, Hofmeyr S, Forrest S (2016) Hype and heavy tails: a closer look at data breaches. J Cybersecur 2(1):3–14. https://doi.org/10.1093/cybsec/tyw003
Eisenbach TM, Kovner A, Lee MJ (2022) Cyber risk and the US financial system: a pre-mortem analysis. J Financ Econ 145(3):802–826. https://doi.org/10.1016/j.jfineco.2021.10.007
Eling M (2020) Cyber risk research in business and actuarial science. Eur Actuar J 10(2):1–31. https://doi.org/10.1007/s13385-020-00250-1
Eling M, Jung K (2018) Copula approaches for modeling cross-sectional dependence of data breach losses. Insur Math Econ 82:167–180. https://doi.org/10.1016/j.insmatheco.2018.07.003
Eling M, Schnell W (2016) What do we know about cyber risk and cyber risk insurance? J Risk Financ 17(5):474–491. https://doi.org/10.1108/JRF-09-2016-0122
Eling M, Schnell W (2020) Extreme cyber risks and the nondiversification trap. Presented at 44th EGRIE seminar, University of St. Gallen, July 2017
Eling M, Wirfs J (2016) Cyber risk: too big to insure? Risk transfer options for a Mercurial risk class. http://www.ivw.unisg.ch/~/media/internet/content/dateien/instituteundcenters/ivw/studien/cyberrisk2016.pdf. Accessed 16 Mar 2023
Eling M, Wirfs J (2019) What are the actual costs of cyber risk events? Eur J Oper Res 272(3):1109–1119. https://doi.org/10.1016/j.ejor.2018.07.021
Eling M, Elvedi M, Falco G (2022) The economic impact of extreme cyber risk scenarios. N Am Actuar J 27:429. https://doi.org/10.1080/10920277.2022.2034507
European Systemic Risk Board (2020) Systemic cyber risk, February 2020. https://doi.org/10.2849/566567
Fahrenwaldt MA, Weber S, Weske K (2018) Pricing of cyber insurance contracts in a network model. ASTIN Bull 48(3):1175–1218. https://doi.org/10.1017/asb.2018.23
Fang Z, Xu M, Xu S, Hu T (2021) A framework for predicting data breach risk: leveraging dependence to cope with sparsity. IEEE Trans Inf Forensics Secur 16:2186–2201. https://doi.org/10.1109/TIFS.2021.3051804
Farkas S, Lopez O, Thomas M (2021) Cyber claim analysis using Generalized Pareto regression trees with applications to insurance. Insur Math Econ 98:92–105
Finkle J (2015) Ace offers $100 million cyber policies with added services, scrutiny. https://www.reuters.com/article/us-ace-ltd-cyberinsurance-idUSKCN0RO2LD20150924. Accesssed 16 Mar 2023
Florackis C, Louca C, Michaely R, Weber M (2023) Cybersecurity risk. Rev Financ Stud 36(1):351–407. https://doi.org/10.1093/rfs/hhac024

Foerderer J, Schuetz SW (2022) Data breach announcements and stock market reactions: a matter of timing? Manag Sci 68(10):7298–7322
Fung B (2017) Actually, every single Yahoo account got hacked in 2013. https://www.washingtonpost.com/news/the-switch/wp/2017/10/03/yahoos-2013-data-breach-affected-all-3-billion-accounts-tripling-its-previous-estimate/. Accessed 30 Mar 2023
Gallagher (2021) 2021 Cyber insurance market conditions report. https://www.ajg.com/us/news-and-insights/2021/jan/2021-cyber-insurance-market-report/. Accessed 16 Mar 2023
Gordon LA, Loeb MP, Sohail T (2003) A framework for using insurance for cyber-risk management. Commun ACM 44(9):70–75. https://doi.org/10.1145/383694.383709
Haimes YY, Jiang P (2001) Leontief-based model of risk in complex interconnected infrastructures. J Infrastruct Syst 7(1):1–12. https://doi.org/10.1061/(asce)1076-0342(2001)7:1(1)
Herath H, Herath T (2011) Copula-based actuarial model for pricing cyber-insurance policies. Insur Mark Co 2(1):7–20
Hofmann A, Ramaj H (2011) Interdependent risk networks: the threat of cyber attack. Int J Manag Decis Mak 11(5/6):312–323. https://doi.org/10.1504/IJMDM.2011.043406
Hovav A, D'Arcy J (2003) The impact of denial-of-service attack announcements on the market value of firms. Risk Manag Insur Rev 6(2):97–121. https://doi.org/10.1046/J.1098-1616.2003.026.x
Howden (2022) Cyber insurance: a hard reset 2.0. https://www.howdengroupholdings.com/themes/custom/howden_group_theme/assets/documents/howden-cyber-insurance-a-hard-reset-2.pdf
Ibragimov R, Prokhorov A (2016) Heavy tails and copulas: limits of diversification revisited. Econ Lett 149:102–107. https://doi.org/10.1016/j.econlet.2016.10.024
Ibragimov R, Jaffee D, Walden J (2009) Nondiversification traps in catastrophe insurance markets. Rev Financ Stud 22(3):959–993. https://doi.org/10.1093/rfs/hhn021
IMARC (2023) Cyber insurance market: global industry trends, share, size, growth, opportunity and forecast 2023–2028. https://www.imarcgroup.com/cyber-insurance- market. Accessed 30 Mar 2023
Iyer S, Mishra S (2021) Understanding risks within Asia Pacific's growing cyber insurance market. https://www.verisk.com/insurance/visualize/understanding-risks-within-asia-pacifics-growing-cyber-insurance-market/. Accessed 30 Mar 2023
Jevtić P, Lanchier N (2020) Dynamic structural percolation model of loss distribution for cyber risk of small and medium-sized enterprises for tree-based LAN topology. Insur Math Econ 91:209–223. https://doi.org/10.1016/j.insmatheco.2020.02.005
Johansmeyer T (2021) Cybersecurity Insurance has a big problem. https://hbr.org/2021/01/cybersecurity-insurance-has-a-big-problem. Accessed 16 Mar 2023
Jung K (2021) Extreme data breach losses: an alternative approach to estimating probable maximum loss for data breach risk. N Am Actuar J 25:580. https://doi.org/10.1007/s13385-023-00341-9
Kamiya S, Kang JK, Kim J, Milidonis A, Stulz RM (2021) Risk management, firm reputation, and the impact of successful cyberattacks on target firms. J Financ Econ 139(3):719–749. https://doi.org/10.1016/j.jfineco.2019.05.019
Kaspersky Lab (2013) Global corporate IT security risks: 2013. https://media.kaspersky.com/en/business-security/kaspersky_ global_it_security_risks_survey_report_eng_final.pdf. Accessed Mar 16 2023
Kaspersky Lab (2019) IT security economics in 2019. https://go.kaspersky.com/rs/802-IJN-240/images/GL_Kaspersky_Report-IT-Security-Economics_report_2019.pdf. Accessed Mar 30 2023
Kelly S, Leverett E, Oughton EJ, Copic J, Thacker S, Pant R, Pryor L, Kassara G, Evan T, Ruffle SJ, Tuveson M, Coburn AW, Ralph D, Halls JW (2016) Integrated infrastructure: cyber resiliency in society: mapping the consequences of an interconnected digital economy. https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/crs-integrated-infrastructure-cyber-resiliency-in-society.pdf. Accessed 16 Mar 2023
KPMG (2013) E-Crime – Computerkriminalität in der deutschen Wirtschaft mit Kennzahlen für Österreich und Schweiz. www.kpmg.com/CH/de/Library/Articles-Publications/Seiten/e-crimesurvey-2013.aspx. Accessed 16 Mar 2023
Lallie HS, Shepherd LA, Nurse JR, Erola A, Epiphaniou G, Maple C, Bellekens X (2021) Cyber security in the age of COVID-19: a timeline and analysis of cyber-crime and cyber-attacks during the pandemic. Comput Secur 105:102248
Leontief WW (1974) Structure of the world economy: outline of a simple input-output formulation. Am Econ Rev 64(6):823–834
Liyanaarachchi G (2020) Online privacy as an integral component of strategy: allaying customer fears and building loyalty. J Bus Strateg 41(5):47–56. https://doi.org/10.1108/jbs-09-2019-0183
Lloyd's (2015a) Business blackout: The insurance implications of a cyber attack on the US power grid. https://www.lloyds.com/news-and-risk-insight/risk-reports/library/society-and-security/business-blackout. Accessed 16 Mar 2023
Lloyd's (2015b) A quick guide to cyber risk. https://www.lloyds.com/news-and-insight/news-and-features/emerging-risk/emerging-risk-2015/a-quick-guide-to-cyber-risk. Accessed 16 Mar 2023

Long Finance (2015) Promoting UK cyber prosperity: public-private cyber-catastrophe reinsurance. http://wwwlongfinancenet/images/Promoting_UK_Cyber_Prosperity_28July2015pdf. Last accessed 16 Mar 2023
Maillart T, Sornette D (2010) Heavy-tailed distribution of cyber-risks. Eur Phys J 75:357–364. https://doi.org/10.1140/epjb/e2010-00120-8
McAfee (2014) Net losses: estimating the global cost of cybercrime. http://www.mcafee.com/mx/resources/reports/rp-economic-impact-cybercrime2.pdf. Accessed 16 Mar 2023
McAfee (2020) The hidden costs of cyber crime. https://www.mcafee.com/enterprise/en-us/assets/reports/rp-hidden-costs-of-cybercrime.pdf. Accessed Mar 16 2023
National Association of Incurance Commissioners (2022) Report on the cyber insurance market. https://content.naic.org/sites/default/files/cmte-c-cyber-supplement-report-2022-for-data-year-2021.pdf. Accessed 31 Mar 2023
NetDiligence (2014) 2014 – Cyber claims study. http://www.netdiligence.com/NetDiligence_2014CyberClaimsStudy.pdf. Accessed 16 Mar 2023
NetDiligence (2019) Cyber claims study. https://www.dfmurphy.com/Portals/0/2019_NetD_Claims_Study_Report_1.2.pdf?ver=2020-06-30-153925-017. Accessed 16 Mar 2023
Panzieri S, Setola R (2008) Failures propagation in critical interdependent infrastructures. Int J Model Identif Control 3(1):69–78. https://doi.org/10.1504/ijmic.2008.018186
Paté-Cornell ME, Kuypers M, Smith M, Keller P (2018) Cyber risk management for critical infrastructure: a risk analysis model and three case studies. Risk Anal 38(2):226–241. https://doi.org/10.1111/risa.12844
Ponemon Institute (2015) 2015 Cost of cyber crime study – global. http://www8.hp.com/us/en/software-solutions/ponemon-cyber-security-report. Accessed 16 Mar 2023
Ponemon Institute (2020) Cost of a data breach report 2020. https://www.ibm.com/security/digital-assets/cost-data-breach-report/1Cost%20of%20a%20Data%20Breach%20Report%202020.pdf. Accessed 30 Mar 2023
Pranggono B, Arabo A (2021) COVID-19 pandemic cybersecurity issues. Internet Technol Lett 4(2):e247. https://doi.org/10.1002/itl2.247
Munich Re (2020) Tropical cyclones causing billions in losses dominate natcat picture of 2019. https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2020/causing-billions-in-losses-dominate-nat-cat-picture-2019.html. Accessed Mar 16 2023
Munich Re (2021) Cyber insurance: risks and trends 2021. https://www.munichre.com/topics-online/en/digitalisation/cyber/cyber-insurance-risks-and-trends-2021.html. Accessed 16 Mar 2023
Risk Management Solutions Inc. (2016) Managing cyber insurance accumulation risk. https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/risk/downloads/crs-rms-managing-cyber-insurance-accumulation-risk.pdf. Accessed 16 Mar 2023
Romanosky S (2016) Examining the costs and causes of cyber incidents. J Cybersecur 2(2):121–135. https://doi.org/10.1093/cybsec/tyw001
Romanosky S, Ablon L, Kuehn A, Jones T (2019) Content analysis of cyber insurance policies: how do carriers price cyber risk? J Cybersecur 5(1):tyz002. https://doi.org/10.1093/cybsec/tyz002
Ruffle SJ, Bowman G, Caccioli F, Coburn AW, Kelly S, Leslie B, Ralph D (2014) Sybil logic bomb cyber catastrophe. https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/crs-sybil-logic-bomb-cyber-catastrophe-stress-test.pdf. Accessed 16 Mar 2023
Shackelford SJ (2012) Should your firm invest in cyber risk insurance. Bus Horiz 55:349–356. https://doi.org/10.1016/j.bushor.2012.02.004
Sinanaj G, Muntermann J (2013) Assessing corporate reputational damage of data breaches: an empirical analysis. Presented at 26th bled econference, bled (Slovenia), 9–13 June 2013
Sornette D (2002) Economies of scale in innovations with block-busters. Quant Financ 2(3):224–227
Sun H, Xu M, Zhao P (2020) Modeling malicious hacking data breach risks. N Am Actuar J 25(2):1–19. https://doi.org/10.1080/10920277.2020.1752255
Symantec Corporation (2013) 2013 Norton report. http://www.symantec.com/content/de/de/about/downloads/2013_Norton_Report_Deck.pdf. https://www.nortonlifelock.com/us/en/newsroom/press-kits/norton-report-2013/
Symantec Corporation (2016) Internet security threat report. https://wwwsymanteccom/security-center/threat-report. https://static.abc.es/gestordocumental/uploads/internacional/symantec.pdf
Verizon (2021) Data breach investigations report. https://www.verizon.com/business/resources/reports/dbir/. Accessed 16 Mar 2023
Wheatley S, Maillart T, Sornette D (2016) The extreme risk of personal data breaches and the erosion of privacy. Eur Phys J B 89(1):1–12. https://doi.org/10.1140/epjb/e2015-60754-4
Wheatley S, Hofmann A, Sornette D (2021) Addressing insurance of data breach cyber risks in the catastrophe framework. Geneva Pap R I-Iss P 46(1):53–78. https://doi.org/10.1057/s41288-020-00163-w
World Economic Forum (2010) Global risk report 2010 – a global risk network report. http://www.weforum.org/pdf/globalrisk/globalrisks2010.pdf. Accessed 16 Mar 2023

World Economic Forum (2015) Global risks report 2015 – 10th edition. http://www3.weforum.org/docs/WEF_Global_Risks_2015_Report15.pdf. Accessed 16 Mar 2023

Xie X, Lee C, Eling M (2020) Cyber insurance offering and performance: an analysis of the US cyber insurance market. Geneva Pap R I-Iss P 45(4):690–736. https://doi.org/10.1057/s41288-020-00176-5

Xu M, Hua L (2019) Cybersecurity insurance: modeling and pricing. N Am Actuar J 23(2):220–249. https://doi.org/10.1080/10920277.2019.1566076

Zeller G, Scherer MA (2022) A comprehensive model for cyber risk based on marked point processes and its application to insurance. Eur Actuar J 12(1):33–85. https://doi.org/10.1007/s13385-021-00290-1

Zukis B (2021) Some vital lessons in how systemic risk is changing cybersecurity. https://www.forbes.com/sites/bobzukis/2021/09/02/some-vital-lessons-in-how-systemic-risk-is-changing-cybersecurity/?sh=56a978e31491. Accessed 30 Mar 2023

Zhang Wu M, Luo J, Fang X, Xu M, Zhao P (2023) Modeling multivariate cyber risks: deep learning dating extreme value theory. J Appl Stat 50(3):610–630.

Digital Insurance and InsurTech

Alexander Braun and Niklas Haeusle

1 Introduction

As many other industries, the insurance sector is facing digital transformation. In this context, insurance companies are experiencing a seismic shift because new technologies impact consumer behavior and expectations, transform the value chain, challenge long-established business models, redefine the ways in which insurance products and services are delivered, and change the market structure (Bauer et al. 2021; Eling et al. 2022; Fleischer et al. 2022; Haeusle 2023). Consequently, traditional insurers find themselves in a race to adapt, innovate, and thrive in this new era of digital insurance.

This chapter provides a comprehensive overview of the repercussions of digital transformation within the insurance industry. We define *digital transformation* as the process of integrating digital technologies into all areas of a business, fundamentally changing how the business operates and rethinking how it delivers value to its customers. Digital transformation is therefore not just about updating technology; it is a cultural shift that requires organizations to continuously challenge the status quo, experiment, and embrace new ways of working. Digital transformation can involve a variety of components, from updating internal systems and processes to adopting new business models and enhancing the customer experience using digital tools and channels. The overarching goal is to become more agile, efficient, and customer-centric in a rapidly evolving digital environment.

We begin our discussion with legacy issues faced by insurers that pose major obstacles for the transformation of the value chain and the exploration of new business models. We also discuss the rewards that can be reaped by digital transformation leaders. In the second section, we consider how the insurance value chain is affected by the digital transformation, leading to an enhanced customer

Alexander Braun is an Associate Professor of Insurance and Capital Markets and Director of the Institute of Insurance Economics at the University of St. Gallen, Tannenstrasse 19, CH-9000 St. Gallen, Switzerland. Niklas Haeusle is an Assistant Professor of Insurance at Leipzig University.

A. Braun (✉)
University of St. Gallen, St. Gallen, Switzerland
e-mail: alexander.braun@unisg.ch

N. Haeusle
Leipzig University, Leipzig, Germany

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_8

experience and operational excellence through automation and efficiency. The third section focuses on innovative business models that transcend the boundaries of the classical insurance logic and create value in novel ways. Finally, in the last section, we will explore the role of InsurTech start-ups in accelerating digital transformation, highlighting their unique value propositions and the competitive dynamics between them and established insurers.

The goal of this chapter is twofold. First, by providing a comprehensive overview of the digital transformation in the insurance industry, we aim to equip readers with a deeper understanding of the forces shaping the future of insurance and the strategies required for success in the digital age. Second, we wish to inspire ideas for academic work, acknowledging that both applied and theoretical research on the digital transformation in the insurance industry are still very scarce compared to other domains within our field.[1]

2 Challenges and Rewards of the Digital Transformation

2.1 Challenges

Across industries, executives today face the challenge of how to successfully navigate the digital transformation (Matt et al. 2016). Insurance is no exception. Owing to a range of legacy issues, the transformation of the value chain poses a substantial challenge to the insurance industry (BCG 2018). First, there is a distribution channel legacy. In many established markets, insurance companies run a large agent force that serves customers in a classical brick-and-mortar distribution logic. These agents, who have significant power in insurance organizations, may perceive the advent of new digital direct channels as a threat and thus oppose transformation efforts.

Second, there is a severe issue with legacy IT systems and infrastructure. Instead of radically switching to a state-of-the-art IT landscape after important technological breakthroughs, many insurance companies have taken care of the front ends, such as customer portals, but just patched up the hardware and software in their core systems over the last two decades. Industry reports, e.g., indicate that in the mid-2010s, the average age of core IT systems in the insurance industry was already 13 years (BCG 2016).[2] This has created a severe technical debt in the form of slow, complex, nonstandardized, and inflexible back ends that are now very difficult to substitute.

Third, product legacy impedes the digital transformation of the value chain, because many policy designs are too cumbersome and therefore unsuited for the digital age. Insurers therefore need to strive for a significant simplification of their products and develop new designs that integrate the possibilities of modern technologies, such as connected devices (Internet of Things). They also need to simplify policy language and coverage options and rethink the customer journey in a digital world. Finally, insurers need to conceive new products that address emerging risks and opportunities in the digital age, such as coverage for cyber risk, gig economy workers, autonomous vehicles, and smart home technologies.

Fourth, insurers struggle with legacy culture and skills. Employees who look back on a long tenure tend to see the digital transformation as a threat rather than an opportunity. The importance of culture as an intangible asset can hardly be overestimated. As Li et al. (2021b) show in a broad study of almost 2,900 US firms, companies with a strong innovation culture are more proactive in their digital transformation efforts.

[1] As a consequence, some of the sections in this chapter need to rely on practitioner sources.

[2] According to the same industry report, approximately a third of all insurance technology stacks are not ready for cloud application and are not well suited for mobile devices.

Apart from cultural issues, the skill sets of insurance employees are based on the business of the past and often lack competence with new technologies. In a recent report, BCG (2021b) estimates that the number of insurance employees in areas such as analytics will have to rise by 50%. Similarly, the overall fraction of roles with a digital focus, such as software programmers and data scientists will increase from approximately 20% today to about 70%. The insurance industry is thus facing a skills gap (Swiss Insurance Association 2021). Acquiring or building the necessary skills and going from the industry's legacy culture to a culture of continuing innovation will likely be the greatest obstacle that insurance companies have to overcome in the digital transformation (McKinsey 2017).

2.2 Rewards

Despite the significant legacy issues discussed in the previous section, evidence on the rewards of digital transformation initiatives suggests that they are worth the effort. Bohnert et al. (2019) consider how a digital agenda in the annual reports impacts the business performance of 41 exchange-listed European insurers in the period from 2007 to 2017. They find a positive effect, which is particularly strong for comprehensive initiatives that cover both the processes within the organization and the interface to the customer.

Similar results are provided by BCG (2020a). According to this report, digital maturity exhibits a strong correlation with performance. Digitally more advanced insurers outperformed their peers in key business indicators such as customer satisfaction, revenue expansion, and operational excellence, which drive shareholder value. To create this success, however, digital leaders commit 16% of their operating expenditures (OPEX) and 18% of full time equivalent (FTE) employees to digital initiatives, compared to 6% of OPEX and FTE for digital laggards.

Another scholarly study on digital transformation as a success factor is contributed by Fritzsch et al. (2021), who analyze the relation between digital transformation and the market capitalization of US insurance companies. Since the degree of digital maturity is difficult to capture, they create a text-based measure that can be computed from annual reports with the help of a machine learning algorithm. In line with the aforementioned studies, their evidence also points to a positive relationship between digital transformation efforts and firm value in the insurance industry.

Most recently, Eckert et al. (2022) concentrate on the connection between digital transformation and customer satisfaction. By examining and evaluating a range of digital applications from an extensive analysis of 106 scholarly articles, industry reports, and regulatory publications, they highlight the potential of digital efforts to enhance customer satisfaction at key interaction points: contract initiation, contract alterations, and claims reporting. Similarly, Lanfranchi and Grassi (2022) conduct a qualitative study on innovation in the insurance industry in general, focusing on the measures implemented by the world's most prominent insurance companies. They highlight that technology-based innovation in core value chain steps determines the success and competitiveness of the insurance business.

3 Transformation of the Insurance Value Chain

3.1 Overview

The digital transformation will affect the whole insurance value chain (Eling and Lehmann 2018). To retain their competitiveness in the future, insurers have to transform their core business with new technologies. The two main goals in this regard are an automation of processes to reduce costs and the

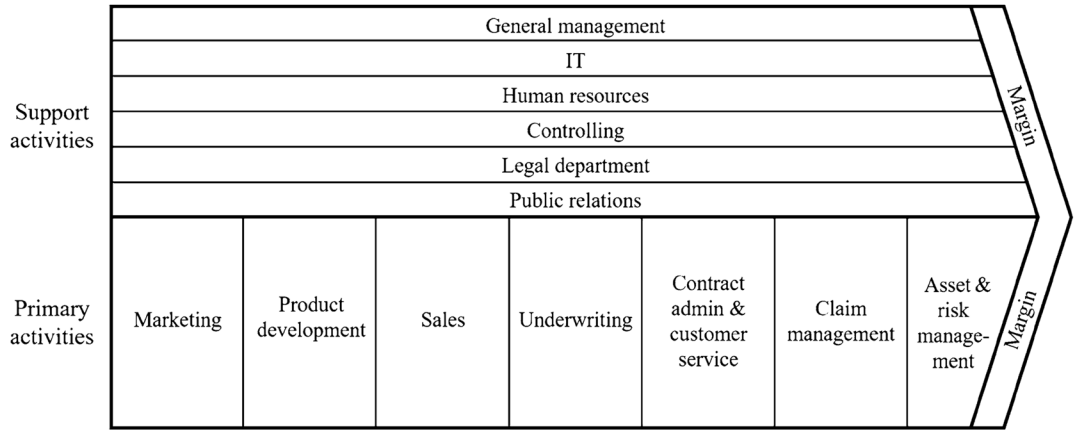


Fig. 1 Insurance value chain based on Porter (1985) and Rahlfs and Döring (2007)

improvement of the experience for the customer. Throughout this section, we revisit the major steps of the value chain and highlight the impact of digital transformation. In line with Eling and Lehmann (2018), we adopt the insurance-specific value chain logic introduced by Rahlfs and Döring (2007), which is shown in Fig. 1. We will concentrate our discussion on the primary activities: marketing, product development, sales (distribution), underwriting, and claims management.[3]

3.2 Marketing

Eling and Lehmann (2018) provide an overview of the possibilities opened by digital transformation in insurance marketing. Innovative technologies are transforming the marketing landscape for insurance companies, empowering them to tap into the potential of artificial intelligence (AI), machine learning, and big data analytics. By leveraging these cutting edge tools, insurers can deliver personalized experiences, streamline marketing campaigns, and optimize customer segmentation and targeting. Advanced data analytics, for instance, enable insurance firms to analyze vast amounts of customer information, unveiling crucial insights for market segmentation and targeted messaging. Furthermore, AI-powered chatbots and virtual assistants are revolutionizing customer service, offering real-time support around the clock and guidance throughout the customer journey.

3.3 Product Development

Telematics-Based Insurance

Telematics, an innovative blend of telecommunications and informatics, enables insurers to better assess risk and tailor policies based on individual behavior (Swiss Re 2017). In automobile insurance, "Pay How You Drive" (PHYD) policies evaluate driving habits, such as speed, acceleration, and braking, to determine premiums.[4] This encourages safer driving practices and reduces risk for

[3] We do not further discuss asset management and risk management, although these core functions of insurance companies also benefit from the digital transformation.

[4] We categorize the more basic "Pay As You Drive" (PAYD) policies as on-demand insurance (see the next section), because they solely focus on charging customers based on the number of miles driven, thus rewarding low-mileage

insurers. In health insurance, telematics devices like wearable fitness trackers and smartphone apps gather data on physical activity, heart rate, and sleep patterns, offering personalized incentives for maintaining a healthy lifestyle. Similarly, home insurance leverages smart home devices such as security systems, smoke detectors, and leak sensors, helping policyholders prevent potential damages while reducing claims for insurers. Overall, telematics technology will transform the insurance landscape by providing data-driven insights that promote responsible behavior and reward customers for their diligence.

There is an amplitude of evidence which shows that telematics is an effective tool to predict crash rates. Eling and Kraft (2020) provide a systematic summary of existing model approaches in telematics insurance. There are several risk factors that can be used to predict those crash rates. For instance, Stipancic et al. (2018) show that hard braking and accelerating events are positively correlated with crash frequency. Weidner et al. (2016) analyze maneuvers, trips, trip sections, and the overall insurance period and show that with this information it is possible to gain valuable insights into an individual's driving behavior. Geyer et al. (2020) conducted a study on driving behavior to examine the relationship between private information in driving characteristics and contract choice and risk in automobile insurance. They define a driving factor based on distance driven, number of car rides, and speeding and find that the number of car rides and relative distance driven on weekends are significant risk factors. This is similar to the findings of Baecke and Bocca (2017), who recommend to base the final premium on the total amount of kilometers driven in a predefined timeslot. If the product can contain more risk factors, time and location of the driving behavior can be used to assess the risk more accurately (Baecke and Bocca 2017). Lastly, Guillen et al. (2020) show the significant explanatory power of near-misses in the context of automobile insurance. Besides risk factors, the model assessing the risk factors also influences the crash rate predictions. Logistic regression models or more complex machine learning techniques might be used to investigate the use of driving behavior variables for predicting the risk and frequency of claims in UBI products (Huang and Meng 2019).

Eling and Kraft (2020) show that telematics technology reduces information asymmetries and improves risk classification, therefore improving the insurability of risks. Interestingly, Geyer et al. (2020) find that speeding and number of car rides were negatively related to high third-party insurance, which contradicts adverse selection. In contrast, the number of car rides and relative distance driven at night are positively related to first-party coverage, which is in line with adverse selection. The study also documents that risk-averse individuals tend to buy more liability coverage and drive their cars more considerately by taking fewer short rides, while telematics policyholders drive less than the average.

Often, personalized feedback alone has the potential to induce significant changes in driving behavior (Ellison et al. 2015). However, the most significant reductions in risk occur when drivers are given a financial incentive to change. Guillen et al. (2021) show how to integrate telematics data in a pay-how-you-drive insurance pricing scheme that penalizes near-miss events. The authors propose a system that combines a basic insurance premium with additional fees for risky driving behaviors, such as hard braking, rapid acceleration, and using a smartphone while driving.

While telematics can be used to screen consumer characteristics, there are some concerns. One is the significant costs caused by the high data volume (Weidner and Transchel 2015), for instance, when the data becomes intractably large. This is especially a problem when the data quality is not perfect and noise influences the data. Another issue is privacy concerns (Gao et al. 2022). Recording personal driving data might lead to large discomfort for a significant share of the population. Gemmo et al. (2019) consider telematics contracts in the presence of consumers that value their privacy and

drivers. In contrast to PHYD insurance, however, the PAYD concept does not use data from the car or connected devices for risk classification, pricing or the mitigation of moral hazard.

dislike sharing private information with insurers. In a second-best efficient Miyazaki–Wilson–Spence (MWS) framework, they show that information disclosure can substitute deductibles for consumers whose transparency aversion is sufficiently low. This can lead to a Pareto improvement of social welfare. Furthermore, they show that telematics may challenge cross-subsidization in insurance markets and introduces negative externalities for consumers who are unwilling to take part in this development.

Finally, despite the potential benefits and theoretical predictions supporting the adoption of a screening technology in equilibrium, market demand for telematics products remains low. Holzapfel et al. (2023) offer a theoretical explanation for this outcome. They explore the impact of insurers' ability to monitor policyholders in a moral hazard framework, arguing that privacy concerns and technology costs are unlikely to be major impediments. The authors modify the traditional moral hazard model of Shavell (1979) with a technology that informs insurers about the prevention efforts of the policyholders. They find that informative monitoring technologies which lack sufficient accuracy will not be adopted in equilibrium. The reason is that the advantage of increased insurance coverage is negated by the risk of being incorrectly classified and thus losing the premium discount.

On-demand Insurance

On-demand insurance is an innovative concept introduced by the InsurTech sector (Braun and Schreiber 2017). It is rapidly gaining traction, with some experts predicting that this emerging market could expand to $190 billion by 2026 (e.g., Garth 2019). Braun et al. (2023) define on-demand insurance as coverage for episodic risks, offering consumers tailored protection during specific periods when they or their possessions are exposed. According to Braun et al. (2023), on-demand insurance comprises two forms: (i) short-term insurance and (ii) usage-based insurance (UBI). Short-term insurance is designed to address risks that are time-bound or recurring and have a known or at least reliably predictable exposure pattern. Common examples of short-term insurance products are single-trip travel insurance, temporary accident coverage, short-term auto insurance, homeshare insurance,[5] and project-based or seasonal business liability insurance. On the other hand, UBI is suitable for situations where consumers' risk exposure varies stochastically over time. Examples of this type of insurance include pay-per-mile (also known as pay-as-you-drive) auto insurance, pay-per-use policies for bicycles and cameras, gig worker liability coverage, and ride-hailing coverage.[6]

Zeier Röschmann et al. (2022) propose a similar definition that also relies on the "coverage as-needed" concept. Instead of the stochasticity of the exposure profile, however, they differentiate subforms according to the mode of activation of coverage. In standard on-demand insurance, the user can switch on coverage manually or is prompted to do so by a smart device. Under UBI, in contrast, coverage is activated fully automatically, conditional on factors such as location, activity, or context. In addition to their definition, Zeier Röschmann et al. (2022) provide an overview of current on-demand insurance offerings, characterize the on-demand insurance business model, assess how on-demand insurance creates value, and gauge the interest of potential customers through a survey in Switzerland. Their results show that on-demand policies are a better fit for certain customer needs than traditional contracts. Thus incumbent insurers should innovate in this direction as well if they want to maintain their customer interfaces in the future.

An empirical analysis of UBI is conducted by Che et al. (2022). Utilizing a difference-in-difference research approach combined with firm fixed effects, they present evidence that UBI enhances the underwriting performance of early UBI adopters among private passenger auto liability insurers

[5] Homeshare insurance risk profiles are known when property owners rent out their space during specific times, such as Munich residents renting out through AirBnB exclusively during Oktoberfest.

[6] Refer to the websites of Cuvva, Lings, and Slice for real-world instances.

through lower loss ratios. Moreover, UBI yields benefits only when it reaches maturity. Their results are robust against self-selection bias and reverse causality tests. Further analyses document a considerable growth in the market share of early UBI adopters. Finally, Che et al. (2022) estimate the total performance impact of UBI programs for early adopters to correspond to a 1% increase in the Return on Assets (ROA) and a 3% increase in the Return on Equity (ROE). Their findings are consistent with a first mover advantage for UBI providers and the ability of UBI to draw low-risk drivers from competing insurers.

Insurance for Emerging Risks

Product innovation in insurance is also essential to address the rapidly evolving landscape of emerging risks (The Geneva Association 2023). As new threats materialize, such as cyber-attacks, consequences of climate change, and pandemics, the insurance industry must adapt its offerings to protect individuals and businesses from these unprecedented challenges. By developing solutions and leveraging advanced technologies, insurers must strive to overcome limits to insurability and provide households and businesses with the necessary protection against these emerging risks. This type of product innovation not only demonstrates the industry's ability to respond to a dynamic risk environment but also highlights the crucial role of insurance in fostering resilience and promoting sustainable growth in an increasingly interconnected world.

3.4 Distribution

Online Sales

A central question in the context of the digital transformation in the insurance sector is how insurance distribution will change. Online insurance sales, whether conducted directly through an insurance carrier's website and apps, through comparison portals (aggregators) or through conversational robo-advisors, have been gaining notable traction in recent years.[7] While the technical possibilities have been around for much longer, this recent growth is primarily driven by changing consumer preferences. Customers increasingly value the convenience, ease, transparency, and speed of purchasing insurance online compared to traditional channels, a trend that has been accelerated by COVID-19 (Braun et al. 2020). In addition, policies sold online are often priced more competitively.

Insurance carriers' websites allow customers to purchase policies, manage their accounts, and file claims directly. These platforms provide a seamless, efficient, and user-friendly experience, which contributes to customer satisfaction and loyalty. In contrast, comparison portals, or aggregators, enable consumers to compare multiple insurance products from different carriers, making it easier to find the best coverage and pricing options that suit their needs. Finally, robo-advisors in the financial sector offer automated, algorithm-driven financial planning services with minimal human intervention. These digital platforms gather information from clients about their financial situation and future goals and then leverage that data to offer advice on coverage, providing a streamlined and cost-effective distribution alternative. Recent research shows that robo-advisors are particularly effective in building trust and creating a positive perception of the company, if they rely on a conversational interface (Hildebrand and Bergner 2021). Customer surveys (Statista 2022b) show that the acceptance of such artificial conversational agents is increasing. More specifically, in 2022, about 12% of the customers in developed nations expressed a willingness to consult a robo-advisor

[7] A study by McKinsey (2017), e.g., revealed that 25% of U.S. consumers who look for auto insurance already purchase their policy online directly from the insurance company.

for insurance advice, whereas in emerging countries like Brazil and India, this percentage is as high as 27%.

The outlook for online insurance sales remains promising, with continuous advancements in technology and shifting consumer preferences contributing to its growth. Insurers should adapt their strategies to capitalize on this trend, enhancing their digital offerings and ensuring a seamless, customer-centric experience. As more and more consumers become used to conducting their essential transactions through digital channels, the expansion of online insurance sales is going to continue (Braun et al. 2020).

Bionic Agent

In many countries, insurers traditionally rely on intermediaries such as agents or brokers to sell their products.[8] Although the importance of online sales is expected to increase in the future, industry experts do not expect agents and brokers to disappear (Swiss Re 2020). Instead, the future is likely to bring about "bionic" insurance agents that combine the best of both human expertise and AI technology to create an enhanced insurance sales and service experience (BCG 2022). This fusion allows insurance agents to capitalize on the strengths of AI, such as data analysis, automation, and pattern recognition while preserving the empathetic and personal side of human agents.

In a bionic insurance agent setup, AI-powered tools are employed to handle routine tasks, such as data entry, policy comparisons, and claims processing. They can even be used to train agents (Luo et al. 2021). This frees up human employees to focus on interacting and building relationships with clients, providing tailored advice, and resolving complex issues. Moreover, AI can analyze vast amounts of data on customer trends, preferences, and risks, to suggest agents the best next step so that they can make better decisions and recommendations. Building up the competency for high-quality advice on the more complex and longer term insurance products (e.g., in life insurance) will be a key success factor for bionic agents because simple short-term products such as automobile insurance will be increasingly sold through online channels.

An important element is the combination of the bionic agent with a seamless omni-channel distribution. Customers frequently desire to change channels throughout the sales process. It is crucial for insurers to predict a customer's unique journey and provide appropriate self-service platforms, digital resources, and remote engagement, e.g., through video call technology or virtual interactions in the metaverse. AI will help to identify the optimal moment at which an in-person consultation is needed (BCG 2020b). All communications between agents and their clients are digitally captured, generating valuable data for the improvement of the insurers' sales processes.

The bionic insurance agent concept benefits both the insurer and the customer. Insurers gain improved efficiency, reduced human error, and increased customer satisfaction. Customers, on the other hand, receive a more accurate and customized service, as well as quicker response times, resulting in an overall enhanced insurance experience. It should be emphasized, however, that the technological boost in agent productivity clearly implies that fewer agents will be required in the future (McKinsey 2021b).

Embedded Insurance

Embedded insurance represents a paradigm shift in the way insurance is offered and consumed. The idea is to integrate insurance coverage seamlessly at the point of sale of noninsurance products and services, creating a streamlined customer experience (see, e.g., Stice and Besio 2022). With the increasing adoption of digital platforms, embedded insurance has emerged as a game-changing

[8] In some markets, particularly throughout Southern Europe and Latin America, bancassurance is the main distribution channel (Braun et al. 2018).

solution that promises to revolutionize the insurance industry. It substantially simplifies the insurance purchasing process for the customers by creating a one-stop-shop experience for their insurance and noninsurance needs and thus eliminating the need for multiple transactions. While embedded insurance accounted for a mere 6% of all global insurance sales in 2020, its potential is considered to be immense (Bewersdorf 2022).

The two primary forms of embedded insurance are linked embedded insurance and bundled embedded insurance. *Linked embedded insurance* refers to the practice of offering insurance as an add-on to a primary product or service. In this model, customers can choose to opt in for insurance coverage tailored to their specific needs, alongside the purchase of the main offering. A well-known example is the purchase of online travel insurance on the website of airlines or tourism companies. *Bundled embedded insurance*, on the other hand, is when insurance is included as an integral part of a product or service. In this case, customers automatically receive insurance coverage as part of the overall package, without the need for a separate purchase. An example is the inclusion of insurance coverage into car subscription models, which are expected to gain a significant market share in the next years (Lang et al. 2022).

From the perspective of insurance companies, embedded insurance provides several benefits. It enables insurers to access new customer segments (through platforms or partner websites), expand their market reach, and increase revenue streams. However, with the integration of insurance products into noninsurance platforms or services, insurance becomes more commoditized, and the customer focus will be mainly on price rather than on the quality or uniqueness of the offering. This could lead to downward pressure on premiums and, ultimately, margins (Standaert and Muylle 2022).

3.5 Underwriting and Pricing

In insurance underwriting and pricing, the digital transformation will have two main impacts: new methods and new data. AI, that is, advanced machine learning and deep learning models, will complement the traditional actuarial and statistical models and allow for a more accurate risk assessment, greater operational efficiency through reduced manual processes, and an enhanced customer experience through faster decisions (Eling and Lehmann 2018).

However, apart from these benefits, the advent of AI also has downsides. The biggest issue is unfair bias, which Smith et al. (2022) define as "unexplained adverse outcomes for marginalized communities." AI models themselves do not inherently produce biases; they arise from the underlying training data. The insurance sector has not been free of discrimination related to class, race, and gender. Minorities, e.g., may pay up to $100 more for car insurance (Kerzner 2023). Using modern AI techniques with historical data can amplify such biases present in the data, inadvertently harming specific demographic groups. To mitigate these biases in insurance, AI systems must be used thoughtfully and they need to be trained on high-quality, relevant, and bias-free data. Otherwise, the insurance industry risks perpetuating historical prejudices that negatively impact both its reputation and its customers. In fact, the issue of unfair bias in insurance AI is already so prevalent that it has called regulators into action (see, e.g., EIOPA 2021; Scism 2023).

New data sources can mitigate the issue of unfair bias and, at the same time, open up a whole new array of possibilities in underwriting. Such data sources comprise, but are not limited to, social media, drones and satellite imagery, smartphone sensor data, and connected devices (Internet of Things, IoT) (McKinsey 2017). IoT devices in particular can provide real-time data feeds on the risk, such as driving behavior and vital signs, enabling insurers to eliminate information asymmetries and apply personalized and dynamic pricing approaches. In many cases, however, the potential will depend on the customers' willingness to share their data with the insurance company. For transparency-averse

customers, insurers will have to offer monetary incentives in telematics policies (Gemmo et al. 2019). Biener et al. (2020) estimate that the decrease in customers' willingness to pay for insurance caused by digital monitoring can be as high as 25–50% of the expected loss.

An interesting new perspective on underwriting opens up with innovative product designs. Braun et al. (2023) show that on-demand insurance contracts can be used for risk classification in markets with adverse selection. They concentrate on the subcategory of short-term insurance (see Sect. 3.3), for which exposure profiles over time are deterministic or reliably predictable. Their model with two consumer types allows for the individual risk to differ across subperiods. In this setting, introducing short-term on-demand insurance contracts improves efficiency compared to the case without it. Moreover, it enables type-specific full fair insurance, if the consumers who are less risky from an annual perspective exhibit the higher risk in the subperiods in which they are exposed. These results remain robust under various additional policy constraints.

3.6 Customer Service

An increasingly notable shift in customer service involves the use of artificial conversational agents to handle interactions and claims settlements, replacing or supplementing human representatives. Artificial conversational agents, or chatbots, are software programs that use computational linguistics to interpret and respond to natural language requests and statements from human users. While they are not capable of autonomous thought like humans, chatbots can be designed to imitate the flow of human conversation and continuously improve their responses based on interactions with users. Lemonade, a US digital insurance company, has developed two conversational agents named Maya and Jim for underwriting and claims settling (Biener 2019). It states that this is associated with significantly reduced labor costs in these areas.

With new advancements in AI, the quality of these services is expected to improve at a fast pace. Chat-GPT alone is expected to hugely transform the industry. However, using chatbots may have some downsides. Besides the preference of some people to talk to actual humans, there are more subtle shortcomings, for instance, when using conversational agents to handle claims. Honest reporting of unobservable or partially observable outcomes is crucial for insurance claims. However, Biener and Waeber (2021) find that in an experimental setup customers are less honest to a chatbot, if the perceived capacity of the chatbot to exert self-control is low. This is contrary to human interactions where individuals are more.

3.7 Claims Management

The use of mobile apps and digital customer portals has allowed customers to submit claims easily and quickly, track their claim status in real time, and communicate with their insurer directly through their mobile devices. Additionally, with the use of AI and machine learning algorithms, insurance companies can now automate and streamline the claims process. This reduces the need for human intervention and accelerates claims handling. Using AI systems, the InsurTech *Lemonade*, for instance, managed to handle claims in as little as 3 seconds. McKinsey (2017) estimates that automatization can reduce the costs of a claim journey by up to 30%. It should be noted, however, that AI-based claims handling can be subject to the same unfair biases that we already discussed above in the context of AI-based underwriting. When training and implementing such systems, insurance companies must therefore take great care that claims of minorities are not rejected on unfair grounds. Given that bias can be avoided, AI also promises to lead to more accurate claims assessments and

a better detection of fraudulent claims. Three technical developments play a significant role in this regard.

The first relevant technical development for fraud detection is the increase in IoT devices. Those are embedded with sensors, software, and connectivity that enable them to collect data and send it over the Internet. For insurance applications, smart home devices such as smart cameras, thermostats or home security systems are very useful. They can be very effective in proving fraudulent behavior. For instance, GPS pattern data might be used to verify that stolen goods were indeed stolen by a third party. The most revolutionary usage in fraud detection is by combining the information of multiple IoT devices with other data sources. For instance, if smart home devices are combined with weather data, fraudulent water damage claims might be detected (Stadelmann et al. 2023). The growth of IoT is extraordinary. It is estimated that there will be more than 50 billion IoT devices by 2025 (McKinsey 2017).

The second development is the usage of blockchain for fraud prevention. Blockchains can be used as a distributed and cross-industry registry of data, allowing to verify identities and to validate the authenticity of documents, the ownership of assets, and the provenance of goods (McKinsey 2017).

The third technical development is in the field of statistical methods and AI. While the statistical methods were often known for a long time, they were not feasible on a large scale due to prohibitively expensive computer power or less collected data.[9] The machine learning methods for fraud detection can be divided into two categories. With supervised learning an algorithm is trained using a labeled dataset that includes both fraudulent and non-fraudulent examples. The model is then able to use the patterns it has learned to accurately classify new, previously unseen transactions as either fraudulent or legitimate. In the context of fraud detection, supervised learning is useful because it allows the model to learn from known examples of fraudulent behavior and to identify similar patterns in new transactions. For instance, Li et al. (2021a) used nonparametric regression model to detect misrepresentation, i.e., exaggeration of claims. To be more specific, they used quantile regression which can be used to model the conditional quantile function of losses. This is different from classic/mean regression where the average amount of losses is estimated.

In contrast to these supervised machine learning methods, unsupervised learning involves training a model on an unlabeled dataset, meaning that the model does not have access to any information about which transactions are fraudulent and which are not. Instead, the model identifies patterns and anomalies in the data on its own. This approach is often used in situations where it is difficult to obtain labeled data for training or when the fraudulent activity is changing and difficult to define. Besides classical unsupervised learning methods like clustering, machine learning in fraud detection has developed some more refined methods. Gomes et al. (2021), for instance, applied deep learning methods to detect fraud. Another example, which highlights the relationships between subjects and accidents, is bipartite networks (Tumminello et al. 2023). Since the counterparty in an accident should be random, anomalies can be used to detect fraud.

4 New Business Models

4.1 Definitions and Overview

A business model is the logic of how a company creates, delivers, and captures value (Gassmann et al. 2014). It outlines the components and relationships that are crucial for the company's success.

[9] As a caveat, while the computing power per dollar increases on average, modern machine learning techniques usually also use much more resources.

Standard definitions characterize a business model along four dimensions (Teece 2010; Gassmann et al. 2014): (i) the target customers, (ii) the products or services that the company offers them, (iii) the value chain or operating model on which it relies to deliver the products and services, and (iv) the revenue mechanism through which the company recaptures some of the created value for itself. To innovate a business model, companies need to change at least two of these four dimensions.[10]

In practice, business model innovation is often hampered by a lack of out-of-the-box thinking and the failure to anticipate future customer needs. Particularly well-performing companies are at risk of falling prey to the innovator's dilemma, because their success tends to blind them toward major shifts in market conditions, consumer behavior, or technology that will put pressure on their existing business model (Bower and Christensen 1995).

Prior to digital transformation, business model innovation has not been a major concern in the insurance industry. This changed with the advent of the InsurTech sector that, from the beginning, has experimented with innovative business models for the insurance industry (Braun and Schreiber 2017). In response to these developments, awareness among incumbents has notably risen in recent years, and many are now exploring new business models as well, in order to protect themselves from the potentially disruptive forces of the digital transformation.

In this section, we will briefly discuss the most important new business models that may change the nature of the insurance industry. Those are the digital attacker, the ecosystem orchestrator, and the prevention partner, as well as the concepts of online mutual aid and decentralized insurance.

4.2 Digital Attacker

Digital attackers rely on a lean organizational structure and use innovative technologies, such as AI, to automate their operations, increase efficiency, offer customized policies, and reduce costs (McKinsey 2021a). Examples include the InsurTechs Lemonade, Hippo (both USA) and Element (Germany) but also subsidiaries of incumbents, such as Smile (Switzerland), Friday and Huk24 (both Germany).

The digital attacker model distinguishes itself from the classical insurance model in the customer dimension, the product dimension, and the value chain dimension. It is targeted at tech-savvy customers who value convenience, transparency, customization, and fast turnaround times. The goal of digital attackers is that customers can access insurance products and services anytime, anywhere, through user-friendly interfaces. Consistent with these target customers, the products sold by digital attackers are typically no-frills, yet customizable, insurance policies.

To deliver their value proposition, digital attackers rely on an end-to-end digitized value chain, as discussed in Sect. 3. In distribution, claims, and services, digital attackers pursue a digital-first approach, prioritizing direct-to-consumer sales through mobile apps and through aggregators or digital platforms (marketplaces). Moreover, digital attackers rely on data-driven decision-making. By harnessing the power of data analytics and AI, they can make rapid, automated, and well-informed decisions on pricing, underwriting, and risk management. This can lead to more accurate risk assessment, personalized pricing, and improved customer experiences. Thus, adopting the digital attacker business model requires insurance companies to master the challenges involved in transforming the value chain (see Sect. 3).

A fertile ground for digital attackers is evident in the shifting preferences of consumers toward online insurance management. Approximately one in four individuals from the United States (24%) has expressed willingness to handle all their insurances exclusively online. Interestingly, these preferences are even more pronounced in most European countries, where nearly 30% of the

[10] Unidimensional changes, e.g., classical product and process innovations, do not constitute a new business model.

population shares this sentiment (Statista 2022c). Furthermore, 45% of insurance customers in the United Kingdom expressed comfort, ranging from moderate to high, with the idea of obtaining a policy from noninsurance companies. This trend aligns with a global shift in customer willingness to explore new avenues for their insurance needs (Statista 2022d). Another Statista report (Statista 2022a) highlights that in developed nations, a growing segment of consumers would consider getting all their insurance from Internet giants like Google or Apple: 12% in the United States, 10% in the UK, and 8% in Germany. Interestingly, emerging markets showcased a notably higher tendency toward this trend, with India at 25%, Brazil at 18%, and Mexico at 16%.

4.3 Ecosystem Orchestrator

The digital transformation has given rise to the concept of business ecosystems (Weill and Woerner 2015). A business ecosystem is a network of interconnected entities that collaborate to create value through a combination of products and services. Access to the ecosystem greatly increases the convenience for customers, as they can satisfy various needs through a one-stop-shop concept or obtain a bundled solution without having to sequentially and separately interact with several providers.

In a business ecosystem, there are two primary roles: the orchestrator and the supplier (Moore 1993). The orchestrator is the central player who creates and manages a digital marketplace (also called platform), facilitating connections and collaborations among various suppliers. Orchestrators set standards for the marketplace and provide the core infrastructure that enables suppliers to engage with customers, ensuring seamless interactions and transactions. Some well-known orchestrators include the big tech companies Amazon, Apple, and Google. In contrast, suppliers depend on the marketplace to provide their products, services, or content. By leveraging the platform provided by the orchestrator, suppliers can reach customers, scale their businesses, and benefit from network effects.

To cope with the headwinds of the digital transformation, the insurance industry has begun to adopt the ecosystem logic (BCG 2021a). A simple way to leverage future ecosystems is to adopt the digital attacker model discussed above, which perfectly matches the supplier role. Many insurers, however, strive to become orchestrators instead (EY 2020). This is a challenging transformation which requires them to change all four dimensions of the business model. More specifically, insurers need to broaden the customer base beyond pure insurance purchasers, fully digitize their value chain (see Sect. 3), and switch to a commission-based instead of a premium-based revenue mechanism. Most importantly, however, the value proposition needs to change: orchestrators command a robust, scalable, and user-friendly platform. This platform needs to attract a large number of suppliers and customers and ensure many touchpoints, which is hardly achievable through classical insurance products.

Until today, the ecosystem strategies of most insurance companies are still in an early stage. The exception is Ping An, a leading Chinese conglomerate with a focus on insurance, banking, and financial services. Ping An has strategically leveraged technology to transform its business and create multiple digital ecosystems in various sectors: financial services, health and wellness, smart city, mobility, and home (Zhu et al. 2022). Below, we briefly consider the health ecosystem of Ping An, highlighting its main characteristics and success factors.[11]

[11] It should be noted that, in 2022, Ping An experienced its most significant profit drop in over a decade, with the fall attributed to the impact of COVID-19 and losses from problematic property developments. The company announced its annual results, highlighting a decline in group net profit to RMB 101 bn ($16 bn), a 29% decrease from the previous year (Smith and Hale 2022). This marks the worst dip since 2008. The insurer's life and health division was particularly affected, with a drop in demand for critical illness insurance products due to the pandemic. However, by the time of writing of this chapter, the firm had recovered with a 48.9% rise in the year-on-year net profit (Reuters 2023).

The core of Ping An's ecosystem is the Ping An Good Doctor App. This app connects the three main actors of the ecosystems: patients, medical service providers, and payers. The patients are the customers, and medical service providers include clinics, doctors, and pharmacies (Braun et al. 2019). The payment side is covered by entities like private insurance firms, including Ping An, and the Chinese government that provides fundamental health coverage. Over 440 million users, 3000 partner hospitals and over 208'000 pharmacies participate in this ecosystem (Ping An 2022). This popularity can be explained by two crucial advantages: a seamless customer journey and efficiency gains. Patients can use a 24/7 online consultation for a first assessment. Patients then have the option to continue their treatment offline with a specialist in a hospital if required. In addition, they have the convenience of buying their prescribed drugs online. The portal seamlessly interfaces with the health insurance provider who handles the expense of the treatment. Moreover, it allows easy access to personal health profiles and tailored health management plans. Therefore, Ping An offers not only continuous access to healthcare services but also a one-stop solution to cover all healthcare needs through a single app. Since this business model is highly scalable, efficiency gains can be realized, with respect to both cost savings and reductions in consumers waiting times for healthcare services.

4.4 Prevention Partner

Prevention as a new business model involves a proactive approach to risk management, where insurance companies focus on the use of digital technology to help their clients avoid losses and mitigate risks rather than just compensating them for incurred damages. Ehrlich and Becker (1972) distinguish two types of prevention efforts: loss reduction and loss prevention. The former reduces the magnitude of the damage given a loss event occurred, whereas the latter reduces the likelihood of the event itself. Hence, prevention as a concept itself is clearly not new. However, advanced analytics, AI, and IoT devices will allow for unprecented prevention measures. An intuitive example is the prevention of property damage due to water leaks. Smart water valves can automatically shut off the water supply to a specific area or the entire house in case of a detected leak. To remain relevant for their customers, insurers will increasingly need to advise them on risk prevention. The disruptive potential of this paradigm shift, which a large European insurance company has dubbed "from payer to partner," is vast. McKinsey (2017), e.g., estimates that profits for traditional auto insurance lines might fall by 40% from their peak, due to increased vehicle safety.

The scholarly literature on the new possibilities of prevention in the digital transformation is still scarce. Two notable contributions are Li and Peter (2021) and Li (2021). Using a subjective expected utility framework, Li and Peter (2021) examine the impact of technology risk on the risk mitigation behavior of consumers. Due to technology risk, decision-makers are often unable to accurately predict the effectiveness of their efforts. For instance, many individuals understand that following a balanced diet and exercising frequently have health benefits. However, they may struggle to determine the specific amount by which these actions reduce their risk of health problems. Li and Peter (2021) find that technology risk gives rise to a trade-off: on the one hand it impairs the effectiveness of risk mitigation, and on the other hand, it generates a precautionary motive. They identify cases, in which the resolution of this trade-off implies an increased effort, in line with the precautionary motive. Similarly, Li (2021) proposes the concept of technological transparency, which refers to the extent to which scientific knowledge enables individuals to predict the success of prevention measures based on observable risk determinants. The more risk determinants are uncovered, the better one can predict the success of any effort. Li (2021) finds that technology transparency unravels the insurance market and may have negative welfare effects, especially when information is only partially disclosed.

4.5 Online Mutual Aid

The online mutual aid model, initially developed in China, offers a cost-effective and inclusive alternative to traditional health insurance (Feng 2023). Participants do not pay a fixed ex ante premium. Instead, everybody pays ex post for the claims of its participants. While the premium amount is not known in advance, online mutual aid models have a significant cost advantage due to the very simple payment structure without much actuarial risk modeling. Furthermore, with no requirement for reserves or risk capital, compliance costs are reduced, contributing to the cost-effectiveness of this funding mechanism. In fact, the study by Chen et al. (2022) revealed that per-dollar coverage cost for comparable health insurance in China can be magnitudes higher than that of a mutual aid platform. For this reason, the model has become popular, particularly among middle- to low-income individuals in small- and medium-sized cities, filling a gap between national health insurance programs and commercial health insurance. Another advantage is the transparency and self-governance mechanisms. In contrast to the opaque operations of traditional insurance, online mutual aid platforms publicly disclose claims, and members are involved in the process to decide which claims are legitimate. This open disclosure procedure is considered an effective deterrent against fraudulent claims.

However, the number of users in China has significantly dwindled from over 350 million in 2021, with only a fraction of that number remaining today (Abdikerimova and Feng 2022). Major platforms such as Xianghubao, Meituan, and Waterdrop Inc. have all discontinued their mutual aid businesses. A common explanation given for this decline is regulatory challenges. Online mutual aid platforms resemble commercial insurance, yet they operate without appropriate regulatory oversight. The China Banking and Insurance Regulatory Commission has implemented stringent measures against these platforms due to the associated risks and their vast user base (Bloomberg 2021). The arguably much larger issue for these platforms is adverse selection. Since no risk classification takes place, adverse selection is especially prevalent on online mutual aid platforms. For example, a significant exodus occurred among the younger demographic from the platform Xianghubao: from June to December 2020, membership of those 39 and younger decreased from 71% to 67.7%, while the 40–59 age bracket saw an increase from 28.81% to 32.3% (The Straits Times 2022). As healthier, younger members departed, shared costs for the remaining participants rose. This erosion of costs compared to insurance products further accelerated the departure of these younger, healthier members.

In summary, online mutual aid presents a more affordable, transparent, and inclusive alternative to traditional health insurance. By leveraging prosocial behavior and low cost structures, it addresses the financial needs of people which have been excluded from traditional insurance options. However, as seen from the failure of this business model in China, it is questionable if such a system would be viable long term.

4.6 Decentralized Insurance

Decentralized insurance platforms are a new emerging approach to insure customers based on blockchain technology. These solutions carry the promise of lower agency costs, as suggested by Feng (2023), and they also offer potential benefits in emerging countries where the insured typically have limited legal protection (Braun et al. 2021).

Figure 2 illustrates the structure of a decentralized insurance system. Customers acquire a product such as an insurance contract, generated by a network of participants interconnected via the blockchain and self-executing contracts. One participant type is the relayer, akin to an insurance broker. The relayer links customers to the decentralized risk transfer system, potentially through a web

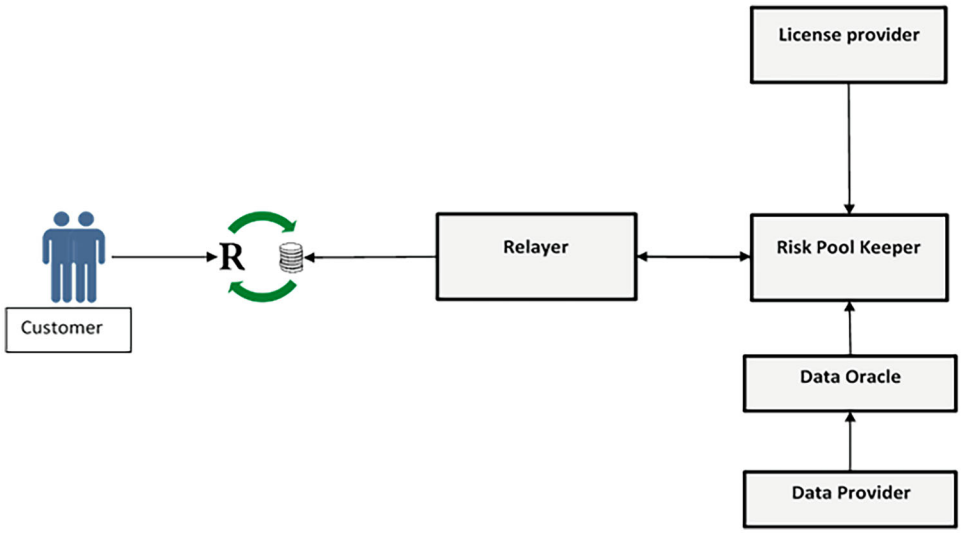


Fig. 2 Structure of a decentralized insurance network

portal where users select their insurance coverage. It is important to note that relayers are not exclusive and can charge distinct prices due to differing customer acquisition costs (Braun et al. 2021). Another role is that of the risk pool keeper, who oversees the portfolio of insurance policies and performs administrative tasks, including the creation of risk models and the establishment of rules for auxiliary contributors in the decentralized risk transfer system. Each risk pool corresponds to a single insurance product and includes the contracts and their premiums. Moreover, the decentralized insurance solution necessitates a license provider. Given the stringent regulation of insurance markets, it is not possible to provide coverage without an insurance license or adherence to regulatory authorities. License providers ensure the decentralized system's operation within this regulated framework. Lastly, the data provider and the data oracle are tasked with supplying raw data to the decentralized insurance system and transforming that data into useful information. This data may, e.g., include weather-related variables for parametric catastrophe insurance, utilized to compute premiums and to decide on potential payouts. These roles are integral to any decentralized insurance system. However, depending on the product and context, additional steps may be required. For instance, in the case of traditional loss-based policies as opposed to parametric products, the network also requires underwriters and claims handlers. Similarly, an additional payment agent could guarantee prompt indemnification to customers in their currency of choice (Braun et al. 2021).

Decentralized insurance has the most potential in emerging economies, where the insurance sector is often not properly developed or regulated. Without an adequate legal system, there are large power asymmetries between policyholders and insurance companies which rent-seeking monopolistic enterprises can take advantage of. For instance, when Hurricane Maria struck Puerto Rico in September 2017, 11,000 citizens were still waiting for their claims to be settled more than 1 year after the event (Andrews and Steklof 2018). Due to the complex nature of some insurance types, decentralized insurance works best for parametric property and casualty products with relatively short contract durations.

5 InsurTech

5.1 InsurTech Business Models

InsurTech, short for insurance technology, refers to start-ups that make innovative use of technology to transform and modernize the traditional insurance industry (Braun and Schreiber 2017).[12] InsurTech start-ups aim to streamline insurance processes, enhance customer experiences, and improve the overall efficiency of the insurance sector. They focus on various aspects of the insurance value chain, including distribution, underwriting, pricing, claims management, and customer service. Braun and Schreiber (2017) identify nine categories of InsurTech start-ups based on their underlying business models and order them along the risk to capital chain of the insurance industry:

- Comparison websites
- Digital brokers
- Embedded insurers
- Peer-to-peer insurers
- On-demand insurers
- Digital attackers
- AI InsurTechs
- Internet of Things InsurTechs
- Blockchain InsurTechs

Comparison websites, digital brokers, and embedded insurance specialists focus on distribution and the customer interface. Comparison websites, also called aggregators, allow customers to compare insurance products from various providers in one place. They simplify the process of finding and purchasing insurance policies, making it more convenient and efficient for customers. Well-known examples are insurify in the USA and Check24 in Germany. Digital brokers act as online intermediaries between insurance companies and customers. They streamline the process of purchasing policies, managing accounts, and filing claims through mobile apps, enhancing customer experience and reducing operational costs. wefox and Getsafe are representatives of this breed, which both expanded their business beyond the digital broker app to become a full-fledged digital carrier.

InsurTechs focusing on embedded insurance integrate insurance products and services into noninsurance platforms, products, or services, creating a seamless, contextual, and convenient customer experience (3.4). Embedded insurance allows customers to access insurance coverage as a part of their everyday transactions or activities, without the need to engage with traditional insurance channels. Firms engaged in this part of the InsurTech sector are CoverGenius, Penni.io, and Qover.

Further InsurTech categories are peer-to-peer (P2P) insurance, on-demand insurance, and digital attackers. As described in Sect. 4.2, digital attackers, such as Lemonade or the Swiss insurer Smile, are a new breed of risk carriers that leverage advanced digital technologies, data analytics, and AI to create faster, simpler, and cheaper insurance products than incumbents.

P2P and on-demand insurance start-ups can be managing general agents (MGAs), focusing on distribution, or they can be full risk carriers in the same way as the digital attackers. P2P insurance models involve groups of customers joining club-like insurance pools. Just like classical mutual

[12] Similar to Braun and Schreiber (2017), Stoeckli et al. (2018) scrutinized the distinct features and transformative potential of InsurTech innovations. They pinpointed the recurring themes in InsurTech's transformative capabilities, which include the implementation of digital and data-centric infrastructure operations and the provision and development of digital services.

insurance companies, digital P2P insurers return any unclaimed premium left in a pool to their members. If the customers in a pool know and trust each other, P2P insurance discourages moral hazard and insurance fraud and can thus lead to lower costs for policyholders. A prime example is the P2P model first introduced by German MGA Friendsurance. On-demand insurance providers, such as Cuvva in the United Kingdom or Lings in Switzerland, offer coverage for specific situations or timeframes (see Sect. 3.3). This enables customers to purchase policies when needed, with flexibility and customization.

Finally, Braun and Schreiber (2017) highlight three categories of InsurTechs that leverage emerging technologies like blockchain, IoT, and AI to innovate and transform various aspects of the insurance industry. Examples are manifold. Established InsurTech players in this segment are Friss, UiPath, and Root. In contrast to the InsurTech start-ups that focus on distribution or those that act as full-fledged risk carriers, these technology providers are enablers for the incumbents. Today, only 5% of InsurTech start-ups act as risk carriers that confront incumbent insurers with the goal of weakening their position and capturing market share (Gallagher 2023). The vast majority are enablers. Disruptors, i.e., firms that threaten the very existence of the insurance industry as we know it today, are virtually nonexistent (Braun and Schreiber 2017).

5.2 Relation to Incumbents

Traditional insurance companies face a strategic decision when dealing with the InsurTech sector: whether to adopt a collaborative approach and leverage the innovations and technologies brought by the start-ups or to confront them and maintain the traditional ways of doing business (Braun and Schreiber 2017). By partnering with InsurTech start-ups, incumbent insurers can benefit from innovative technologies, products, and services that help to transform their value chain, enhance their offerings, and streamline their operations. More specifically, Sosa and Montes (2022) identified five advantages InsurTechs might have. These include the implementation of digital capabilities, data-driven solutions, integrated services and customer experience, creation and development of digital insurance offers, and digital insurance distribution. Therefore, collaborating with InsurTech companies can help traditional insurers to foster an innovation culture, create a more customer-centric approach, and provide personalized and seamless experiences to policyholders.

Today, collaboration with start-ups has become an established part of the industry's innovation toolbox, helping incumbents to adapt more quickly to the changing market dynamics and evolving customer needs. Chang (2023) examines the impact of incumbent investments into InsurTech. They find that InsurTech-oriented investments have a negative impact on their short-term performance. This can be explained by the investment effect, due to the increase in spending. They also find a positive correlation between long-term performance and InsurTech investments. A likely explanation for this is that the collaborative advantages take some time to materialize, hence the time delay in performance.

There is, however, an area in which competition between InsurTech start-ups and incumbents is becoming more fierce. Both compete for talent in the insurance industry as they seek to attract skilled professionals who can drive innovation, growth, and success. InsurTech start-ups have several advantages when competing for talent in the insurance industry. First, they are typically more agile and innovative than traditional insurers, offering a fast-paced work environment where employees can make a significant impact and drive change. Second, a start-up culture promotes collaboration, creativity, and risk-taking, which can be attractive to professionals looking for a more entrepreneurial and less hierarchical work environment. Flat organizational structures and open communication channels can empower employees to contribute ideas and take ownership of their work. Third,

InsurTech start-ups work with the latest technologies, such as AI, blockchain, and IoT, providing employees with opportunities to hone their expertise in these areas and work on groundbreaking projects that can transform the industry. Finally, as start-ups grow and scale, employees often have the chance to take on new responsibilities, advance their careers, and potentially become part of the company's leadership team. This can be appealing to ambitious professionals seeking rapid career growth.

5.3 Facts and Figures

The InsurTech sector experienced rapid growth in recent years, with regard to both the number of start-ups and the invested capital. Figure 3 shows that the number of transactions and the total invested volume in US dollars grew until it reached a peak in 2021. In 2022, volume and number of transactions were smaller, likely due to the rising interest rates and macroeconomic uncertainty. Investments in InsurTech are conducted not only by venture capital (VC) firms but also by incumbent insurers and reinsurers. Gallagher (2023) reports only 1 year with less than 100 transactions by incumbents since 2016. Insurance companies investing in InsurTech are recognizing the potential for strategic advantages, such as innovation, growth, and increased competitiveness in the rapidly evolving insurance landscape. Through their investments, they get access to new technologies, ideas, and business models that can enhance their products and services, streamline operations, and improve customer experiences. Due to the legacy issues discussed in Sect. 2, such innovations are difficult to create in the organizational environments of incumbents.

Figure 4 breaks the InsurTech investments between 2012 and the fourth quarter of 2022 down by country and investment stage. VC financing is typically categorized into different stages based on the maturity of a start-up and the purpose of the investment. The initial stage, seed/angel investment, is where early-stage innovators receive capital to support product development and market research. This capital often comes from individual investors or angel groups. As the start-up progresses, it might secure a Series A round, which is designed to optimize product offerings and fund initial marketing efforts. With a proven business model and increased operational growth, a company can then move to a Series B round. This stage often focuses on expanding market reach, scaling production, and growing the team. Subsequent rounds, like Series C and beyond, generally cater to more established companies aiming for activities like mergers and acquisitions, expanding to new markets, or preparing

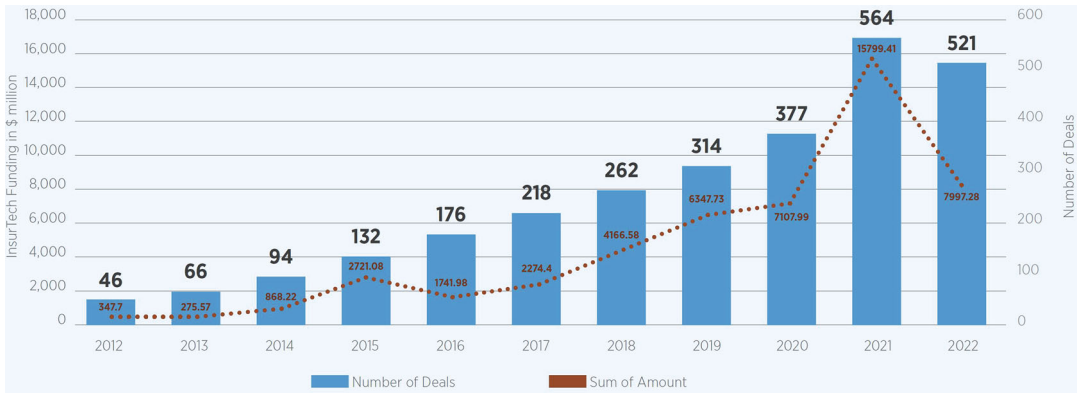


Fig. 3 InsurTech investment by year. (Source: Gallagher 2023)

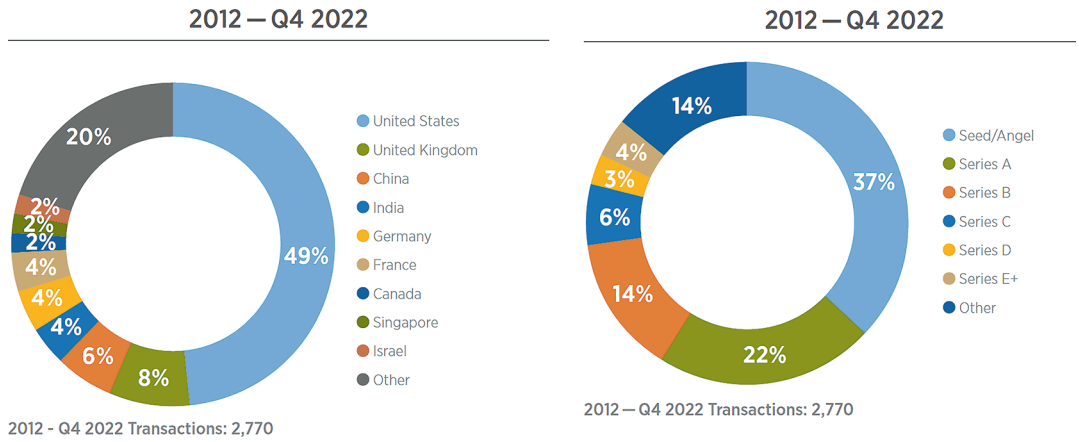


Fig. 4 InsurTech investment by target country and stage. (Source: Gallagher 2023)

for an initial public offering (IPO). Each stage represents a step forward in the company's lifecycle, often indicating increased valuations and decreasing levels of investment risk.

About 50% of the InsurTech activity relates to the United States, which corresponds to the global importance of Silicon Valley as an innovation and technology hub. Further hotspots are the UK, China, India, Germany, and France. In terms of the start-up life cycle, investments still focus on the early stages seed and series A. This reflects the general funnel logic of the VC industry in which only a fraction of the start-ups in each stage are successful in acquiring a follow-up funding.

Neale et al. (2020) identified the used technology in InsurTech-investments based on data collected from Willis & CB Insights InsurTech briefings, Google searches, and Lexis-Nexis in 2016–2018. A total of 174 instances of InsurTech investments were analyzed. The majority of these investments (38%) were associated with management software technologies, including automated or distribution platforms. AI technology was the second most common investment target during this period, making up 34% of the total investments. Predictive technology is also featured prominently, accounting for 15% of the investments.

Since 2012, half of the funded InsurTechs have focused on the "distribution" part of the insurance value chain. This involves roles such as lead generators, brokers, or originators of risk, but without any risk-bearing functionality (Gallagher 2023). Following closely, full-stack InsurTechs, which tackle the entire end-to-end insurer value chain, account for 26% of fundraises. B2B technology vendors operating on a SaaS model and addressing specific links in the value chain comprise the remaining 24% (Gallagher 2023).

Figure 5 categorizes the total VC transactions in the InsurTech sector during the time period from 2012 to the end of 2022 into property & casualty (P&C) insurance (upper subfigure) versus life insurance (lower subfigure) and by role in the industry (distribution, risk carrier, or business-to-business focus). The first observation is that the P&C sector is currently more comprehensively covered by InsurTech than the life sector. In both cases, the vast majority of start-ups focus on distribution. Risk carriers, in contrast, only account for a small fraction of the market. There are several reasons for the focus of InsurTechs on distribution (see, e.g., Braun and Schreiber 2017). First, traditional insurance buying processes can be complex and time-consuming. InsurTechs aim to enhance the customer experience by simplifying and digitizing this process, making it more intuitive and user-friendly. Second, compared to the creation of a full-stack risk carrier, the focus on distribution is associated with lower barriers to entry. It does not require significant capital reserves

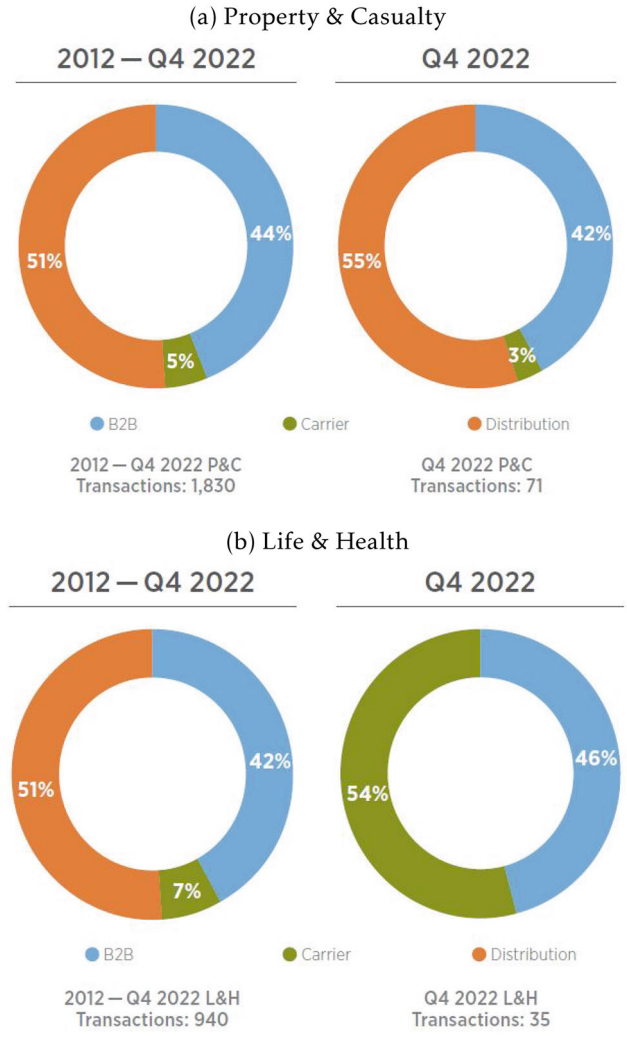


Fig. 5 InsurTech investment by insurance subsector. (Source: Gallagher 2023)

or complex risk models, making it an attractive starting point for start-ups. Third, by focusing on distribution, InsurTechs are able to own the customer relationship and collect a wealth of customer data. This data can become highly valuable for refining marketing strategies, understanding customer behavior, and potentially informing future expansions into underwriting or other areas of insurance.

6 Conclusion

In conclusion, the digital transformation in the insurance industry has ushered in a new era of innovation, presenting both challenges and rewards for incumbent insurers. As we have explored in this chapter, the transformation journey encompasses various aspects, including the transformation of the insurance value chain, the emergence of new business models, and the pivotal role of InsurTech in driving change.

The challenges of digital transformation are manifold, ranging from modernizing legacy systems over establishing a digital culture in incumbent insurance companies to acquiring the necessary skills for the future of insurance. However, the rewards are significant, with insurers standing to benefit from increased operational efficiency, a substantially enhanced customer experience, and evidence-based decision-making through data analytics. The insurance value chain is being transformed by the integration of advanced technologies, which streamline operations, automate repetitive activities, improve risk assessment, and facilitate seamless interactions with customers. Moreover, the emergence of new business models, such as digital attackers, ecosystem orchestrators, and prevention consultants, is challenging in the traditional industry logic. To cope with these developments, insurers need to adapt and evolve, or they risk being left behind. Lastly, InsurTech start-ups have played a critical role in driving digital transformation, bringing fresh ideas, agility, and technological prowess to the industry. By collaborating with or investing in InsurTechs, traditional insurers can accelerate their own transformation efforts, access new markets, and harness the power of innovation to stay competitive in the digital age.

In light of these developments, it is clear that the insurance industry is in the midst of a profound shift, with digital transformation poised to redefine the very fundamentals of the business. Success in this new era hinges on embracing change, fostering a culture of innovation, and relentlessly focusing on delivering value to customers. As the industry continues to evolve, those who can navigate the challenges and seize the opportunities presented by digital transformation will be well-positioned to thrive in the future of insurance.

References

Abdikerimova S, Feng R (2022) Peer-to-peer multi-risk insurance and mutual aid. Eur J Oper Res 299(2):735–749

Andrews WJ, Steklof CD (2018) Unpaid hurricane maria insurance claims, new laws in puerto rico, and the lesson for all policyholders. https://www.huntoninsurancerecoveryblog.com/2018/12/articles/bad-faith/unpaid-hurricane-maria-insurance-claims-new-laws-puerto-rico-lesson-policyholders

Baecke P, Bocca L (2017) The value of vehicle telematics data in insurance risk selection processes. Decis Support Syst 98:69–79

Bauer D, Tyler Leverty J, Schmit J, Sydnor J (2021) Symposium on insure-tech, digitalization, and big-data techniques in risk management and insurance. J Risk Insurance 88(3):525–528

BCG (2016) Building a digital technology foundation in insurance. https://www.bcg.com/publications/2016/building-a-digital-technology-foundation-in-insurance

BCG (2018) Start, enable, and scale digital transformation in insurance. https://www.bcg.com/industries/insurance/digital-transformation-insurance

BCG (2020a) Digital underinvestment hurts insurers. https://www.bcg.com/publications/2020/digital-underinvestment-hurts-insurers

BCG (2020b) What lies beyond digital for insurance operations? https://www.bcg.com/publications/2020/bionic-insurance-could-be-the-future-of-insurance

BCG (2021a) Digital partnerships – unlocking the $10 billion opportunity in insurance. https://www.bcg.com/digital-partnerships-unlocking-10-bn-opportunity-for-insurance

BCG (2021b) Embracing the human side of the bionic insurer. https://www.bcg.com/publications/2021/how-insurers-can-build-human-bionic-capabilities

BCG (2022) Long live the insurance agency. https://www.bcg.com/publications/2022/long-live-future-of-insurance-agencies

Bewersdorf C (2022) Shaping the future of embedded insurance through partnerships. https://www.swissre.com/risk-knowledge/driving-digital-insurance-solutions/shaping-future-embedded-insurance-through-partnerships.html

Biener C (2019) Artificial conversational agents–changing the norms? I VW Management Information

Biener C, Waeber A (2021) Would I lie to you? How interaction with chatbots induces dishonesty. SSRN working paper

Biener C, Eling M, Lehmann M (2020) Balancing the desire for privacy against the desire to hedge risk. J Econ Behav Organ 180:608–620

Bloomberg (2021) Ant group shuts mutual aid platform xianghubao amid crackdown. https://www.bloomberg.com/news/articles/2021-12-28/ant-group-shuts-mutual-aid-platform-xianghubao-amid-crackdown#xj4y7vzkg

Bohnert A, Fritzsche A, Gregor S (2019) Digital agendas in the insurance industry: The importance of comprehensive approaches. Geneva Papers Risk Insurance Issues Pract 44:1–19
Bower JL, Christensen CM (1995) Disruptive technologies: Catching the wave. Harvard Business Review Januar-February
Braun A, Eling M, Häusle NM, Lehmann M, Schmeiser H, Ungar L (2019) China's tech giants transform healthcare sector with cutting-edge technologies. Credit Suisse Research Institute Report
Braun A, Fleischer M, Ufert A (2020) Digitalisierungsbeschleuniger Corona: Warum Vergleichsplattformen und Ökosysteme gewinnen. Versicherungswirtschaft Heute
Braun A, Haeusle N, Karpischek S (2021) Decentralized autonomous risk transfer on the blockchain. Routledge
Braun A, Haeusle N, Paul T (2023) Risk classification with on-demand insurance. J Risk Insurance (forthcoming). https://doi.org/10.1111/jori.12429
Braun A, Peter W, Xu J (2018) Introduction to bancassurance. Institute of Insurance Economics (University of St Gallen)
Braun A, Schreiber F (2017) The current insurtech landscape: Business models and disruptive potential. Institute of Insurance Economics (University of St Gallen)
Chang VY (2023) Technology investments and firm performance under the wave of insurtech. Geneva Papers Risk Insurance Issues Pract 49:501–536
Che X, Liebenberg A, Xu J (2022) Usage-based insurance – Impact on insurers and potential implications for InsurTech. North Am Actuar J 26(3):428–455
Chen Z, Feng R, Wei L, Zhao J (2022) Cost-effectiveness, fairness and adverse selection in mutual aid. SSRN working paper
Eckert C, Neunsinger C, Osterrieder K (2022) Managing customer satisfaction: digital applications for insurance companies. Geneva Papers Risk Insurance Issues Pract 47(3):569–602
Ehrlich I, Becker GS (1972) Market insurance, self-insurance, and self-protection. J Polit Econ 80(4):623–648
EIOPA (2021) Artificial intelligence governance principles: Towards ethical and trustworthy artificial intelligence in the european insurance sector. Report from EIOPA's Consultative Expert Group on Digital Ethics in insurance
Eling M, Jia R, Lin J, Rothschild C (2022) Technology heterogeneity and market structure. J Risk Insurance 89(2):427–448
Eling M, Kraft M (2020) The impact of telematics on the insurability of risks. J Risk Finance 21(2):77–109
Eling M, Lehmann M (2018) The impact of digitalization on the insurance value chain and the insurability of risks. Geneva Papers Risk Insurance Issues Pract 43(3):359–396
Ellison AB, Bliemer MC, Greaves SP (2015) Evaluating changes in driver behaviour: A risk profiling approach. Accident Anal Prevent 75:298–309
EY (2020) Ecosystems in insurance: What winners do differently. https://www.ey.com/en_nl/financial-services-emeia/ecosystems-in-insurance-what-winners-do-differently
Feng R (2023) Decentralized insurance: Technical foundation of business models
Fleischer M, Braun A, Ufert A, Walthes F (2022) Der vollkommene Markt–ein nahendes Dilemma der Assekuranz? Zeitschrift für die gesamte Versicherungswissenschaft 111(3):217–243
Fritzsch S, Scharner P, Weiß G (2021) Estimating the relation between digitalization and the market value of insurers. J Risk Insurance 88(3):529–567
Gallagher (2023) Gallagher Re global InsurTech report. https://www.ajg.com/gallagherre/-/media/files/gallagher/gallagherre/gallagher-re-insurtech-report-q4-2022.pdf
Gao G, Meng S, Wüthrich MV (2022) What can we learn from telematics car driving data: A survey. Insurance Math Econ 104:185–199
Garth D (2019) The future of insurance in an on-demand world. Coverager
Gassmann O, Frankenberger K, Csik M (2014) The business model navigator: 55 models that will revolutionise your business. Pearson UK
Gemmo I, Browne MJ, Gründl H (2019) Privacy concerns in insurance markets: Implications for market equilibria and social welfare. SSRN working paper
Geyer A, Kremslehner D, Muermann A (2020) Asymmetric information in automobile insurance: Evidence from driving behavior. J Risk Insurance 87(4):969–995
Gomes C, Jin Z, Yang H (2021) Insurance fraud detection with unsupervised deep learning. J Risk Insurance 88(3):591–624
Guillen M, Nielsen JP, Pérez-Marín AM (2021) Near-miss telematics in motor insurance. J Risk Insurance 88(3):569–589
Guillen M, Nielsen JP, Pérez-Marín AM, Elpidorou V (2020) Can automobile insurance telematics predict the risk of near-miss events? North Am Actuar J 24(1):141–152
Haeusle N (2023) The interplay of quality and information in insurance markets. Working paper n/a(n/a)
Hildebrand C, Bergner A (2021) Conversational robo advisors as surrogates of trust: onboarding experience, firm perception, and consumer financial decision making. J Acad Market Sci 49:659–676

Holzapfel J, Peter R, Richter A (2023) Mitigating moral hazard with usage-based insurance. J Risk Insurance (forthcoming). https://doi.org/10.1111/jori.12433
Huang Y, Meng S (2019) Automobile insurance classification ratemaking based on telematics driving data. Decision Support Syst 127:113–156
Kerzner I (2023) Marketing data can reduce AI bias in the insurance industry. Insurance J. https://www.insurancejournal.com/magazines/mag-features/2023/02/20/708144.htm#:~:text=Only%20by%20pursuing%20cutting%20edge,disservice%20to%20themselves%20and%20their
Lanfranchi D, Grassi L (2022) Examining insurance companies' use of technology for innovation. Geneva Papers Risk Insurance Issues Pract 47(3):520–537
Lang N, Schellong D, Sadek P, Wachtmeister A (2022) Driving success in car subscriptions. https://www.bcg.com/publications/2023/the-rise-of-the-car-subscription-market
Li H, Song Q, Su J (2021a) Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. J Risk Insurance 88(3):625–663
Li K, Liu X, Mai F, Zhang T (2021b) The role of corporate culture in bad times: Evidence from the covid-19 pandemic. J Financ Quantitat Anal 56(7):2545–2583
Li L (2021) Opening up the black box: Technological transparency and prevention. J Risk Insurance 88(3):665–693
Li L, Peter R (2021) Should we do more when we know less? The effect of technology risk on optimal effort. J Risk Insurance 88(3):695–725
Luo X, Qin MS, Fang Z, Qu Z (2021) Artificial intelligence coaches for sales agents: Caveats and solutions. J Market 85(2):14–32
Matt C, Hess T, Benlian A, Wiesbock F (2016) Options for formulating a digital transformation strategy. MIS Q Execut 15(2):123–139
McKinsey (2017). Digital disruption in insurance: Cutting through the noise. https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/time%20for%20insurance%20companies%20to%20face%20digital%20reality/digital-disruption-in-insurance.ashx
McKinsey (2021a) Building new digital businesses in insurance. https://www.mckinsey.com/industries/financial-services/our-insights/building-new-digital-businesses-in-insurance
McKinsey (2021b) Insurance 2030—The impact of AI on the future of insurance. https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance
Moore JF (1993) Predators and prey: A new ecology of competition. Harvard Bus Rev 71(3):75–86
Neale FR, Drake PP, Konstantopoulos T (2020) Insurtech and the disruption of the insurance industry. J Insurance Issues 43(2):64–96
Ping An (2022) Company overview. https://www.pagd.net/allPage/aboutUs/47?lang=EN_US
Porter ME (1985) Competitive advantage. Creating and sustaining superior performance. Free Press, New York
Rahlfs C, Döring U (2007) Redefinition der Wertschöpfungskette von Versicherungsunternehmen. Springer
Reuters (2023) China's Ping An insurance posts 48.9% rise in Q1 profit. News report
Scism L (2023) Regulators look at potential bias in life-insurance algorithms. Wall Street J
Shavell S (1979) On moral hazard and insurance. Q J Econ 93(4):541–562
Smith I, Hale T (2022) Ping an profits tumble due to covid and property woes. Financial Times
Smith LT, Pirchalski E, Golbin I (2022) Avoiding unfair bias in insurance applications of ai models. SOA Research Institute. https://www.soa.org/4a36e6/globalassets/assets/files/resources/research-report/2022/avoid-unfair-bias-ai.pdf
Sosa I, Montes Ó (2022) Understanding the insurtech dynamics in the transformation of the insurance sector. Risk Manag Insurance Rev 25(1):35–68
Stadelmann C, Ueli K, Hostettler L (2023) Digital transformation in the insurance industry: The internet of things as a main driver. https://www.bearingpoint.com/en-ch/insights-events/insights/digital-transformation-in-the-insurance-industry-the-internet-of-things-as-a-main-driver/
Standaert W, Muylle S (2022) Framework for open insurance strategy: Insights from a European study. Geneva Papers Risk Insurance Issues Pract 47(3):643–668
Statista (2022a) Share of consumers who could imagine getting all their insurances from an internet company like google or apple globally in 2022, by country. https://www.statista.com/statistics/1332136/share-consumers-willing-to-get-insurance-from-an-internet-company/
Statista (2022b) Share of consumers willing to consult a robo-advisor for insurance advice globally in 2022. https://www.statista.com/statistics/1332131/share-consumers-willing-to-consult-robo-advisor-for-insurance-advice-globally/
Statista (2022c) Share of consumers willing to manage insurances exclusively online globally in 2022. https://www.statista.com/statistics/1332139/share-consumers-willing-to-manage-insurances-exclusively-online-globally/
Statista (2022d) Share of insurance customers who would take out a policy from non-insurance companies in the united kingdom in 2022. https://www.statista.com/statistics/1333793/embedded-insurance-comfort-customers-uk/
Stice S, Besio B (2022) Embedded insurance: a brief overview. https://insuranceblog.accenture.com/embedded-insurance-brief-overview

Stipancic J, Miranda-Moreno L, Saunier N (2018) Vehicle manoeuvers as surrogate safety measures: Extracting data from the GPS-enabled smartphones of regular drivers. Accident Anal Prevent 115:160–169
Stoeckli E, Dremel C, Uebernickel F (2018) Exploring characteristics and transformational capabilities of InsurTech innovations to understand insurance value creation in a digital world. Electron Markets 28:287–305
Swiss Insurance Association (2021) Skills der Zukunft für die Assekuranz. Industry report
Swiss Re (2017) Unveiling the full potential of telematics. Industry report
Swiss Re (2020) Why insurers need to transform digital distribution and how to do it in the digital age. Industry report
Teece DJ (2010) Business models, business strategy and innovation. Long Range Plann 43(2-3):172–194
The Geneva Association (2023) The value of insurance in a changing risk landscape, Research summary November 2023 by Kai-Uwe Schanz. https://www.genevaassociation.org/sites/default/files/2023-11/value_of_insurance_summary_web.pdf. The Geneva Association, Zurich, Switzerland
The Straits Times (2022) How china's healthcare mutual aid industry folded in just 3 years. https://www.straitstimes.com/asia/east-asia/how-chinas-health-care-mutual-aid-industry-folded-in-just-3-years
Tumminello M, Consiglio A, Vassallo P, Cesari R, Farabullini F (2023) Insurance fraud detection: a statistically validated network approach. J Risk Insur 90:381–419. https://doi.org/10.1111/jori.12415
Weidner W, Transchel FWG (2015) Aktuarielle Besonderheiten bei der Kalkulation von Telematik-Tarifen in der Kfz-Versicherung. Zeitschrift für die gesamte Versicherungswissenschaft 104:595–614
Weidner W, Transchel FWG, Weidner R (2016) Classification of scale-sensitive telematic observables for risk-individual pricing. Eur Actuar J 6:3–24
Weill P, Woerner SL (2015) Thriving in an increasingly digital ecosystem. MIT Sloan Manag Rev
Zeier Röschmann A, Erny M, Wagner J (2022) On the (future) role of on-demand insurance: Market landscape, business model and customer perception. Geneva Papers Risk Insurance Issues Pract 47(3):603–642
Zhu F, Woo AK, Dai NH (2022) Ping An: Pioneering the new model of technology-driven finance. https://www.hbs.edu/faculty/Pages/item.aspx?num=57837



Pandemic Risk Management and Insurance

Gunther Kraut, Paulina La Bonté, and Andreas Richter

1 Introduction

The recent COVID-19 pandemic had far-reaching effects around the world, leading to a high death toll and severely impacting economies. Government measures such as lockdowns that were taken to contain the spread of the virus negatively affected businesses, in particular small- and medium-sized enterprises. While the COVID-19 pandemic took the general population by surprise, it did not come unexpectedly to the insurance industry, which is exposed to epidemic and pandemic risk in multiple lines of business. Since disease outbreaks that entailed a risk of turning into a pandemic had already been occurring regularly pre-COVID-19, insurance companies included these scenarios in their risk modeling, in particular in the life insurance lines where in many markets this is even a regulatory requirement (see Kraut and Richter 2015). However, some major developments were not correctly anticipated, including the magnitude of business interruption claims and the uncertainty in the legislation surrounding these claims. COVID-19 underlined the challenges that pandemic risk represents to both the insurance industry and governments and sparked debates about the role of insurance. Pandemic insurance can potentially not only protect from economic consequences by compensating for losses but also contribute to risk mitigation and management. Nevertheless, the insurance industry's capacity for taking on pandemic risk is limited. Since future disease outbreaks with pandemic potential are inevitable, further investigating pandemic risk and possible risk management strategies is crucial for both insurance companies and society as a whole.

G. Kraut
Munich Re, Munich, Germany
e-mail: gkraut@munichre.com

P. La Bonté · A. Richter (✉)
Ludwig-Maximilians-Universität München (LMU Munich), Munich, Germany
e-mail: p.bonte@lmu.de; richter@lmu.de

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_9

Literature on pandemic risk management and insurance was already available pre-COVID-19, discussing previous epidemic and pandemic outbreaks, possible scenarios, and their impact on the insurance industry (see, e.g., European Actuarial Consultative Group 2006; SOA 2007; Keogh-Brown and Smith 2008; World Health Organization 2017). Since the COVID-19 outbreak turned into a global pandemic, a multitude of new articles and reports on pandemic insurance have been published. The purpose of this chapter is to provide an introduction to pandemic risk management and insurance, with a focus on risk transfer solutions and an emphasis on recent developments, and to derive implications for a path forward for pandemic risk management and insurance. This chapter is organized as follows: Sect. 2 presents an introduction to epidemic and pandemic risk, addressing both previous outbreaks and the COVID-19 pandemic in terms of fatalities and economic losses. Furthermore, predictions concerning the frequency and economic impact of future outbreaks are considered. Section 3 reviews the insurability of pandemic risk, before Sect. 4 describes the impact of the COVID-19 pandemic on the insurance industry, discussing the effect on different lines of business and the impacts via the financial markets. This section additionally illustrates examples of insurance market developments in the aftermath of COVID-19. Next, Sect. 5 explores pandemic risk management solutions and provides suggestions on how insurance companies, capital markets, and the public sector can interact in order to overcome the limits of insurability and prepare for the next pandemic. While we briefly address other risk management strategies related to public preparedness and corporate risk management, we focus on risk transfer mechanisms. Section 6 concludes.

2 Epidemic and Pandemic Outbreaks

2.1 Classification of Outbreaks

While the term *epidemic* refers to "the occurrence in a community or region of cases of an illness, specific health-related behavior, or other health-related events clearly in excess of normal expectancy" (Porta 2014, p. 93), a *pandemic* is "an epidemic occurring over a very wide area, crossing international boundaries, and usually affecting a large number of people" (Porta 2014, p. 209). This does, however, not provide a precise definition of the conditions that have to be fulfilled for an epidemic to be considered a pandemic.

In order to classify outbreaks, the WHO has designed several instruments that serve as warning systems, indicating the need for preparedness and mitigation measures. These instruments include the *WHO pandemic phases*, the *WHO Disease Outbreak News* (DONs), the declaration of a *Public Health Emergency of International Concern* (PHEIC), and the declaration of a pandemic made by the Director-General. The WHO pandemic phases (World Health Organization 2017) serve as a tool for global risk assessment for influenza pandemics and comprise four stages that detail the spread of a new influenza subtype with pandemic potential that infects humans. Following an *interpandemic phase*, the *alert phase* describes the identification of a new influenza subtype in humans. The *pandemic phase* is reached when this subtype spreads globally, while risk reduces again in the *transition phase*. These phases provide a framework for the definition of stages of risk management actions, including preparedness, response, and recovery, and are also incorporated in Disease Outbreak News, of which there are about 200 issues per year (Kraut and de Kuiper 2021). According to the World Health Organization (2023b), DONs provide "information on confirmed acute public health events or potential events of concern." In addition, serious events constituting public health risks on an international scale are declared a Public Health Emergency of International Concern. This declaration is the highest alert level under the *International Health Regulations* (IHRs), following a defined process of assessing whether the event is unusual or unexpected, has a large

public health impact, and contains a high risk for international spread and international travel or trade restrictions (Wilder-Smith and Osman 2020). The determination of a PHEIC signals the necessity of measures to contain the spread of the disease, includes temporary recommendations for actions, and may also legally require the WHO member countries to react (World Health Organization 2017). The 2002-2004 SARS coronavirus outbreak was the triggering event leading to the definition of the PHEIC system. Since the term was then introduced in 2005, there have been eight PHEICs, with the 2024 Mpox outbreak being the most recent announcement (World Health Organization 2024). Serving as another instrument that might entail regulatory actions by the WHO and member countries, if deemed appropriate, the WHO Director-General can also make a declaration of a pandemic (World Health Organization 2017).

Regarding the potential cause of a pandemic, it should be noted that there is a large variety of pathogens and microbial agents. Their characteristics will influence their base risk of causing a pandemic, the prospects of mitigating or even containing the pandemic, and the choice of prevention and control measures. Two key characteristics of pathogens contributing to their epidemic spread are their *transmissibility* and their *virulence*. These two dimensions typically form the basis for modeling epidemic spread with compartmental epidemiological models such as *SIR models* (where SIR stands for Susceptible, Infectious, or Recovered).[1] The most relevant underlying drivers influencing these factors include the type of pathogen and its pathogen-specific characteristics.[2] While Adalja et al. (2022) state that, given the right context, any microbial organism could be a global catastrophic biological risk, they also specify that the most likely manifestation will be a virus. For specific virus categories of particular concern, see Table 1. Contrasting viruses, it will be much less likely that, e.g., bacteria cause pandemic spread, as antibiotics are available as an immediate treatment.

Pathogen-specific characteristics such as the mode of transmission will also influence which response and containment measures may be effective, thus determining the pandemic threat posed by a pathogen.[3] Gavi (2021) provides summaries of pandemic threat analyses for several viruses. Holmes (2022) raises the additional question of whether the viruses that are most likely to emerge in humans can be identified. Although metagenomic sequencing reveals an increasingly large virosphere, i.e., more and more viruses are known, he notes mammals are carrying many thousands of different viruses, most of which remain undocumented.

2.2 A Quick Overview of Important (Recent) Outbreaks

Several notable epidemic and pandemic outbreaks have occurred since the previous century. These outbreaks include the Spanish flu influenza pandemic (1918–1920), the Asian flu (1957–1958), the Swine flu pandemic (2009), SARS (2002–2004), MERS (2012–present/2015), and Ebola (2014–2016/ 2018–2020). For a more detailed overview of noteworthy recent outbreaks, see Fig. 1. This figure shows a list of selected large epidemics (with at least 40 fatalities) caused by infectious diseases that are on or have been on the *WHO priority list*[4] and have been identified as having a high PHEIC risk. For yellow fever and dengue, the years with record epidemics/resurgence are shown. The 2002-

[1] See RMS (2021) for an example of pathogen characteristics being used in risk modeling.

[2] Relevant pathogen-specific characteristics that influence the development of an outbreak include the transmission mode of pathogens, the timing of contagiosity of a disease compared to the onset of symptoms, and the preexisting immunity of the population to a pathogen (see Adalja et al. 2022).

[3] See Sect. 2.4 for further details.

[4] This list includes pathogens with the greatest public health risk due to their epidemic or pandemic potential, and for which there is no, or insufficient, treatment or vaccines (see World Health Organization 2023e).

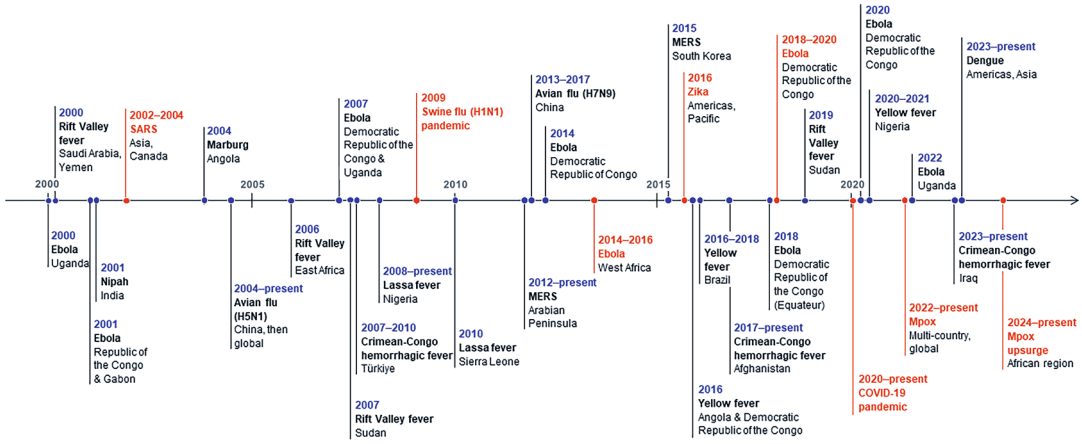


Fig. 1 Overview of noteworthy recent outbreaks of WHO priority diseases, including both PHEIC declarations and "near-misses" (own illustration, based on Munich Re 2023b)

2004 SARS epidemic and outbreaks declared PHEICs are marked in orange, while other outbreaks are marked in blue. Years indicate the start of the outbreak and when it was declared over. For endemic diseases such as Lassa fever, years indicate peak years. Infection and mortality rates and the economic impact of each of the depicted outbreaks differ significantly.

2.3 COVID-19

The COVID-19 outbreak was declared a PHEIC by the WHO on January 30, 2020 and characterized as a global pandemic outbreak on March 11, 2020 (World Health Organization 2020b; World Health Organization 2020c). The disease is caused by the *SARS-CoV-2* pathogen, which is a virus that stems from the coronavirus family (European Centre for Disease Prevention and Control 2023). By May 2023, about 760 million cases and 6.8 million deaths were reported to the WHO (World Health Organization 2023d). However, due to underreporting, excess deaths associated with the COVID-19 pandemic (including deaths that were indirectly caused by the pandemic, e.g., due to disruptions to health-care services, see World Health Organization 2023f) are assumed to be substantially higher: The WHO announced that the death toll had most likely exceeded 20 million by May 2023 (Fatunmole 2023)—even with strong containment measures being in place. Comparing COVID-19 to the two previous coronavirus outbreaks, it is significantly more widespread but characterized by a lower case fatality rate than both the SARS epidemic, which killed approximately 10% of the 8.219 cases, and the MERS epidemic, killing around 36% of approximately 2.600 cases since first reported (World Health Organization 2015; World Health Organization 2023c). In comparison to this, the Spanish flu is estimated to have infected about 500 million people, which was a third of the world's population at that time (Centers for Disease Control and Prevention 2018). There is, however, severe uncertainty about the number of deaths, with estimates ranging from 20 to 100 million (Madhav et al. 2017, Chap. 17). The most recent global pandemic before the COVID-19 outbreak, the 2009 influenza pandemic, was highly transmissible (with an estimated global infection rate of 11%-21% according to Kelly et al. 2011) but exhibited a lower case fatality rate (see, e.g., Da Costa et al. 2020 for an epidemiological comparison between the two diseases).

Estimates by the IMF suggest that the cumulative economic losses from the COVID-19 pandemic will reach an amount of 13.8 trillion USD by the end of 2024. The pandemic is estimated to have caused global real GDP to drop by about 6% in 2020 and by about 4% in 2021 (Agarwal et al. 2022). Comparing this number to the economic impact of the Spanish flu (Barro et al. 2020 estimate that the losses amount to about 6% of global real GDP cumulated over the period from 1918 to 1920) and the SARS epidemic (40 billion USD total losses according to Lee and McKibbin 2004, with percentage real GDP decreases in 2003 being significantly lower than 1% for all analyzed regions except mainland China and Hong Kong) further illustrates the proportions of the COVID-19 crisis.

2.4 Future Pandemics: Estimates and Scenarios

Considerations about future epidemic and pandemic risk must take into account that outbreaks are increasing in both frequency and severity, with multiple factors influencing the likelihood that novel viruses jump from animals to humans. LMU (2023) describes different contributing aspects: Population growth causes humans to settle in new areas, which increases contact between people and wild animals and exposes humans to viruses which they had previously not come into contact with. Destabilized animal populations and intensive farming practices lead to a higher virus density within animal populations and increase the occurrence of mutations. Globalization makes large companies contribute to this phenomenon by mining resources and clearing woodland in sensitive areas. Additionally, high mobility of people enables a rapid spread of viruses. Another contributing factor is global warming, leading transmitters of infectious diseases to seek out new habitats.

Despite mitigating factors such as a better understanding of microbial agents and improved outbreak response measures, recent estimates show that extreme outbreaks such as COVID-19 will not be rare events in the future. According to risk modeler Metabiota, the likelihood of experiencing a pandemic that is at least of the same magnitude as the recent COVID-19 pandemic is between 2.5% and 3.3% each year and ranges from 22% to 28% in the next 10 years and from 47% to 57% in the next 25 years (Cheney 2021; Smitham and Glassman 2021). Similar estimates are found by Marani et al. (2021), who suggest a yearly probability of 2.63%, and by Airfinity (2023), who provide a probability of 27.5% for the next 10 years. Another aspect of concern is that low-income countries are estimated to face the highest risk for spillover events from wildlife to humans, while these countries are also least prepared when it comes to managing and containing outbreaks (Smitham and Glassman 2021).

It is important to note that there is no reason why the next pandemic could not be caused by a pathogen with much more severe characteristics than the SARS-CoV-2 virus (see, e.g., Reuters Staff 2021; Smitham and Glassman 2021; Weintraub 2023). Equally, it could be milder, such as the 2009 Swine flu influenza pandemic, whose infection and case fatality rates were discussed in Sect. 2.3. Estimates for future deaths resulting from different classes of infectious diseases underline the potential severity of outbreaks. Metabiota finds 2,200,200 potential future deaths per decade from pandemic influenza (excluding seasonal flu), 1,500,000 potential deaths from coronaviruses, and 300,000 deaths from viral hemorrhagic fevers. In comparison, future deaths for non-disease natural catastrophes (including floods, storms, droughts, and earthquakes) are estimated at 84,200 per decade (Stefan et al. 2023b).

Understanding the interplay of epidemiological factors with available countermeasures and response options is crucial for understanding both the public health and economic consequences of an outbreak caused by a specific pathogen. Table 1 outlines scenarios for pathogens which can be

Table 1 Selected pathogens that can be considered "candidates" for causing the next pandemic and their countermeasures and plausible response options

	Mild scenario	Moderate scenario	Severe scenario
Likely infectious	• Pandemic influenza	• Novel coronavirus	• Novel paramyxoviruses
Most at risk for severe disease	• Any age group. Unlike seasonal flu, pandemic flu does not always particularly affect the old and chronically ill	• Infections with SARS, MERS, and COVID-19 have been more severe in older people and those with preexisting health conditions	• Any age group would be at risk
	• A high mortality was seen in healthy individuals and 20-40 year olds during the 1918 Spanish flu		• Particularly severe consequences are to be expected if a new virus emerged that was as contagious as measles and as deadly as Nipah
Available countermeasures	• Influenza vaccines exist but require adapting	• No vaccine or specific treatment is currently available	• No vaccine or specific treatment is currently available
	• It will take at least 5-6 months before an approved pandemic flu vaccine can be rolled out	• It took around 1 year for the SARS-CoV-2 vaccine and specific treatments to be licensed	• New vaccines and treatments would have to be developed, which is unlikely to occur within less than a year
	• Flu treatments such as antivirals exist; however, the supply may not be adequate to meet early stage demand		
Use of civil authority restrictions	• High probability of limited restrictions in early stages of outbreak in affected areas and likely advice to avoid traveling in these areas for nonessential reasons	• High probability of limited restrictions in early stages of outbreak in affected areas and likely advice to avoid traveling in these areas for nonessential reasons	• High probability of limited restrictions in early stages of outbreak in affected areas and likely advice to avoid traveling in these areas for nonessential reasons
	• Low probability of widespread lockdowns	• Moderate to high probability of widespread lockdowns	• High probability of widespread lockdowns if the disease is severe
	• If a severe pandemic influenza such as the 1918 Spanish flu occurs, the probability of some restrictions will be higher	• Reduced probability if pan-coronavirus vaccines are developed. This, however, may not happen for many years	

considered candidates for causing a fast-spreading global pandemic, with available countermeasures and plausible response options (which will be addressed again in Sect. 5.1).[5]

The economic impact of COVID-19 was devastating and has been widely analyzed (see also Sect. 2.3). For estimates of the economic impact of other epidemic outbreaks, see, e.g., Lee and McKibbin (2004), Keogh-Brown and Smith (2008), World Bank (2014), Ippolito et al. (2015), United Nations Development Programme (2017), and Barro et al. (2020). Note that significant economic damage is caused not only by global pandemics but also by regional epidemics—the frequency of which has been increasing significantly (see Rohr et al. 2019). Obviously, a pathogen economic impact is not only influenced by its initial transmissibility and virulence. Other characteristics play an

[5] These examples are not understood as a prediction. Pathogens causing smaller regional epidemics may have similar characteristics and countermeasures and may be successfully contained at a regional level. Other pathogens may be so mild that they do not require specific interventions. Additionally, different modes of transmission require a different type of intervention, as, for example, in the case of sexually transmitted diseases or diseases transmitted via vectors.

important role as well. For example, it is expected that a novel pathogen is feared more than a known pathogen, in particular if it is combined with certain types of severe, visible, or painful symptoms, and might thus yield stronger behavioral responses which cause economic losses. An attempt to quantify such effects has led to the development of Metabiota's *Pathogen Sentiment Index* (see Metabiota 2018).

Annual expected losses in economic output resulting from pandemics are estimated to lie at about 176 billion USD according to Glennerster et al. (2022). This figure not only emphasizes the need for investments in risk mitigation and preparedness to reduce the size of losses and the number of deaths but also explains why affordability is a topic of consideration for risk transfer solutions for pandemic risk, both of which will be addressed again in Sect. 5. Naturally, the chosen response will influence the economic impact, in particular because responding to an outbreak typically reduces the pathogen's transmissibility and virulence. As an example, it could be clearly observed that the measures imposed to contain the COVID-19 pandemic did, as a side effect, very effectively reduce the spread of other infectious diseases such as seasonal influenza (for statistics on the number of laboratory-confirmed influenza cases during the COVID-19 pandemic, see, e.g., World Health Organization 2019; World Health Organization 2020a; Robert Koch-Institut 2021; World Health Organization 2023a). In analyzing the potential economic impact of epidemic and pandemic events, it is important to consider first-order and second-order effects. If an outbreak is contained early on through strict response and containment measures (which we will address in more detail in Sect. 5.1), it may appear that the economic impact is mostly caused by the response to the outbreak.[6] If, however, the outbreak is not contained and the death toll increases, fear and voluntary social distancing may well lead to worse economic consequences—combined with a higher death toll. As the response influences the outbreak itself, a precise evaluation for a specific outbreak remains challenging.

3 Pandemic Risk and the Limits of Insurability

Catastrophic losses, and in particular situations where losses significantly exceed existing insurance coverage, as was the case with COVID-19, tend to trigger discussions about the insurability of such risks. This can additionally be motivated by (even worse) extreme scenarios and considerations regarding capacity limits of the (re)insurance industry (see, e.g., Cummins et al. 2002).

Once the limitations of insurability have been identified, product design can be used to reduce or even exclude insurability problems of the underlying risk. In addition to the properties of the risk itself, the question of whether a specific type of exposure can be underwritten is driven by various characteristics of the market and its players: Among other factors, a potential carrier's existing book of business, organizational form and ownership structure, risk management (in particular, reinsurance) options, and regulatory constraints, as well as market structure and the degree of competition, define the framework within which underwriting decisions are made. The limits of insurability can be moved substantially by the right setup and, in particular, by the proper division of responsibilities between public and private institutions. We will get back to this line of thought in Sect. 5 which addresses the question of how the capacity for pandemic risk should be provided. In this section, we will primarily focus on the characteristics of the risk itself. We will identify features of pandemic risk that make it particularly difficult to insure.

Insurability analysis typically checks a set of criteria that a risk should ideally fulfill if it is to be covered in private insurance markets (see, for instance, Karten 1972; Berliner 1985; Karten 1997;

[6] Or the economic damage spreads differently over time, as response measures may also influence the timeline of an outbreak.

Harrington and Niehaus 2003; Karten et al. 2018 and for recent work specifically on the insurability of pandemic risk Hartwig et al. 2020; Klein and Weston 2020; Richter and Wilson 2020; Schanz 2020; Gründl et al. 2021). These criteria classify technical or economic limitations that constrain the efficiency of insurance markets for the specific type of risk. Following the spirit of this literature, we look for potential insurability problems resulting from correlation and/or the size of potential losses, information asymmetries (in particular, moral hazard), insufficient knowledge about the loss distribution, and difficulties in precisely defining the insured risk.

Significant correlation prevents the law of large numbers from working. The positive effects of pooling cannot materialize, and the risk of catastrophic cumulative losses increases. Clearly, correlation is a key issue in insuring epidemic and especially pandemic risk. Similar to natural or man-made disasters, an epidemic typically affects a large number of people and businesses simultaneously and, thus, potentially causes large total amounts of insured losses. To address this issue, insurance markets rely on regional diversification strategies that include global underwriting, reinsurance, and insurance-linked securities.[7]

The aforementioned strategies help diversify locally correlated exposures. Pandemics, however, are defined by their global nature, which creates a systemic risk (Hartwig et al. 2020). The rapid spread of COVID-19 illustrates how large parts of the world can be affected almost simultaneously. Hence, regional diversification will be less effective in managing this risk. Because of the large potential for loss accumulation, Richter and Wilson (2020) therefore consider the correlation problem one of the two central impediments to the insurability of pandemic risk (along with the problem of external moral hazard, see below). Note that similar problems can arise in the context of cyber risks, where the interconnected nature of computer systems allows malware to spread around the world in a very short time frame (see, for instance, Munich Re 2017; Swiss Re 2017). Gründl et al. (2021) analyze hypothetical pandemic insurance contracts and find that the large associated *accumulation risk*[8] results in significantly higher expected shortfalls and thus in an increase in insurance prices, which reduces the affordability of these contracts.

Where the effects of regional diversification are limited, more capital will be required to back insurance coverage for accumulation risk.[9] Another tool to manage accumulation risk can be pooling over time, if losses over time are independent, i.e., losses in 1 year are stochastically independent of losses in another year. Gründl et al. (2021) analyze intertemporal risk-sharing, assuming that insurers participate in a pooling arrangement over 50 years, and find that the expected shortfall that insurers face decreases significantly. However, several problems make it difficult to create viable arrangements for risk pooling over time. In particular, it would be difficult to implement a system that makes a sufficient capital buffer available each year. Large losses might occur in an early year before the necessary buffer has been accumulated (Hartwig et al. 2020). Also, tax reasons and capital costs can make building up the required capital buffer rather expensive. Hartwig et al. (2020) refer to specific adverse selection and moral hazard problems associated with (private) arrangements for pooling over time and discuss whether government solutions are superior to private ones. Similar arguments concerning challenges in designing risk-sharing schemes that preserve incentives of both governments for preparedness and insurers for risk-based pricing are also provided by Gründl et al. (2021).

[7] For an analysis how risk pooling strategies for illiquid catastrophe risk markets among several stakeholders can be implemented in a way that limits adverse selection and moral hazard, see Kraut (2014) and Kraut (2015).

[8] Accumulation risk refers to highly correlated potential claims (Gründl et al. 2021), with losses occurring due to one single event or peril.

[9] For an analysis of the efficient use of collateral in certain specialty risk markets, see Lakdawalla and Zanjani (2012).

While highly correlated claims in life and health insurance have always been considered a defining characteristic or a direct potential consequence of a pandemic, the COVID-19 crisis has also highlighted the accumulation potential of financial market losses and of property and casualty losses.

The low interest rates caused by the COVID-19 pandemic increased insurers' liabilities due to lower discount rates, while at the same time on the asset side negative market moves were driven by lower equity values (due to lower expected earnings), lower investment returns, and credit risk. This demonstrates that financial market developments, correlated with pandemic risk, can result in lower asset values and higher economic liability valuations (especially problematic for insurers with long-term retirement businesses) and therefore adversely impact solvency ratios.[10] (Sect. 4.4 provides a more detailed picture of the financial impact.) Also, insurance company investors might be reluctant to support pandemic risk coverage, if it is positively correlated with their investment portfolio (Richter and Wilson 2020; Qiu 2020).

Although property and casualty losses caused by a pandemic, in particular non-property damage business interruption and event cancelation, had been taken into account by pre-COVID-19 scenario analysis, at the onset of the crisis the insurance industry did not expect severe accumulations of insured losses to result from these lines. Scenarios did not fully anticipate the broad impact of governments' pandemic containment activities. Also, in the industry's initial estimates, the assumption was that non-property damage business interruption losses were not covered under standard property business interruption policies, and for coverage that was specifically designed to include non-property business interruption there was only a small market (Ralph 2020). The underestimated accumulation potential, combined with legal uncertainty and issues with imprecise wording, caused an unpleasant "surprise" for the industry (Richter and Wilson 2020).

The consequences of pandemic lockdowns in many cases led to discussions about whether there was actually coverage to be found under policies that the industry had thought of as being immune to pandemic-related business interruption claims (Ralph 2020). Although unambiguous and precise wording of the contract is a straightforward required property of an insurable risk (as well as for any functioning financial product), COVID-19 demonstrated that in practice this requirement is often challenged, especially when new risks or unprecedented events are concerned. See Sect. 4.1 for further discussions on COVID-19 and business interruption insurance and the litigation around policy wording. Unintended exposure to the consequences of a general lockdown under business interruption insurance can be compared to the problem of so-called silent cyber risk (see Guy Carpenter 2018) and also bears similarities with emerging threats and unexpected court rulings in the context of liability insurance (an example is the well-known Jackson Township case; see, e.g., Kunreuther 1987).

The fact that there are many examples of imprecise wording concerning the question of whether specifically the risk of lockdown-related business interruptions was covered suggests that these scenarios were underestimated or ignored by the insurance industry. This also demonstrates the challenges of adequate quantification and pricing of such risk. The false perception of the potentially catastrophic consequences of a pandemic resulted in shortcomings in pandemic risk modeling for property and casualty insurance. Qiu (2020) emphasizes that pandemic models before COVID-19 often focused on numbers of fatalities and casualties and were not built for property and casualty products. The use of standard epidemic risk models, including established biomathematical tools, enables insurers to assess the frequency and severity of a pandemic event in terms of, e.g., excess

[10] Richter and Wilson (2020) point out that during times of crisis, management actions can only materially reduce solvency capital requirements but cannot replace lost solvency capital, as this takes a long time and much effort to generate and retain. This, according to the authors, underlines the importance of building up initial resilience versus contingency plans as part of an insurer's crisis risk management.

mortality or disease spread. In a second step, this information must be combined with underwriting expertise and different sources of financial or economic loss data from former epidemics to model the economic impact of pandemic outbreaks (see Kraut and de Kuiper 2021). Qiu (2020) compares the impact the COVID-19 crisis had on the industries' risk awareness to the impact of the 9-11 attack, in the sense that "[t]he sector suddenly realized a material risk and significant loss potential that was never predicted before" (Qiu 2020, p. 7). He expects that similar to terrorism risk, post-COVID-19 pandemic risk modeling and pricing will significantly improve and become more robust and particularly highlights the importance of correlation as a core element of pandemic risk models. At this point, though, there is still significant uncertainty associated with modeling the frequency and severity of a pandemic, in particular because specific risk modeling expertise is not yet widespread among market participants.

Contrasting most natural and man-made catastrophes, pandemic losses materialize over time. During this period, governmental and personal decisions are made to contain or suppress the pandemic. With respect to the insurability of pandemic risk, the key question is whether insurance coverage affects these decisions, or in other words, whether moral hazard can be an issue. Obviously, where policy triggers involve the extent of containment measures in place, attention must be paid to the threat of *external moral hazard*[11] that can result from political decision-makers' incentives to look for deep pockets (Richter and Wilson 2020): Decisions about containment measures might be influenced by the degree to which affected businesses are insured. Even worse, there is a risk that there would be pressure on the insurance industry to cover losses even if they are excluded or to rewrite contracts to cover business interruption retroactively. This latter problem is of course more likely where public perception is that policy wording has been ambiguous.

The COVID-19 crisis showed how social pressure on the industry to pay claims can become a major factor in determining the potential losses from the pandemic, even if such payments require a retroactive change in terms and conditions. As an example, in the USA this pressure began in social media with tweets from US President Trump in early March 2020 suggesting that business interruption should, in any case, be covered by insurers (Richter and Wilson 2020) and continued with proposed legislation.

In addition to the issue of external moral hazard, incentive alignment problems can of course also occur between the insured and the insurer: In event cancelation insurance, for instance, the insured typically has to decide whether or not to cancel an event. As discussed, e.g., in Turner (2020) and Clarey (2020), the 2020 Wimbledon tennis championships were canceled, while the French Open was merely postponed—a difference that may in theory well have been influenced by the fact that only the former had cancelation coverage.[12]

The preceding paragraphs focused on ex post moral hazard, as decisions that are made after a pandemic has started were considered. In the future, it will also be crucial to keep an eye on the effect that the availability of protection will have on businesses' efforts in preparing for a pandemic, given the increased awareness of pandemic risk and potentially higher demand for explicit coverage for property and casualty consequences. On the other hand, with a growing market for pandemic property and casualty risk, the industry is likely to limit coverage due to the accumulation potential, which would mitigate such ex ante moral hazard problems (see Sect. 4.5 for possible market developments).

Naturally, standard tools such as coinsurance, deductibles, waiting periods, etc. will also be used to contain these moral hazard problems. The abovementioned example from the event cancelation

[11] The term external moral hazard refers to insurance-induced changes in the behavior of third parties which affect the size of insurance payments (see Nell 1993; Nell et al. 2009).

[12] For the specific case of event cancelation insurance, product design could support a stronger alignment of interests in case of pandemic events by, e.g., also providing cover for a tournament taking place with reduced attendance.

market segment illustrates the importance of one product design aspect in particular: policy triggers that are beyond the insured's control. Defining triggers that, e.g., refer to parameters of a pandemic, will not be trivial but should become easier as in the aftermath of the COVID-19 crisis public response mechanisms and information collection will be reviewed and improved. As was emphasized above, however, it must be kept in mind that attention is paid not just to the insured's potential influence on the loss distribution, but also, and even more importantly, to the fact that triggers involving the extent of containment measures in place carry the severe threat of external moral hazard. As much as these triggers are transparent and easy to use and although they will typically provide a good match with actual losses, there is a significant downside that limits their applicability. The design of triggers will again be addressed in Sect. 5.

4 The Impact of COVID-19 on the Insurance Industry

Initial estimates for total insured losses from COVID-19 varied widely. While Willis Towers Watson (2020) put them at 38 billion USD in a moderate scenario, estimates at the upper end were higher than 100 billion USD. Actual insured losses resulted in 44 billion USD by the beginning of 2022, representing the third-largest catastrophe-related amount after Hurricane Katrina in 2005 and the 9-11 terrorist attacks (Cohn 2022). However, the impact of the COVID-19 pandemic on the insurance industry and the extent to which insured losses match predictions vary greatly with different lines of business.

4.1 Property and Casualty Insurance

Government measures that were taken to contain the spread of the virus led to large income losses for businesses. Stay-at-home orders, supply chain failures, business closures, and restrictions, as well as reduced demand, have particularly affected small- and medium-sized enterprises. In addition, events had to be postponed or suspended, severely affecting the event industry. In their moderate scenario, Willis Towers Watson (2020) estimated 49% of insured losses to result from business interruption claims. The actual impact of these claims is however still to be determined.

Epidemic and pandemic insurance covering event cancelation and business interruption losses was available pre-COVID-19 only to a limited extent. As discussed in Sect. 3, insurers face a large accumulation risk, and decisions by governments are hard to predict, which makes modeling and pricing very challenging. Thus, according to Hartwig et al. (2020), most event cancelation policies contained exclusions for communicable diseases that could in some cases be removed by paying additional premiums. Demand for these removals was, however, rather low, with publicly known examples including the abovementioned policy purchased by the organizers of the Wimbledon Tennis Tournament. Insurers denied coverage for event cancelation losses resulting from COVID-19 for policies that did not contain the add-on, either because such losses were explicitly excluded or arguing that the cause of the loss was the pandemic itself and not government orders (Aviles and DeField 2022).

Regular business interruption policies typically cover losses sustained during an interruption caused by a direct physical loss or property damage. Richter and Wilson (2020) argue that most insurers did not view pandemic-related losses as covered under these policies. In addition, many policies even specifically excluded pandemic exposures. On the other hand, demand for explicit pandemic coverage or add-ons to existing policies was low (Glynn et al. 2021, Chap. 14). Offered products have either been viewed as too expensive (Ziady 2020) or businesses were not aware

of pandemic risk, which left them unprepared. The most widely discussed example of pandemic insurance products, *PathogenRx*, was introduced to the US market in 2018 by Marsh LLC in cooperation with reinsurer Munich Re and risk modeler Metabiota. The policy was indemnity-based and used triggers such as fatalities within predefined areas to determine the severity of a covered epidemic or pandemic event in order to provide coverage for the loss of revenue, loss of gross profits, and extra expenses following the spread of an outbreak in a certain area. In addition, the product also offered companies the opportunity to model their infectious disease risk (see Goyder and Smythe 2020 for a comprehensive description of the policy). Although hardly any such pandemic business interruption policy was sold pre-COVID, an increasing demand was reported during the recent pandemic (Ziady 2020). Product design for epidemic and pandemic risk insurance has gradually evolved to become easier to understand and apply for the policyholder,[13] and some new products have been made available for different market segments such as business interruption (see, e.g., Sompo International 2021; Munich Re 2023a; Munich Re 2023c) and contingency (see, e.g., Dieleman 2022). However, a strongly relevant factor for the market success of pandemic risk insurance products has been the dependency on the ongoing COVID-19 pandemic. Adhering to the principle that an insurance policy should not cover events that are already ongoing at policy inception, the inability to cover the next wave of the ongoing pandemic event has been a significant sales hurdle. Early adopters of pandemic risk insurance products may have been financially strong firms that wanted to demonstrate to their stakeholders that they now have profoundly integrated pandemic risk into their risk management processes.[14] Footfall-dependent industries, however, such as hospitality or the event business, which on average experienced a strong impact of the pandemic on their cash flows and business models, will only purchase cover for the next pandemic once they perceive the ongoing event (which cannot be covered under the new policy) as over. Thus, at the time of writing, it remains uncertain if epidemic and pandemic risk insurance products will develop into a stable and sustainable specialty risk market segment.

Low take-up rates for pandemic business interruption insurance before COVID-19 might be explained by the fact that policyholders were expecting their regular business interruption policies to cover the losses from a pandemic.[15] The misalignment of insurers' and policyholders' expectations concerning losses resulting from suppression measures caused global litigation, with rulings varying by country. By the end of 2022, more than 2300 cases had been filed in the USA. Litigation was conducted both over policies including specific virus exclusions and policies without explicit exclusions. However, the majority of claims have been dismissed (for a detailed case overview, see University of Pennsylvania 2023a; University of Pennsylvania 2023b).[16] Industry groups argued that legislature that forces coverage threatens insurers' solvency, with initial estimates from the American Property and Casualty Insurance Association (APCIA) suggesting that losses for small businesses (with less than 100 employees) holding business interruption policies were between 52 billion USD and 223 billion USD per month. In comparison, the combined surplus of the property and casualty

[13] For current considerations on product design, see also Sect. 5.3.

[14] Specific motivations could also have a link to ESG considerations, to integrated risk management services, to project finance or to merger and acquisition activities.

[15] As mentioned above, additional factors contributing to low take-up rates for pandemic business interruption are high premiums and lack of awareness or underestimation of the risk by potential policyholders.

[16] Note that the plaintiffs in these lawsuits were heavily concentrated in a small number of industries which were disproportionately affected by government-ordered shutdowns or other restrictions, with about a third of all cases being related to Food Services and Drinking Establishments. As a consequence, insurers might perceive these industries as more likely to give rise to litigation and change related underwriting practices for such businesses substantially. However, since (as described in Sect. 2) future pandemics might have different characteristics and thus lead to different government responses, litigation patterns might also change.

insurance industry at the end of 2019 amounted to about 890 billion USD and would have been exhausted quickly if these losses were covered. Monthly premiums collected for all commercial property lines would also only make up for a small fraction of these losses (see Hartwig and Gordon 2020 for a more detailed overview of estimates of business interruption losses). However, jurisdiction varies globally, with business interruption test cases being considered in multiple countries in order to provide businesses with clarity over whether policies cover losses related to COVID-19. In the UK, for example, the Financial Conduct Authority brought a test case that involved 21 policy wordings from eight different insurers. The Supreme Court ruled in favor of policyholders in some cases, but not consistently, as 12 policies were found to have the potential to provide cover in response to the pandemic (FCA 2021). The test case, however, did not resolve all of the uncertainty surrounding pandemic-related claims: Litigation over payouts resulting from policies with slightly different wordings compared to the ones in the test case sample is still ongoing. Insurers in the UK paid out more than 1.38 billion USD in COVID-19-related business interruption claims by October 2022 (The Insurer 2023).

As discussed in Sect. 3, business interruption claims were severely underestimated, which has led many insurers to change their policy wordings and add pandemic exclusions to business interruption policies or to increase renewal prices. In addition, only a few insurers decided to launch new products that specifically offer protection for epidemic and pandemic risk. On the other hand, the number of businesses seeking coverage is increasing (Ziady 2020). In order to align supply and demand, capacity and affordability issues have to be addressed. Possible risk transfer solutions will be discussed in Sect. 5.

4.2 Life Insurance

The impact on the life insurance sector was estimated to be moderate based on historical SARS and Spanish influenza scenarios and COVID-19 stress scenarios ran in early 2020 (Richter and Wilson 2020), even though life insurance policies typically do not entail exclusions for infectious diseases. Actual developments turned out to be in line with these estimates and did not generally come as a surprise to life insurers, who already included pandemic risk as part of their pricing model pre-COVID-19: Mortality catastrophe shock scenarios that are part of the standard model for the determination of life insurers' solvency capital requirement were developed with a focus on pandemic risk under Solvency II (see Kraut and Richter 2015). The need for pandemic preparedness of the life insurance sector has already been addressed in 2007 by the Society of Actuaries (SOA 2007). In addition, compared to business interruption policies, life insurance contracts also clearly define the payment trigger and the amount paid out when the trigger is met, which facilitates modeling the risk (Hillier 2022).

Reported claims due to COVID-19 amount to 3.5 billion USD for the global life insurance industry in 2020 and 5.5 billion USD for the first 9 months of 2021 (Cohn and Hussain 2022). However, different countries also show some variation in the amount of premiums earned and benefits paid out.[17] According to Insurance Europe (2021), gross written premiums in Europe fell by about 9.6% in 2020, with policyholders interrupting their pension savings due to economic hardships. Nevertheless, some European countries such as Germany reported stagnation or even saw growth in gross premiums. At the same time, benefits paid out decreased by 9.1%, amounting to 593 billion EUR in 2020 compared to 653 billion EUR in 2019. The fact that returns on unit-linked products were smaller

[17] Note that all figures presented in Sect. 4 are not adjusted for inflation, which also adds to the variation between different countries.

compared to previous years resulted in more policyholders holding on to these plans, which largely contributed to the reduction in payouts. Note that excess mortality (according to Eurostat 2023, the number of total deaths in the EU increased by 11.8% compared to the 2016-2019 average) was particularly high among the elderly, with about 366.000 of 410.000 estimated COVID-19-related deaths occurring among the 65 and above age group. Since life insurance penetration among the elderly is lower, higher death tolls in this age group did not significantly add to the benefits paid out. The higher excess mortality among this age group also resulted in a decrease in annuity payouts for some markets, which contributes to the reduction of benefits paid out. On the other hand, people surrendering their private pensions before maturity due to economic hardship created the opposite impact on payouts.

In contrast to this, payments under life insurance policies and annuity contracts in the USA stayed relatively constant in 2020 compared to the previous year. The payout of death benefits increased by about 15.4% (American Council of Life Insurers 2021), which is driven by the fact that the USA experienced a higher excess mortality than most European countries. Silverman et al. (2022) report a total mortality increase of 18% in 2020 compared to the average mortality of 2017-2019 (which is mainly attributed to deaths resulting from COVID-19 infections and societal changes related to the pandemic), with the biggest increase occurring in the 35-44 years age group. At the beginning of 2021, some life insurers introduced restrictions, limiting coverage to certain age groups and imposing waiting periods before COVID-19 patients were able to apply for coverage (Barlyn et al. 2021). These restrictions were mostly reversed by the end of 2021. However, in the USA death benefits still increased by 10.8% compared to the previous year (American Council of Life Insurers 2022), while lower payouts were expected due to vaccine rollouts (Cohn and Hussain 2022). Excess mortality further increased compared to 2020, with the increase being the highest among the middle-aged group, which is also the most likely to hold life insurance policies (Silverman et al. 2022). Overall payments under life insurance policies and annuity contracts saw an increase of about 10.4%, which was additionally driven by an increase in payouts resulting from annuity surrenders. Death benefits decreased again by 8.3% in 2022 (National Association of Insurance Commissioners 2023b), resulting from lower excess mortality (The Economist 2023).

According to Insurance Europe (2022), preliminary figures for 2021 also show a growth in benefits paid out in most European countries, since mortality among policyholders further increased (with an additional mortality of 14.0% compared to the 2016-2019 average in the EU according to Eurostat 2023), resulting in higher death benefits which were only partially offset by lower annuity payouts. Some countries saw an increase in annuity surrenders induced by policyholders suffering economic hardship. Gross written premiums increased in most European countries, with higher investments in life insurance products made from household savings that were accrued during the previous year. However, some countries such as Germany also saw a decrease in gross premiums, since household savings were also affected by high inflation and high energy prices.

In other regions, the impact on the life insurance sector differs widely from what was observed in the USA and Europe. As an example, Swiss Re (2021b) reports that Latin America faced a particularly high death toll in both 2020 and 2021 also among age groups with high life insurance penetration (see, e.g., Brazil, where 15% of COVID-19 deaths were covered under life insurance policies by September 2021), leading to a large number of pandemic claims.

With an increase in risk awareness, demand for life insurance policies might rise. While premiums were not strongly adjusted immediately (see Cohn and Hussain 2021; Harris et al. 2021), increases due to higher mortality rates might be observable in the long term, resulting from chronic complications from COVID-19 infections and the adverse impact of delayed preventive healthcare visits (Barlyn et al. 2021). However, there may also be offsetting effects due to behavioral changes and better medical care, yielding higher life expectancies. In addition, COVID-19 may have accelerated

mortality among individuals with preexisting conditions, which in turn could result in lower death tolls in the future (see, e.g., Office for National Statistics 2022, who provide evidence for this phenomenon in England and Wales).

4.3 Health Insurance

Since health insurance systems vary across different countries, it is difficult to give a comprehensive overview of how the health insurance sector was impacted by the recent pandemic. Differences in infection rates and hospitalizations add to this challenge. In general, an increase in hospitalizations and testing led to additional costs for health insurers in 2020. This increase was, however, mitigated by restricted health care (including delays in elective and nonessential procedures) and by governments covering part of the costs. Additionally, it should be noted that some private health insurance contracts at least partly excluded pandemic situations (Díaz Llavona 2023). EIOPA (2021) reports that no major decreases or increases in claims were observed in Europe. According to Insurance Europe (2021), paid health claims increased by only 1.1% compared to 2019. In addition, gross written premiums increased by 3.4%, reflecting increasing awareness and a higher demand for private health insurance. In the USA, the profitability of the health insurance sector (with the lines of business including comprehensive hospital and medical insurance, federal employee health benefit plans, Medicare/Medicare supplement, Medicaid, dental insurance, vision insurance, and other health insurance) increased significantly according to the National Association of Insurance Commissioners (2021), with a change in net income in 2020 by 41.6% compared to 2019. This result was largely driven by an 11% increase in net earned premiums that was caused by higher premium rates and rising enrollment. Hospital and medical benefits increased by 6.6% (which is similar to the changes experienced in prior years), offsetting the growth in net earned premiums only to a very limited extent.

However, the industry was affected severely in 2021. In the USA, figures from the National Association of Insurance Commissioners (2022) indicate that profits for 2021 decreased by 41% compared to 2020, making 2021 the least profitable year since 2017. The decrease in profits was driven by high claims, including expenses for medical care and procedures that were delayed in 2020. Total medical and hospital expenses grew by 13.5% compared to the previous year and were only partially offset by an increase in net earned premiums of 7.8%. An upward trend was reported in 2022, with a 26% profitability increase driven by an 11% increase in net earned premium compared to only a 10% increase in hospital and medical expenses (National Association of Insurance Commissioners 2023a). In Europe, preliminary figures for 2021 from Insurance Europe (2022) indicate a rise in claims for most countries, at least partly resulting from procedures and healthcare services that had been postponed in 2020. Additionally, gross premiums grew further in many countries, reflecting increasing take-up for supplementary and complementary health insurance and higher premium levels resulting from rising costs of medical services.

Compared to Europe and the USA, the pandemic impact was more severe in some countries in Asia. In Japan, insurers limited eligibility for COVID-19-related benefits resulting from medical insurance policies to patients at risk of severe illness in 2022. Tightening of eligibility criteria was a result of insurers facing large payouts due to policies not only including benefits paid out in case of hospitalizations but also in case of "deemed hospitalizations," which cover infections where policyholders are quarantined at home (Koide 2022). Similarly, insurance companies in Thailand stopped selling COVID-19 medical insurance policies in June 2021. Four out of the top ten insurance companies have gone bankrupt due to selling policies offering payouts to COVID-19 patients at a low cost (Phoonphongphiphat 2022).

Díaz Llavona (2023) states that future developments in the health insurance sector depend on the impact of delays in routine healthcare visits and the long-term consequences of COVID-19 infections, which pose a risk of higher future claims. While rising awareness will probably lead to a further increase in the demand for additional or voluntary health coverage against treatment costs caused by COVID-19 and future pandemics, health insurers might try to further limit or exclude protection against pandemics or tighten eligibility criteria, e.g., by excluding individuals that have been infected with COVID-19.

4.4 Financial Market Impact and Solvency Ratios

The COVID-19 pandemic additionally affected insurers via its impact on financial markets, with a strong decline in asset values and increasing liabilities in early 2020 (see Sect. 3). However, the consequences vary both globally and by line of business. Thus, we only provide a rough overview.

Analyzing insurers' stock prices shows that the industry performed worse than the overall stock market in the first months after the outbreak. As noted by Richter and Wilson (2020), the Eurostoxx insurance index fell by 23.9% over a 1-year horizon ending early May 2020, compared to a 14.9% decline in the aggregate index. The sharp decline in the Eurostoxx insurance index started in the middle of February 2020 and lasted until April 2020, when an upward trend in the index began to show. Pre-COVID-19 levels were reached again in January 2022, while the Eurostoxx aggregate index recovered faster, reaching pre-COVID-19 levels in mid-March 2021. Similarly, the Dow Jones US Insurance index fell by 19.5% over a 1-year horizon ending at the beginning of May 2020, compared to a 10.2% decline in the Dow Jones Industrial Average. Pre-COVID-19 levels were reached in March 2021 for the insurance index and in November 2020 for the average index.

Solvency ratios were also affected by the COVID-19 outbreak. However, capitalization levels remained relatively stable, with analysts at JPM estimating the average European (re)insurer solvency ratio for the first half of 2020 at 189%, compared to an estimate of 204% for 2019 (Wells 2023). Richter and Wilson (2020) state that this decline was mostly driven by a reduction in own funds, with impacts stemming from lower bond yields, equity market movements, and wider credit spreads. In addition, recovery was quick, with average solvency ratios being estimated at 207% for the first half of 2021 and at 210% for the whole year (Wells 2023).

4.5 Paths for Insurance Market Development

Preparing for future pandemics, and in particular with respect to the necessity of involving the public sector, it is crucial to understand in which direction insurance markets might develop in the aftermath of COVID-19. In the following, we will provide stylized examples for market development. As the impact of a pandemic on life (and health) insurance markets seems to be relatively well-understood, we focus on non-life markets, which, as was previously argued, were to some extent surprised by COVID-19.

Market Scenario 1: Enduring Underwriting Discipline

- We have seen underwriting discipline with regard to communicable diseases tighten since the beginning of the pandemic. As this coincides with a general hardening of the markets, there is no immediate pressure to particularly weaken the underwriting discipline for communicable diseases. The insurance markets have incorporated the lessons learned from the COVID-19 pandemic and remain very restrictive, even when general market conditions might soften at some point in time in the future. If the next severe pandemic hits, there is a high chance that the industry can absorb the losses in line with underwriting expectations. While this is a sustainable scenario for the insurance market, a protection gap for the economy remains.

Market Scenario 2: Return of Soft Market Conditions

- As time goes by, the immediate memory of the COVID-19 pandemic fades away. As a result, at some point insurance markets might start to soften again and the general population continues to underestimate the threat of another global pandemic. Underwriting discipline around communicable diseases weakens, and cover is provided to the market without stringent accumulation control and is underpriced. The protection gap hence is reduced, in particular as wordings now clearly include pandemics, due to the lessons learned by corporate risk managers during the litigation in the aftermath of the COVID-19 pandemic. However, the market environment is not sustainable for the insurance industry. When the next severe pandemic hits, industry losses may well be higher than the ones from the COVID-19 pandemic.

Market Scenario 3: Development of a Sustainable Specialty Market Segment

- This market scenario is an extension of market scenario 1, as a sustainable specialty market segment to a degree requires a sustainable risk assessment by insurance markets in general. The lessons learned from the COVID-19 pandemic have been incorporated by the markets, and the underwriting discipline remains high even beyond the hard market phase, recognizing the fundamental challenge posed by accumulation. While pandemic risk remains excluded from commodity products such as property damage business interruption, a specialty segment develops with strict accumulation control. This way the insurance industry is not overexposed but can deploy its capabilities and limited capacity in a sustainable way, allowing more and more market participants to build up the required underwriting expertise for pandemic risk. The protection gap for commercial customers is reduced by an increasing amount of "survival covers" for pandemic risk (see Sect. 5 and, for general related considerations around the protection gap, Kawai 2022). As affordability of epidemic insurance for corporate clients remains challenging, they develop risk management strategies that combine insurance with other instruments. One further interesting aspect of this market scenario is that the existence of a specialty market segment and the corresponding price signals may increase the awareness of pandemic risk exposure even among firms that do not participate in this specialty market segment.[18] This bears the potential for an overall reduced vulnerability of economies, even beyond the aggregate insurance capacity purchased by firms.

In any of these scenarios, insurance markets will only be able to provide partial cover for the total losses associated with a major pandemic. Ways of integrating additional sources of capacity will be addressed in the next section.

[18] These firms may employ other means of reducing exposure, e.g., by diversifying their supply chain and revenue stream.

5 Implications for Future Pandemic Risk Management

COVID-19 has highlighted insurability problems associated with pandemic risk, and we have seen why the insurance markets' capacity for this type of risk is limited. To some extent, insurability concerns can be addressed via product design in order to expand the boundaries of insurability. More broadly, however, given the COVID-19 experience and with epidemic and pandemic outbreaks increasing in frequency and severity, pandemic risk management as a whole must be improved. In the following we discuss implications for societal preparedness and government responses and for individual risk management, with a focus on the design of risk transfer solutions and the role of the different channels that are or can be involved: the insurance industry, capital markets, and the public sector. We only briefly address other risk management strategies related to public preparedness and corporate risk management. While these aspects are very important for finding a path forward, they require extensive discussion. Further research will be necessary to specify the optimal design of these strategies and to quantify their value.

5.1 Public Health Preparedness and Response

In the context of epidemic outbreaks, investments in emergency preparedness may lead to the detection and earlier containment of new pathogens. Some outbreaks may even be fully prevented. The project *Predict*, which was initiated by the US Agency for International Development in 2009, provides an example of an early warning program identifying viruses with the potential to develop into pandemics. It was, however, ended in September 2019, i.e., just shortly before the beginning of the COVID-19 crisis (Milman 2020).

If well prepared, the response to new outbreaks will be quicker and more effective. Response measures for epidemic outbreaks include physical mitigation strategies (e.g., mask wearing), medical measures (e.g., vaccines, antivirals), measures to reduce the vector population (e.g., fogging against mosquitoes), or actions influencing human behavior (e.g., social distancing, communication to prevent panic). Lockdowns will most likely continue to be an important instrument of epidemic and pandemic containment. If pharmaceutical interventions (such as antiviral drugs or vaccinations) are available, these together with contact tracing and advisory non-pharmaceutical interventions will naturally be the first choice for governments in responding to epidemics. In this context, investments in research and development, collaborative genomic surveillance and the surveillance and mitigation of supply chain risks in the production of vaccines, and other medical countermeasures will be necessary (Agarwal and Gopinath 2021; Agarwal et al. 2022). However, if no pharmaceutical interventions are available and advisory non-pharmaceutical interventions, such as encouraging the population to adopt good hand and respiratory hygiene, wearing a face mask, and staying away from sick people, are not sufficient to contain the epidemic and healthcare systems are at risk of being overwhelmed, governments may have no choice than to apply mandatory measures which severely limit people's interactions (such as civil authority restrictions and lockdowns).[19] They could potentially occur more frequently on a smaller, regional scale in order to prevent the spread via spill-over events to other regions early on. For an example of a detailed overview of response measures to a respiratory pathogen, we refer to the UK Influenza Pandemic Preparedness Strategy (United Kingdom

[19] Both the effectiveness of lockdowns in controlling pandemics and the economic efficiency are subject to controversial discussions (see Yanovskiy and Socol 2022 for a review of studies analyzing lockdowns in different countries concerning their impact on containment and economic losses). Further research on the design of lockdowns as risk management tools for future pandemics will be necessary.

Government Department of Health and Social Care 2011). Different approaches to analyzing optimal policy interventions are used by Echazu and Nocetti (2020), Gollier (2020) and Smetters (2020). In addition, Salanié and Treich (2020) model the externalities resulting from changes in individual behavior that compulsory self-protection measures induce, and Courbage and Peter (2021) investigate personal vaccination decisions in the context of ambiguity.

The World Bank (2022) provides guidance on investing in the prevention of pandemics and on overcoming the lack of investment incentives. The COVID-19 pandemic has shown the consequences of systematically underinvesting in resilience. The expert community had warned for decades that pandemic risk is on the rise. With SARS, Zika, and the 2009 Swine flu event, sizeable outbreaks of severe pathogens and even a mild global pandemic had occurred. Nevertheless, preparedness was clearly not at the level where it could have been, considering a risk that now has proven to substantially dampen economies, impact work, and trade, as well as social interactions for an extended period of time (Munich Re 2020). A report by the The Independent Panel for Pandemic Preparedness and Response (2021) highlights that there is a need for investment in preparedness now, before the next crisis hits, and for access to financial resources, in particular to be able to quickly inject funds at the very beginning of another (potential) pandemic. Similarly, in light of the COVID-19 pandemic, the OECD (2023) emphasizes the importance of pandemic preparedness in their recommendation on disaster risks and pandemic resilience. Prevention, preparedness, and response financing has previously been insufficient and fragmented (Stefan et al. 2023a). In order to scale up capacities, the G7 (G7 Health Ministers 2023) proposes a three-layered response funding system, including the mobilization of domestic resources as a first step, followed by the utilization of existing multilateral mechanisms and financing instruments. As a third layer, a new "surge" financing framework complements the previous layers in order to deploy funds quickly and in a coordinated manner. Donors commit to disbursement upon pre-agreed triggers. Stefan et al. (2023a) also emphasize the role of a disaster risk finance approach (i.e., financing upon pre-agreed triggers, with allocation of the funds already being planned before disasters happen). By directly providing liquidity in the initial stages of an outbreak, the "surge" financing facility could complement the Pandemic Fund, which was set up in 2022 by the World Bank and is directed at investments in preparedness and detection.

The experience of governments, business entities, and individuals forms a basis for learning from the past and for collaboratively adapting to increased risk exposure. All stakeholders play an important role in developing resilience, thus also defining the speed of a community's recovery post-event. Understanding the interdependence between epidemiological factors and available countermeasures and response options is crucial for understanding both the public health and economic consequences of an outbreak caused by a specific pathogen. Further research is needed to detail the quantitative reduction of economic consequences for a future outbreak of a given severity which can be achieved by better preparedness. To our knowledge, most of the existing literature evaluating the benefit of pandemic preparedness is focused on public health in terms of the value of lives saved and medical costs (see, e.g., Meltzer et al. 1999; Bernstein et al. 2022), while only Gopinath (2020) and Glennerster et al. (2022) capture the reduction in economic output losses.

5.2 Corporate Risk Management

From the perspective of a corporate risk manager, the characteristics of pandemic risk include the following aspects: First, a pandemic occurs with a significant probability per year. Second, if the risk materializes, the duration of the event is expected to be several months to potentially years. Finally, the economic impact on business operations is difficult to mitigate for the individual business even

with good individual risk management, especially for footfall-dependent industries.[20] Possible risk management strategies are (1) do nothing. For sufficiently financially strong businesses with strong risk identification, this may be the efficient choice. Related to this is (2) building capital and cash reserves.[21] This strategy is important in order to increase resilience. However, it will typically be subject to tax constraints, and the business requires a high degree of financial discipline. (3) Based on an evaluation of the risk exposure, alter the business strategy where necessary, and introduce business-continuity plans that cover critical operational processes, supply chains, and distribution.[22] (4) Risk transfer. However, for potentially catastrophic losses occurring with a probability in the single-digit percentage range each year (see Sect. 2.4), full risk transfer may neither be feasible nor even preferred compared to a partial risk transfer focused on the financial survival of the business operations.

Captive insurance companies provide large businesses with the opportunity to self-insure against pandemics. Formation of these entities nearly doubled in 2020, with pandemic disruptions being a contributing factor to this growth (Dunsavage 2022). For small- and medium-sized enterprises, however, captive insurance might not be feasible. Instead, Louaas and Picard (2023) suggest a self-insurance mechanism based on a portfolio diversification strategy that exploits differences in the reaction of stock returns to pandemic events. The capacity of businesses for self-insurance will, however, often be limited compared to the potentially severe economic losses resulting from a pandemic.

In general, it is crucial for businesses to accumulate capital in order to absorb losses arising in the initial period following an outbreak.[23] Investments in risk mitigation and pandemic preparedness additionally lower the economic impact of the next event. Further research regarding risk management strategies and their impact, which also depends on the nature of the outbreak, is necessary. According to Kekulé (2017), epidemic business continuity management should incorporate a pre-outbreak stage in which plans for surveillance, response, and recovery are devised, including an assessment of potential threats and their impact. Surveillance involves paying attention to alerts from the WHO and local health authorities. He suggests that in the initial phase of an outbreak, response measures should include assessing the possible impact of the incident on the business and determining and implementing countermeasures or alternative strategies (which are ideally derived from plans devised in the pre-outbreak stage). In addition, to prevent the disease from affecting the business, it is crucial to protect employees from infection. Prevention continues to be important when an epidemic outbreak starts to affect business operations. In this stage, the goal should be to reduce the impact on the business and maintain operations (at least at a reduced level) if still profitable. Dispensable operations could be closed to move the workforce to other sections and secure critical infrastructure. The protection of employees also remains a crucial aspect.

Businesses adapted to constraints during the COVID-19 pandemic, e.g., by shifting to work from home and through online sales, which may be an advantage during future lockdowns. In addition, some businesses have mandated vaccines for existing employees or new hires during the COVID-19 pandemic or imposed health insurance surcharges on unvaccinated employees (Japsen 2021; Messenger 2021). As an example, starting in November 2021, unvaccinated employees of Delta Air Lines faced a 200 USD monthly health insurance surcharge, while all new employees were

[20] For example, consider two hotels in the same town. One may be managed very well with particular hygiene measures in place, while for the other hotel, only poor risk management strategies apply. However, if an epidemic outbreak or fear thereof leads to a general reduction of travel, both hotels may experience similar reductions in bookings.

[21] For larger corporates, this could be implemented within a captive.

[22] Sticking to the same example as before, a hotel could try to reduce dependency on foreign tourism.

[23] Even in markets where government support was widely deployed, this often came with a time lag, straining the financial liquidity of businesses.

required to be vaccinated. The policy was ended in April 2022, when more than 95% of employees had been vaccinated according to the company. The higher premiums were justified by the fact that unvaccinated individuals are more likely to be hospitalized when infected, causing high costs (Josephs 2022). An additional strategy is provided by mandatory testing, for which some employers also covered the costs. The abovementioned policies have, however, been widely debated (Yang and Chen 2022). As an alternative, companies were providing financial incentives by paying their employees to get vaccinated (Maruf 2021).

5.3 The Role of Insurers and Reinsurers

While the insurance industry can build up the expertise to underwrite and price pandemic risk, accumulation risk remains the fundamental challenge that needs to be addressed. Strict accumulation control across all lines of business will be required, particularly if comprehensive covers for pandemic business interruption are offered. Structured (re)insurance solutions may support the efficient deployment of capacity, given a limited accumulation budget. As was discussed in Sect. 3, a fundamental difference between pandemic risk and typical natural disaster threats is the fact that regional diversification will usually not work well in the pandemic risk case. While reinsurance and similar risk transfer mechanisms enable the worldwide spread of natural disaster exposures that allows for the law of large numbers to work, this will not be possible with significant worldwide correlation. This observation, however, affects the potential role of the (re)insurance industry. Since the risk appetite of the industry is limited and if the goal is to achieve broader economic resilience, the role of insurers might be to primarily cover a small share of losses ("survival covers"), with payments enabling businesses to remain viable in the initial stages of a pandemic. Beyond this, insurers and reinsurers can use their expertise in structuring and pricing to help transfer the risk to the capital markets and the public sectors. Furthermore, the industry takes on an important role in providing incentives and assistance for risk mitigation and risk management. Product design can motivate risk mitigation through risk-based pricing and tools such as coinsurance or deductibles which were already mentioned in Sect. 3. In addition, insurance payouts could also be made contingent upon meeting certain requirements, e.g., implementing predefined preparedness and response measures and adhering to public health guidelines, which was also recommended by the US *Business Continuity Protection Programme* (BCPP) proposal (if monitoring of adherence to these requirements is possible for insurers) (OECD 2021). In some cases, risk management services may be integrated into insurance products. As an example, see the services provided to policyholders by companies such as International SOS, which assist strategic decision-making related to risk management (Richardson 2023).[24]

But ultimately, the supply side hurdle of accumulation risk cannot be resolved within the insurance industry alone. If, beyond a small niche insurance market segment, a larger share of economic losses needs to be transferred, additional risk-bearing stakeholders need to be activated that are not yet invested in the risk. This can be achieved via the capital markets, again with appropriate accumulation control in place, and the public sector, which currently already holds a large share of pandemic risk exposure.[25]

[24] Similar solutions are included in cyber insurance policies, which provide the insured with an assistance hotline to call when under cyberattack.

[25] This is in particular true if ad hoc risk bearing, such as the disbursement of subsidies to stabilize the economy, is included in this consideration.

An open question that has to be addressed is whether insurers should be forced to offer pandemic insurance through mandatory requirements. While this approach at first sight ensures sufficient availability of coverage, the mandate would, considering the concerns regarding the accumulation risk and limited capacity, generate a major threat of overexposing the industry. Depending on the design of the mandate, it would create particular challenges for insurers who lack the tools to underwrite pandemic coverage and might severely increase these carriers' insolvency risk (Klein and Weston 2020). At the same time, insurers might decide to withdraw from the respective market segments to avoid the increased insolvency risk. An overexposure of the insurance industry might be mitigated partly if offering the risk is mandatory only at a risk-adequate price (obviously the abovementioned problems with pricing need to be kept in mind). Given the potentially high premiums (note that not only expected losses but also correlation-based loadings will typically be substantial), demand might be limited, due to, among other factors, non-affordability and/or insufficient risk perception. While the market might develop into a state where some individual businesses may be (better) protected against epidemic risk, the pickup rate might be too low to achieve the goal of a broadly increased economic resilience.

If sufficient capacity can be provided successfully, it remains to be discussed whether there should be a requirement for businesses to purchase coverage in order to ensure broad take-up, which was proposed, for example, by the German Insurance Association (Reuters 2021). As in most potential applications, mandatory coverage is discussed rather controversially. Since businesses are reluctant to spend too high a share of their profit margin on protection against one single risk factor, typically premium subsidies are discussed in this context as well. In the so-called *government pooling discussions* observable in many countries and markets during the COVID-19 pandemic,[26] this dilemma could not be resolved. If the goal is to increase economic resilience of the whole economy from day 1, any risk transfer mechanism would require too much capacity (even for the public sector). If affordability should be addressed within an insurance mechanism to generate broad pickup, substantial premium subsidies appear to be required. Hence, any solution with the goal to increase the pandemic resilience of an economy in practice may need to start small as a specialty market segment, with the ambition to crowd in further capacity from different sources over time. Substantially increasing demand by overcoming the affordability hurdle may require the combination of different risk transfer instruments beyond insurance.

According to Nebolsina (2021), demand for pandemic business interruption insurance in the USA is increasing, with the COVID-19 pandemic leading to greater risk awareness among businesses. This is an indication that market development may be possible. However, the effect of the pandemic on insurance demand might be different in other countries. In addition, fundamental aspects that should be considered when analyzing potential product designs and market developments are the incentives of the stakeholders. For instance, given the experience from the COVID-19 crisis, firms could anticipate government post-event aid, a phenomenon that is known in the literature as "charity hazard" or the "Samaritan's dilemma" (see, e.g., Buchanan 1972). A mandate for businesses to purchase insurance would naturally reduce the issue of charity hazard (Medders and Schwarcz 2022). In addition, it might also contribute to improving affordability by decreasing the risk of adverse selection (Allianz SE 2020). However, as mentioned above, mandatory coverage tends to be very controversial and might not be a good solution for business interruption risk. The more desirable route seems to be implementing voluntary coverage and ensuring broad take-up by increasing risk awareness and

[26] For clarification purposes, the term pooling in these discussions in most cases refers to the government acting as ultimate risk bearer, rather than making use of diversification benefits by pooling risks together and jointly bearing it (cf. Kraut 2014; Kraut 2015).

improving affordability, with premiums that will be viewed as reasonable by businesses (Dixon and Saunders-Medina 2020).

With any form of pandemic business interruption protection provided by the private sector, a core issue will be the incentives for political decision-makers. As argued in Sect. 3, it will be easier for a government to implement strong containment measures if insurance protection is in place that covers a significant portion of the consequences. To address this external moral hazard, ideally the insurance industry should avoid the first layer of the loss. And where insurers are involved in covering the first loss, other stakeholders must participate substantially as well to avoid distortions in political decision-makers' incentives. If a market scenario materializes where the industry provides only the "survival covers" that were mentioned above and if no other risk transfer mechanisms are in place, the insured will continue to bear a substantial share of economic losses themselves; hence the incentives for the government might not fundamentally change. If, however, in order to achieve widespread economic risk transfer and thereby to increase economic resilience further, the public sector participates in the risk bearing (see Sect. 5.5), it is important to ensure that incentives are aligned. This could, for example, be achieved by proportional risk-sharing between the private and the public sector.

Nevertheless, the government action itself should not be the sole trigger for insurance payments. Rather, international public health agencies may be more suitable to decide upon the severity of an epidemic outbreak. The smaller the impact of government action is on the insured event and on the insured losses, the broader insurance solutions can be. Clearly, a situation where the insurer is the only stakeholder in covering the first loss layer with the payout fully dependent on government actions needs to be avoided.

In terms of coverage design, it is also important that insurers' wordings are clear to avoid uncertainty around claims. Triggers in epidemic and pandemic insurance policies must be transparent and should thus not involve too much complexity. Glynn et al. (2021, Chap. 14) propose a combination of three different triggers, with the first trigger being the WHO reporting a new epidemic outbreak via Disease Outbreak News, the second trigger being a measure of the severity of the outbreak, for example, based on casualties or fatalities or relying on a global alert such as the Public Health Emergency of International Concern,[27] and the last trigger linking the outbreak to the disease impact in a certain area, which could be via evidencing the presence of the disease in the covered area, or legal restrictions imposed as a response to the epidemic outbreak.[28] This approach would reduce external moral hazard since payouts are not simply contingent on lockdown declarations but reflect the underlying event structure. If implemented in a parametric way, such as determining the magnitude of payouts based on local industry loss indicators, these triggers also limit or exclude moral hazard that arises, when, e.g., a company's decision to cancel an event triggers insurance coverage (see Sect. 3). However, the approach also entails some basis risk, with payouts not necessarily matching the actual needs of the insureds. In contrast, indemnity-based policies that determine payouts according to the losses that were actually incurred minimize basis risk but also contain the risk of moral hazard (Doherty and Richter 2002). Since parametric products facilitate claims handling and provide swift payouts, they might be used to cover a first layer of losses that will keep businesses viable during the initial stages of a pandemic,[29] with traditional indemnity products compensating for further losses arising (see Hillier 2022; Jerry 2023) or being used for larger capacities. A parametric first loss layer

[27] Under consideration of external moral hazard, national alerts that are influenced by political decisions are not suitable as policy triggers.

[28] For clarity, such a product design will cover both epidemic and pandemic events. Once the policy is activated, there is no differentiation depending on further outbreak development.

[29] Or until government support starts at a later point in time.

may in particular address the need for quick liquidity inflow in the moment of crisis to better manage the ongoing pandemic event.

Glynn et al. (2021, Chap. 14) additionally address the fact that businesses might be impacted differently during the course of a pandemic, with some companies suffering losses during early stages, while others are affected later on. In terms of the period of cover, they thus consider an indemnity period of a fixed length that is flexible concerning its beginning: Once the triggers are met, the policyholder is allowed to choose when to start the interval during which claims can be made.

Regardless of how policies are designed, it is important to keep in mind that the insurance markets cannot cover pandemic risk in its entirety. (Re)insurers will have to put limits on the amount of risk that they are willing and able to take on, according to their risk appetite and capacity.

5.4 Capital Market Solutions

The potential magnitude of pandemic losses is clearly too large to only be taken on by the insurance industry. Risk has to be transferred to capital markets and the public sector. Alternative risk transfer instruments, including derivatives, swaps, funds, and insurance-linked securities, provide solutions to obtain capital from the private market. Additionally, private sector entities such as banks could come in as contingent lenders (Kraut and de Kuiper 2021).

Reinsurers can issue and subscribe bonds on the ILS market in order to off-load their risk exposure. Extreme mortality risks, including fatalities resulting from pandemics, were managed through mortality cat bonds as early as the 2000s, with the first bond being issued by Swiss Re in 2003 (MacMinn and Richter 2018; Swiss Re 2021a). Demand for these bonds has, however, been quite limited (Schiller 2022). Additionally, for insurance-linked securities covering excess mortality risk or other life- and health-related risks, risk capital issued has not significantly increased during the COVID-19 pandemic compared to previous years (Artemis 2023b).[30]

Further examples of capital market solutions include the World Bank's *Pandemic Emergency Financing Facility* (PEF), which was set up in 2017 as an instrument to quickly provide funds to the world's poorest countries when facing potential pandemic outbreaks. The World Bank (2020) provides a description of the PEF's two independent payment mechanisms, a funding window and an insurance window which are supposed to complement each other. While the funding window was financed through contributions from Japan, Australia, Germany, and member countries of the *International Development Association* (IDA), the insurance window provided coverage through pandemic risk-linked swaps executed with insurance companies and through pandemic catastrophe bonds that were sold to capital market investors. Activation criteria for the insurance window in case of the outbreak of a non-flu virus included a total ongoing case amount and a total number of fatalities of 250 or more. In addition, at least 12 weeks had to pass from the start of the outbreak, which had to affect at least two countries with 20 fatalities or more each. In contrast, the funding window was designed to provide fast payouts to countries fighting disease outbreaks upon approval of a steering body. While the 2018-2020 Ebola outbreak in the Democratic Republic of Congo (DRC) and Uganda only triggered payments from the funding window due to the limited international spread of the outbreak,[31] the COVID-19 outbreak triggered both the funding window and the insurance window. Activation criteria for the insurance window were met on March 31, 2020. Payouts that

[30] Artemis (2023a) and Artemis (2023b) provide a continually updated deal directory containing details about catastrophe bond and ILS transactions issued, including date of issuance and covered risk type.

[31] It is assumed that the outbreak escalated in the DRC also due to the simultaneously ongoing civil conflict.

were disbursed to IDA countries via the insurance window amounted to around 196 million USD. With pandemic bonds and swaps maturing in July 2020, the mechanism was not renewed during the ongoing pandemic and was officially closed by the World Bank on April 30, 2021.

The PEF has been widely criticized for both its costs and the time that it took to be triggered. However, addressing the critique concerning the timing of payouts, one should bear in mind that the structure was designed for IDA countries only (where the COVID-19 pandemic did not start) and also was designed to attract capital market participants (Evans 2020). Triggers that have a high risk of being met quickly will discourage investors or lead to high risk premiums. While payouts may have been too slow, the insurance window still served its purpose of providing relief to the poorest countries. Two further aspects concerning the costs of the PEF should be addressed: First, Hartwig et al. (2020) argue that relative to the opportunity cost of capital, the pricing of the bonds was fair. Second, there was criticism stating that issuing insurance-linked securities is an expensive tool to provide disaster relief. Lane and Beckwith (2020) point out, however, that alternatively disbursing funds via post-event borrowing will put the World Bank at default risk and question whether post-event funding will be available at affordable rates.

A similar concept has been launched by the *African Risk Capacity* (ARC) in 2022, with a sovereign parametric insurance product that can cover member states of the African Union for the costs of interventions necessary to manage and contain outbreaks of Ebola, Marburg virus disease, and meningitis. Senegal was the first country to take up protection. Significant response capacity has to be built up by creating a sovereign risk pool and involving reinsurers and capital markets as capacity providers (see African Risk Capacity Group 2022a; Agyarko et al. 2022; Evans 2022). For the pilot transaction under the ARC program, the participation of several renowned capacity providers could be achieved (see African Risk Capacity Group 2022b; Global Parametrics 2022).

In general, increasing capacity will only be possible if the supply challenge of creating a sufficiently large investor base is adequately addressed. Securitized catastrophe risk usually presents low correlation with market risk (see Litzenberger et al. 1997). However, this argument is questionable for cat bonds covering extreme events (Gürtler et al. 2016). As already discussed in the previous sections, pandemic risks are correlated with financial market developments. While the markets recovered relatively fast during the COVID-19 crisis, there is still uncertainty around financial market developments resulting from future pandemics. The risk that pandemic losses have a negative influence on financial market movements might lead capital market participants to refrain from investments or to demand very high risk premiums. On the other hand, during a financial market crisis that is not caused by an epidemic or pandemic outbreak, capital market solutions for pandemic risks will remain unaffected (Medders and Schwarcz 2022). Hence, adding a defined amount of pandemic risk to an investment portfolio can still improve portfolio diversification. In addition, capital market participants might be attracted by ESG aspects, with investors seeing a societal value in the fact that capital market solutions will contribute to containing epidemic outbreaks or to keeping businesses viable during a pandemic outbreak. However, since the recent pandemic is still very present in investors' minds, there might be an initial aversion to pandemic capital market products, which could be reduced if other catastrophic events gain focus (Arenz et al. 2023).

Medders and Schwarcz (2022) argue that pricing pandemic cat bonds is feasible, since the loss magnitude can be limited ex ante and the probability of a triggering event is smaller compared to other catastrophic events, whereby the latter naturally depends on the specific calibration of the instrument. Triggers have to be designed carefully. While it would be desirable that triggers are met quickly upon epidemic or pandemic outbreaks to provide swift payments, it is necessary to strike a balance with the risk appetite of investors. The design of pandemic capital market instruments focusing on pandemic business interruption might follow the approach of Glynn et al. (2021, Chap. 14) using a combined trigger (see Sect. 5.3), which will offer a high amount of transparency to investors. A trigger for

pandemic cat bonds with similar characteristics is also proposed by Spaeter (2021) and Schmitt and Spaeter (2023), who provide a comprehensive discussion both on the associated basis risk and moral hazard and on how to implement these bonds in a risk-sharing scheme that also involves (re)insurers and the public sector. Capacity of the capital market will be limited to some extent, underlining the necessity of public sector participation in the risk transfer scheme.

5.5 The Role of the Public Sector

Taking into account the capacity and affordability issues highlighted in this chapter, political decisions about the right design of pandemic risk management solutions should also involve the public sector as a potential risk-bearer. Multiple approaches have been proposed in light of the COVID-19 pandemic. In France, insurers suggested a concept for a national insurance solution called *CATEX*, which is designed as a public-private partnership providing coverage for disasters including pandemics with a backstop guaranteed by a state-owned reinsurer (EIOPA 2021). The German Insurance Association put forward two different models for a pandemic support system, highlighting the role of government support (see Richter and Wilson 2020 for more details on the proposal). Furthermore, the USA saw multiple proposals for public–private partnerships, including the aforementioned BCPP and the *Pandemic Risk and Insurance Act* (PRIA). While the BCPP shows similarities to the *National Flood Insurance Program* (NFIP), the PRIA was designed based on the *Terrorism Risk Insurance Act* (TRIA) that was implemented after the 9/11 attacks. A comprehensive evaluation of the US proposals can be found in Klein and Weston (2020). Richter and Wilson (2020) emphasize that under the BCPP proposal, the insurance industry would help administer protection but would not assume the actual risk (and therefore avoid the problem of external moral hazard).

Hartwig et al. (2020) broadly classify different government involvement proposals into four categories. Lacking specific preparedness plans, most governments only provided post-pandemic aid during the recent pandemic, assigning funds to impacted businesses. Alternative approaches for governments are to either act as a reinsurer to insurers who provide businesses with coverage against pandemic losses or to directly insure businesses. As a fourth category, a social insurance scheme will require many or all entities to make pre-event payments, with an administrative system disbursing the funds during a pandemic event. While each approach comes with different advantages and disadvantages, there is no clear conclusion on whether one of the approaches is superior. However, according to the authors, post-event aid and social insurance (assuming payouts are disbursed in relatively uniform amounts) perform poorly in matching losses and compensation and provide low incentives for risk mitigation. As argued in Sect. 5.3, using the resources and the existing infrastructure provided by insurers will be more efficient both in matching funds and needs and in incentivizing preparedness. In addition, post-pandemic aid will be a burden to future generations, e.g., through tax increases.

In a system that is purely based on post-event aid, the need to rapidly disburse funds will lead to a poor alignment between payouts and losses. With bureaucratic delays and lacking expertise, efficient disbursement of government aid might be very slow. Thus, in order to reduce the impact on the overall economy, the government could act as a reinsurer and additionally provide post-event aid. While governmental reinsurance (making use of the insurers' infrastructure) contributes to swift payouts of claims, the latter can additionally provide targeted help that can be thoroughly planned and adjusted when the amount of losses is determined (see Hartwig et al. 2020 for a comprehensive description and evaluation of the approaches). Wolfrom (2022) provides further analysis on the example of the COVID-19 pandemic, exploring how insurance could provide an alternative to fiscal support in crisis response, also stating the importance of rapid payment mechanisms, limiting public exposure through

leveraging available private sector capacity and targeting companies in need, as well as providing incentives (or requirements) for risk reduction and employment stability (see Wolfrom 2022, pp. 30-31).

Government pooling discussions received a large amount of attention during the initial phase of the COVID-19 pandemic. They have, however, significantly slowed down over time, with the implementation of proposals being delayed or rejected. A contributing factor to this slowdown might be that even for governments, pandemic losses are too large to take on, especially if affordability for insured businesses would additionally be addressed via premium subsidies. Proposals have either been limited to payouts that only cover a small fraction of losses or a certain economic sector or have been severely underpriced (Kraut and de Kuiper 2021). In order to address this failure, another (proportional) risk layer in addition to insurance transfer solutions could be provided by governments coming in as lenders at market rates alongside the private sector. Contingent lending at pre-agreed terms will make funds available upon pre-agreed triggers, which can be designed in the same way as triggers for pandemic cat bonds and insurance products. Additionally, premiums for businesses facing the largest risks could be subsidized. However, improving affordability purely through lending without using subsidies would enable premiums to remain risk-based, thus both reducing adverse selection and setting incentives to increase preparedness and economic resilience, which could in turn contribute to increasing the capacity that is provided by private sector investors and to mitigating overall economic losses to a certain degree. A comprehensive description of a risk transfer market platform which includes both the private and the public sector as investors and lenders is given by Kraut and de Kuiper (2021).[32]

Interestingly, most proposals for government involvement did not consider the inclusion of capital markets. In addition to the alternative risk transfer instruments that were presented in the previous subsection, risk allocation could be further improved by governments issuing sovereign cat bonds. Making these cat bonds available to both institutional and individual investors and limiting an individual's maximum loss would enable broad participation, allowing losses to be borne by many entities (Richter and Wilson 2020). Another approach of government involvement is proposed by Medders and Schwarcz (2022). The authors suggest that government support takes place in the form of guarantees (at least partially) for pandemic cat bonds. As an alternative, governments could purchase the shortfall between the capital that is needed and the risks that investors are willing to take on. To enhance capital market participation, governments should purchase more subordinated bonds, with bonds acquired by more risk-averse private investors being more senior.

While both government and capital market participation are essential to improve affordability and increase capacity, it should be noted that due to the magnitude of potential pandemic losses, even a functioning risk transfer system will initially only create a specialty market segment that covers a small part of pandemic losses. However, this segment will have the potential to grow over time, attracting further investors and accumulating capital (Kraut and de Kuiper 2021).

Lastly, the public sector potentially also has a role to play in setting a regulatory framework that incentivizes or requires pandemic risk management, as illustrated earlier by the regulatory differences of taking pandemic risk into account for life and non-life insurance markets (see International Association of Insurance Supervisors 2022 for current considerations). Further research may be useful to evaluate different policy measures.

[32] The recommendation to implement such an open Epidemic Risk Markets Platform has been endorsed by the APEC Finance Ministers during the 29th APEC Finance Ministers' Meeting in Thailand in October 2022 (see ABAC 2022a; ABAC 2022b and APEC 2022 for the endorsement and ABAC 2022c for further details on the original recommendation of the APEC Business Advisory Council (ABAC)).

6 Conclusion

COVID-19 has changed the perception of pandemic risk. Estimates of the probability and potential severity of outbreaks underline that it is necessary to reflect on lessons learned and look ahead to future pandemics. Risk modeling and management will thus be crucial in preparing for the next pandemic.

There seems to be a broad consensus that pandemic risk is systematic in nature and that it violates important requirements of insurability. The impact of the COVID-19 pandemic on the insurance industry is in line with these findings, with unexpected claims and a significant amount of uncertainty concerning business interruption policies leading insurers to specifically exclude pandemic risk. While other lines of business were not severely affected, long-term effects on the industry are yet to be seen.

As the responses to COVID-19 varied widely around the world, further economic research is needed that analyzes the effectiveness and efficiency of different interventions and risk management tools. Is there empirical evidence regarding parameters that were particularly relevant for the time of economic recovery? How important was the existence of developed insurance markets and the timing of financial support? Can we draw conclusions regarding the way in which governmental aid should best be distributed?

It should, however, be noted that future pandemic outbreaks might differ significantly from the recent COVID-19 outbreak. Infection and mortality rates and governmental measures will possibly vary strongly, which poses additional challenges in risk modeling. Nevertheless, it will be essential to increase capacity and improve the affordability of pandemic insurance. Solutions have to involve both the private and the public sector. While government pooling discussions have recently slowed down significantly, it is necessary to implement risk transfer solutions as quickly as possible. This approach will initially only create a specialty market segment. However, with time, it might be possible to attract further investors and build up a significant amount of capacity. In order to do so, it is crucial to find ways to enhance the appetite of capital market participants and, if the goal is further scale beyond "survival covers," to combine different financial instruments. Beyond risk-bearing, insurers and reinsurers can play a very important role, assisting insureds in risk mitigation, incentivizing preparedness, and using their experience and resources in structuring and pricing solutions. A functioning risk transfer system that involves participants covering different layers of risk, proportionally or nonproportionally, will contribute to lowering the magnitude of pandemic losses and thus reduce the burden on society, especially on future taxpayers.

Acknowledgments The authors would like to thank Christophe Courbage, Tobias Huber, Kerry McCabe, Tracey Wood, and one anonymous referee for very helpful comments and discussions. All errors and omissions remain the authors' responsibility. Opinions stated in this chapter are the authors' personal views.

References

ABAC (2022a) 2022 report to APEC finance ministers: a call to concerted action. http://www2.abaconline.org/assets/2022/Ministerial_Letters/ABAC_Letter_to_APEC_Finance_Ministers_2022.pdf. Accessed 12 July 2023

ABAC (2022b) Building a resilient future: enabling financial services to promote sustainability, digitalization and risk mitigation. http://www2.abaconline.org/assets/2022/2022_Progress_Report_of_APFF_APFIF_and_APIP.pdf. Accessed 12 July 2023

ABAC (2022c) Enhancing the region's resilience through PPPs in pandemic risk transfer. http://www2.abaconline.org/assets/2022/Reference_Documents/Reference%2003%20RT%20Report%20Pandemic%20Risk%20Transfer.pdf. Accessed 12 May 2023

Adalja A, Watson M, Toner ES et al (2022) The characteristics of pandemic pathogens. Johns Hopkins Center for Health Security. https://centerforhealthsecurity.org/sites/default/files/2022-12/180510-pandemic-pathogens-report.pdf. Accessed 06 May 2023

African Risk Capacity Group (2022a) The African Risk Capacity launches its first parametric insurance product against high-impact epidemic risks. https://arc.int/news/african-risk-capacity-launches-its-first-parametric-insurance-product-against-high-impact. Accessed 04 Apr 2023

African Risk Capacity Group (2022b) Today, the government of Senegal officially signed the first-ever sovereign parametric insurance policy. https://www.linkedin.com/feed/update/urn:li:activity:7011265304999657472/. Accessed 14 May 2023

Agarwal R, Gopinath G (2021) A proposal to end the COVID-19 pandemic, vol SDN/2021, 004 (May 2021). International Monetary Fund, Washington, D.C.

Agarwal R, Gopinath G, Farrar J (2022) Global strategy to manage the long-term risks of COVID-19. Working Paper 68, International Monetary Fund

Agyarko RK, Bah A, Kavuma EG (2022) Epidemic preparedness and response - Solutions for Africa's capacity needs. allAfrica. https://allafrica.com/stories/202212090263.html. Accessed 4 Apr 2023

Airfinity (2023) A strong pandemic defence system could reduce the chance of another COVID like pandemic in the next ten years from 27.5% to 8%. https://www.airfinity.com/articles/a-strong-pandemic-defence-system-could-reduce-the-chance-of-another-covid. Accessed 5 May 2023

Allianz SE (2020) Are pandemics insurable? https://www.allianz.com/en/press/news/business/insurance/201130_Allianz-are-pandemics-insurable.html. Accessed 29 Mar 2023

American Council of Life Insurers (2021) Life insurers factbook 2021. https://www.acli.com/-/media/acli/files/fact-books-public/2021lifeinsurersfactbook.pdf. Accessed 5 Apr 2023

American Council of Life Insurers (2022) Life insurers factbook 2022. https://www.acli.com/-/media/acli/public/files/factbook/2022lifeinsurersfactbook_v2.pdf. Accessed 5 Apr 2023

APEC (2022) Chair's statement of the 29th APEC finance ministers' meeting. https://www.apec.org/meeting-papers/sectoral-ministerial-meetings/finance/chair-s-statement-of-the-29th-apec-finance-ministers-meeting. Accessed 11 May 2023

Arenz J, Bierwirth M, Schüssler R et al (2023) Grenzen der Versicherbarkeit. In: Arnold R, Berg M, Goecke O et al (eds) Risiko im Wandel: Herausforderung für die Versicherungswirtschaft. Springer Fachmedien Wiesbaden, Wiesbaden, pp 165–187

Artemis (2023a) Catastrophe bond & insurance-linked securities deal directory. https://www.artemis.bm/deal-directory/. Accessed 11 July 2023

Artemis (2023b) Catastrophe bonds & ILS issued by type and year. https://www.artemis.bm/dashboard/catastrophe-bonds-ils-issued-by-type-and-year/. Accessed 11 July 2023

Aviles J, DeField A (2022) Examining event cancellation coverage as COVID lingers. Law360. https://www.huntonak.com/images/content/8/2/v2/82049/examining-event-cancellation-coverage-as-covid-lingers.pdf. Accessed 22 Aug 2023

Barlyn S, Cohn C, Hussain NZ (2021) Global life insurers impose restrictions, worried about long-term pandemic risks. Reuters Media. https://www.reuters.com/article/us-health-coronavirus-insurance-idUSKBN29U0FC. Accessed 05 Apr 2023

Barro RJ, Ursúa JF, Weng J (2020) The coronavirus and the Great Influenza pandemic: Lessons from the "Spanish Flu" for the coronavirus's potential effects on mortality and economic activity. Working Paper 26866, National Bureau of Economic Research

Berliner B (1985) Large risks and limits of insurability. Geneva Papers Risk Insur Issues Practice 10(4):313–329

Bernstein AS, Ando AW, Loch-Temzelides T et al (2022) The costs and benefits of primary prevention of zoonotic pandemics. Sci Adv 8(5):eabl4183

Buchanan J (1972) The Samaritan's dilemma. Reprinted in: Buchanan, JM (1977): Freedom in Constitutional Contract. Texas A&M University Press, Texas, pp 169–185

Centers for Disease Control and Prevention (2018) History of 1918 flu pandemic. https://www.cdc.gov/flu/pandemic-resources/1918-commemoration/1918-pandemic-history.htm. Accessed 23 Aug 2023

Cheney C (2021) How might probability inform policy on pandemics? Metabiota has ideas. Devex. https://www.devex.com/news/how-might-probability-inform-policy-on-pandemics-metabiota-has-ideas-100427. Accessed 1 Feb 2023

Clarey C (2020) To safeguard its future, tennis seeks unity. Even with the French Open. The New York Times. https://www.nytimes.com/2020/04/03/sports/wimbledon-grand-slams-canceled.html. Accessed 28 Apr 2023

Cohn C (2022) COVID-19 loss of $44 bln is 3rd largest catastrophe cost to insurers - Howden. Reuters Media. https://www.reuters.com/markets/commodities/covid-19-loss-44-bln-is-3rd-largest-catastrophe-cost-insurers-howden-2022-01-04/. Accessed 12 Apr 2023

Cohn C, Hussain NZ (2021) Life insurers shift to pre-pandemic norms after COVID vaccine roll-outs. Reuters Media. https://www.reuters.com/business/healthcare-pharmaceuticals/life-insurers-shift-pre-pandemic-norms-after-covid-vaccine-roll-outs-2021-09-02/. Accessed 3 Apr 2023

Cohn C, Hussain NZ (2022) Life insurers adapt pandemic risk models after claims jump. Reuters Media. https://www.reuters.com/business/life-insurers-adapt-pandemic-risk-models-after-claims-jump-2022-01-13/. Accessed 13 Apr 2023

Courbage C, Peter R (2021) On the effect of uncertainty on personal vaccination decisions. Health Econ 30(11):2937–2942

Cummins J, Doherty N, Lo A (2002) Can insurers pay for the "big one"? Measuring the capacity of the insurance market to respond to catastrophic losses. J Bank Financ 26(2):557–583

Da Costa VG, Saivish MV, Santos DER et al (2020) Comparative epidemiology between the 2009 H1N1 influenza and COVID-19 pandemics. J Infect Public Health 13(12):1797–1804

Díaz Llavona C (2023) Impact of Covid-19 on travel and health insurance. In: Muñoz Paredes ML, Tarasiuk A (eds) Covid-19 and insurance. Springer, Cham, pp 271–299

Dieleman C (2022) Klap No Risk biedt entertainmentbranche opnieuw pandemiedekking. AMweb. https://www.amweb.nl/137384/klap-no-risk-biedt-entertainmentbranche-opnieuw-pandemiedekking. Accessed 14 May 2023

Dixon L, Saunders-Medina B (2020) Key questions to ask in designing a pandemic risk insurance program. The RAND Blog. https://www.rand.org/blog/2020/07/key-questions-to-ask-in-designing-a-pandemic-risk-insurance.html. Accessed 28 Mar 2023

Doherty NA, Richter A (2002) Moral hazard, basis risk, and gap insurance. J Risk Insur 69(1):9–24

Dunsavage J (2022) Pandemic fuels growth in captive insurance. Insurance Information Institute. https://www.iii.org/insuranceindustryblog/pandemic-fuels-growth-in-captive-insurance/. Accessed 8 Apr 2023

Echazu L, Nocetti DC (2020) Willingness to pay for morbidity and mortality risk reductions during an epidemic. Theory and preliminary evidence from COVID-19. Geneva Risk Insur Rev 45(2):114–133

EIOPA (2021) 2020 consumer trends report. https://www.eiopa.europa.eu/system/files/2021-01/consumer-trends-report-2020.pdf. Accessed 5 Apr 2023

EIOPA (2021) Insurance and reinsurance stakeholder group: advice on measures to improve the insurability of business interruption risk in light of pandemics. https://www.eiopa.europa.eu/system/files/2021-04/irsg-21-18-insurability-in-pandemics.pdf. Accessed 11 May 2023

European Actuarial Consultative Group (2006) Actuarial reflections on pandemic risk and its consequences. https://actuary.eu/documents/pandemics_web.pdf. Accessed 26 July 2023

European Centre for Disease Prevention and Control (2023) The pathogen SARS-COV-2. https://www.ecdc.europa.eu/en/infectious-disease-topics/z-disease-list/covid-19/facts/pathogen-sars-cov-2. Accessed 11 July 2023

Eurostat (2023) Excess mortality - statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Excess_mortality_-_statistics#Excess_mortality_in_the_EU_between_January_2020_and_February_2023. Accessed 06 May 2023

Evans S (2020) World bank pandemic bonds & swaps triggered, will pay out $195.84m. Artemis. https://www.artemis.bm/news/world-bank-pandemic-bonds-swaps-triggered-will-pay-out-195-84m/. Accessed 28 Mar 2023

Evans S (2022) ARC launches parametric epidemic cover with Munich Re as lead reinsurer. Artemis. https://www.artemis.bm/news/arc-launches-parametric-epidemic-cover-with-munich-re-as-lead-reinsurer/. Accessed 29 Mar 2023

Fatunmole M (2023) WHO raises global COVID-19 deaths to 20 million. The ICIR. https://www.icirnigeria.org/who-raises-global-covid-19-deaths-to-20-million/. Accessed 6 May 2023

FCA (2021) Business interruption insurance test case – Insurer claims data. https://www.fca.org.uk/data/bi-insurance-test-case-insurer-claims-data. Accessed 5 Apr 2023

G7 Health Ministers (2023) G7 shared understanding on enhanced finance-health coordination and PPR financing. https://www.mof.go.jp/english/policy/international_policy/convention/g7/g7_20230513_2.pdf. Accessed 24 Aug 2023

Gavi (2021) Next pandemic. https://www.gavi.org/vaccineswork/tag/next-pandemic. Accessed 06 May 2023

Glennerster R, Snyder CM, Tan BJ (2022) Calculating the costs and benefits of advance preparations for future pandemics. Working Paper 30565, National Bureau of Economic Research

Global Parametrics (2022) Global Parametrics is proud to support a global first for the parametric industry. https://www.linkedin.com/posts/global-parametrics_today-the-government-of-senegal-officially-activity-7018502115849302016-4xxL/. Accessed 14 May 2023

Glynn D, Riley D, Rogers T (2021) Riley on business interruption insurance, 11th edn. Sweet & Maxwell, London

Gollier C (2020) Pandemic economics: optimal dynamic confinement under uncertainty and learning. Geneva Risk Insur Rev 45(2):80–93

Gopinath G (2020) A long, uneven and uncertain ascent. International Monetary Fund. https://www.imf.org/en/Blogs/Articles/2020/10/13/blog-a-long-uneven-and-uncertain-ascent. Accessed 15 Aug 2023

Goyder B, Smythe C (2020) Insider in full: Pandemic product saw no takers before coronavirus: Marsh broker. Slipcase. https://www.slipcase.com/view/inside-in-full-pandemic-product-saw-no-takers-before-coronavirus-marsh-broker. Accessed 04 Apr 2023

Gründl H, Guxha D, Kartasheva A et al (2021) Insurability of pandemic risks. J Risk Insur 88(4):863–902

Gürtler M, Hibbeln M, Winkelvos C (2016) The impact of the financial crisis and natural catastrophes on CAT bonds. J Risk Insur 83(3):579–612
Guy Carpenter (2018) Affirmative vs. silent cyber: an overview. https://www.guycarp.com/content/dam/guycarp-rebrand/pdf/Affirmative%20vs.%20Silent%20Cyber%20An%20Overview.pdf. Accessed 28 Apr 2023
Harrington S, Niehaus G (2003) Risk management and insurance. International edition, McGraw-Hill
Harris TF, Yelowitz A, Courtemanche C (2021) Did COVID-19 change life insurance offerings? J Risk Insur 88(4):831–861
Hartwig R, Gordon R (2020) Uninsurability of mass market business continuity risks from viral pandemics. https://www.banking.senate.gov/imo/media/doc/Hartwig%20Testimony%20Addendum%207-22-21.pdf. Accessed 04 Apr 2023
Hartwig R, Niehaus G, Qiu J (2020) Insurance for economic losses caused by pandemics. Geneva Risk Insur Rev 45(2):134–170
Hillier R (2022) The legal challenges of insuring against a pandemic. In: Del Boado-Penas MC, Eisenberg J, Sule S (eds) Pandemics: Insurance and social protection. Springer Actuarial Ser, Springer, Cham, pp 267–286
Holmes EC (2022) The ecology of viral emergence. Ann Rev Virol 9(1):173–192
Insurance Europe (2021) European insurance in figures - 2020 data. https://insuranceeurope.eu/publications/2569/european-insurance-in-figures-2020-data. Accessed 7 Apr 2023
Insurance Europe (2022) European insurance: Preliminary figures 2021. https://www.insuranceeurope.eu/publications/2674/european-insurance-preliminary-figures-2021/download/Preliminary+figures%202021.pdf. Accessed 5 Apr 2023
International Association of Insurance Supervisors (2022) The role of insurance supervisors in multi-stakeholder approaches to address pandemic protection gaps. https://www.iaisweb.org/uploads/2022/09/220929-The-role-of-insurance-supervisors-in-multi-stakeholder-approaches-to-address-pandemic-protection-gaps.pdf. Accessed 14 May 2023
Ippolito G, Puro V, Piselli P (2015) Ebola in West Africa: who pays for what in the outbreak? New Microbiol 38(1):1–3
Japsen B (2021) Coming soon for the unvaccinated: A $50 monthly paycheck deduction from your employer. Forbes. https://www.forbes.com/sites/brucejapsen/2021/08/08/coming-soon-for-the-unvaccinated-a-50-monthly-paycheck-deduction-from-your-employer/?sh=4a77400a47ae. Accessed 12 July 2023
Jerry RH (2023) Understanding parametric insurance: A potential tool to help manage pandemic risk. In: Muñoz Paredes ML, Tarasiuk A (eds) Covid-19 and insurance. Springer, Cham, pp 17–62
Josephs L (2022) Delta ends $200 monthly health insurance surcharge on unvaccinated employees after Covid cases drop. CNBC. https://www.cnbc.com/2022/04/13/delta-ends-200-health-insurance-surcharge-on-unvaccinated-employees.html. Accessed 11 July 2023
Karten W (1972) Zum Problem der Versicherbarkeit und zur Risikopolitik des Versicherungsunternehmens–betriebswirtschaftliche Aspekte. Zeitschrift für die gesamte Versicherungswissenschaft 61:279–299
Karten W (1997) How to expand the limits of insurability. Geneva Papers Risk Insur Issues Practice 22(4):512–522
Karten W, Nell M, Richter A et al (2018) Risiko und Versicherungstechnik: Eine ökonomische Einführung. Springer Gabler, Wiesbaden
Kawai Y (2022) Idome “protection gap”, kyodai risuku sonaeru keiei wo. Nikkei Financial. https://financial.nikkei.com/article/DGXZQOUB120MC0S2A810C2000000. Accessed 14 May 2023
Kekulé A (2017) Controlling disease outbreak risks for industry operations in sub-Saharan Africa
Kelly H, Peck HA, Laurie KL et al (2011) The age-specific cumulative incidence of infection with pandemic influenza H1N1 2009 was similar in various countries prior to vaccination. PLoS One 6(8):e21828
Keogh-Brown MR, Smith RD (2008) The economic impact of SARS: how does the reality match the predictions? Health Policy (Amsterdam, Netherlands) 88(1):110–120
Klein RW, Weston H (2020) Government insurance for business interruption losses from pandemics: an evaluation of its feasibility and possible frameworks. Risk Manag Insur Rev 23(4):401–440
Koide T (2022) Insurers to limit COVID-19 benefit payouts mainly to elderly people. Asahi Shimbun. https://www.asahi.com/ajw/articles/14707235. Accessed 1 Aug 2023
Kraut G (2014) A fair pool sharing mechanism for illiquid catastrophe risk markets. Munich Risk and Insurance Center Working Papers 19
Kraut G (2015) Essays on current topics in life insurance, Münchener Reihe, vol 70, 1st edn. Verlag Versicherungswirtschaft Karlsruhe GmbH
Kraut G, de Kuiper R (2021) Epidemic and pandemic risk transfer solutions and options for public sector support. Working Paper 41, Munich Risk and Insurance Center
Kraut G, Richter A (2015) Insurance regulation and life catastrophe risk: Treatment of life catastrophe risk under the SCR standard formula of Solvency II and the necessity of partial internal models. Geneva Papers Risk Insur Issues Practice 40(2):256–278
Kunreuther H (1987) Problems and issues of environmental liability insurance. Geneva Papers Risk Insur 12(44):180–197

Lakdawalla D, Zanjani G (2012) Catastrophe bonds, reinsurance, and the optimal collateralization of risk transfer. J Risk Insur 79(2):449–476
Lane M, Beckwith R (2020) World Bank pandemic bonds redux and quarterly market performance report - Q2 2020. http://www.lanefinancialllc.com/images/stories/Publications/2020-06-30%20Quarterly%20Review.pdf. Accessed 11 May 2023
Lee JW, McKibbin WJ (2004) Estimating the global economic costs of SARS. In: Lee JW, McKibbin WJ (eds) Learning from SARS: preparing for the next disease outbreak: workshop summary. National Academies Press (US)
Litzenberger RH, Beaglehole DR, Reynolds CE (1997) Assessing catastrophe reinsurance-linked securities as a new asset class. J Portfolio Manag 23(5):76–86
LMU (2023) Viral zoonoses can cause a pandemic. https://www.story.lmu.de/en/zoonoses-pandemics/. Accessed 11 Apr 2023
Louaas A, Picard P (2023) A pandemic business interruption insurance. Geneva Risk Insur Rev 48(1):1–30
MacMinn R, Richter A (2018) The choice of trigger in an insurance linked security: the mortality risk case. Insur Math Econ 78:174–182
Madhav N, Oppenheim B, Gallivan M et al (eds) (2017) Disease control priorities: improving health and reducing poverty, 3rd edn. The International Bank for Reconstruction and Development/The World Bank, Washington, DC
Marani M, Katul GG, Pan WK et al (2021) Intensity and frequency of extreme novel epidemics. Proceedings of the National Academy of Sciences of the United States of America 118(35)
Maruf R (2021) These companies are paying their employees to receive the Covid-19 vaccine. CNN. https://edition.cnn.com/2021/03/24/business/covid-vaccine-incentives-companies/index.html. Accessed 11 July 2023
Medders LA, Schwarcz SL (2022) Securitizing pandemic-risk insurance. Risk Manag Insur Rev 25(4):551–583
Meltzer MI, Cox NJ, Fukuda K (1999) The economic impact of pandemic influenza in the United States: priorities for intervention. Emer Infect Diseases 5(5):659–671
Messenger H (2021) From McDonald's to Goldman Sachs, here are the companies mandating vaccines for all or some employees. NBC News. https://www.nbcnews.com/business/business-news/here-are-companies-mandating-vaccines-all-or-some-employees-n1275808. Accessed 12 July 2023
Metabiota (2018) Infectious disease outbreaks can inflict enormous social and economic disruption. https://metabiota.com/sites/default/files/presentation_files/Metabiota%20Pathogen%20Sentiment%20Index%20Flyer_Apr%202018.pdf. Accessed 6 May 2023
Milman O (2020) Trump administration cut pandemic early warning program in September. The Guardian. https://www.theguardian.com/world/2020/apr/03/trump-scrapped-pandemic-early-warning-program-system-before-coronavirus. Accessed 6 May 2023
Munich Re (2017) Cyber update: Global ransomware attack: WannaCry. https://www.munichre.com/content/dam/munichre/contentlounge/website-pieces/documents/HSBEI-1754.pdf/_jcr_content/renditions/original./HSBEI-1754.pdf. Accessed 26 Apr 2023
Munich Re (2020) Re-imagining resilience in a post pandemic world. https://reliefweb.int/report/world/re-imagining-resilience-post-pandemic-world. Accessed 12 May 2023
Munich Re (2023a) Epidemic Risk Solutions. https://www.munichre.com/en/solutions/for-industry-clients/epidemic-risk-solutions.html. Accessed 14 May 2023
Munich Re (2023b) Market presentation | Epidemic Risk Solutions
Munich Re (2023c) Pilot transactions for Epidemic Risk Markets Platform. https://www.linkedin.com/feed/update/urn:li:activity:6997876235691290624/. Accessed 14 May 2023
National Association of Insurance Commissioners (2021) U.S. health insurance industry | 2020 annual results. https://content.naic.org/sites/default/files/inline-files/2020-Annual-Health-Insurance-Industry-Analysis-Report.pdf. Accessed 12 Apr 2023
National Association of Insurance Commissioners (2022) U.S. health insurance industry | 2021 annual results. https://content.naic.org/sites/default/files/2021-Annual-Health-Insurance-Industry-Analysis-Report.pdf. Accessed 12 Apr 2023
National Association of Insurance Commissioners (2023a) U.S. health insurance industry | 2022 annual results. https://content.naic.org/sites/default/files/inline-files/Health%202022%20Annual%20Industry%20Report.pdf. Accessed 31 July 2023
National Association of Insurance Commissioners (2023b) U.S. life and A&H insurance industry | 2022 annual results. https://content.naic.org/sites/default/files/inline-files/2022%20Life%20Annual%20Industry%20Commentary_1.pdf. Accessed 31 July 2023
Nebolsina E (2021) The impact of the Covid-19 pandemic on the business interruption insurance demand in the United States. Heliyon 7(11):e08357
Nell M (1993) Versicherungsinduzierte Verhaltensänderungen von Versicherungsnehmern: Eine Analyse der Substitutions-, Moral Hazard- und Markteffekte unter besonderer Berücksichtigung der Krankenversicherung. Karlsruhe

Nell M, Richter A, Schiller J (2009) When prices hardly matter: incomplete insurance contracts and markets for repair goods. Eur Econ Rev 53(3):343–354
OECD (2021) Responding to the COVID-19 and pandemic protection gap in insurance. https://read.oecd-ilibrary.org/view/?ref=133_133327-3tdsqdiu7y&title=Responding-to-the-COVID-19-and-pandemic-protection-gap-in-insurance. Accessed 2 Aug 2023
OECD (2023) Draft recommendation on building financial resilience to disaster risk. https://www.oecd.org/daf/fin/insurance/Recommendation-Building-Financial-Resilience-Disaster-Risks-for-comment%20.pdf. Accessed 25 July 2023
Office for National Statistics (2022) Excess deaths in England and Wales. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/excessdeathsinenglandandwales/march2020todecember2021. Accessed 6 May 2023
Phoonphongphiphat A (2022) Thai insurers sink into bankruptcy due to COVID claims. Nikkei Asia. https://asia.nikkei.com/Spotlight/Coronavirus/Thai-insurers-sink-into-bankruptcy-due-to-COVID-claims. Accessed 1 Aug 2023
Porta M (2014) A dictionary of epidemiology, 6th edn. Oxford University Press, Oxford
Qiu J (2020) Pandemic risk: Impact, modeling, and transfer. Risk Manag Insur Rev 23(4):293–304
Ralph O (2020) Business disruption insurance: can it help with coronavirus? Financial Times. https://www.ft.com/content/1d359ef3-9104-4441-a039-82c6b9683ad2. Accessed 26 Apr 2023
Reuters (2021) GDV schlägt Pflichtversicherung für Pandemie-Folgen vor. Reuters Media. https://www.reuters.com/article/deutschland-gdv-pandemie-idDEKBN29P1BU. Accessed 27 July 2023
Reuters Staff (2021) Next pandemic could be more lethal than COVID, vaccine creator says. https://www.reuters.com/business/healthcare-pharmaceuticals/next-pandemic-could-be-more-lethal-than-covid-oxford-vaccine-creator-says-2021-12-06/. Accessed 9 May 2023
Richardson M (2023) Pandemic parametric risk-transfer solutions for captives. Captive Review. https://captivereview.com/features/pandemic-parametric-risk-transfer-solutions-for-captives/. Accessed 24 July 2023
Richter A, Wilson TC (2020) Covid-19: Implications for insurer risk management and the insurability of pandemic risk. Geneva Risk Insur Rev 45(2):171–199
RMS (2021) The next pandemic risk: Risk drivers for life, health, and business interruption insurance. https://www.rms.com/sites/default/files/content/rms-files/pdf-files/INTL_Pandemic_Drivers_WP2.pdf. Accessed 11 May 2023
Robert Koch-Institut (2021) Influenza-Wochenbericht. https://influenza.rki.de/Wochenberichte/2020_2021/2021-20.pdf. Accessed 7 May 2023
Rohr JR, Barrett CB, Civitello DJ et al (2019) Emerging human infectious diseases and the links to global food production. Nat Sustain 2(6):445–456
Salanié F, Treich N (2020) Public and private incentives for self-protection. Geneva Risk Insur Rev 45(2):104–113
Schanz KU (2020) An investigation into the insurability of pandemic risk. The Geneva Association. https://www.genevaassociation.org/sites/default/files/research-topics-document-type/pdf_public/insurability_report_web.pdf. Accessed 23 Apr 2023
Schiller F (2022) An actuary's opinion: How to get through a pandemic. In: Del Boado-Penas MC, Eisenberg J, Sule S (eds) Pandemics: Insurance and social protection. Springer Actuarial Ser, Springer, Cham, pp 287–298
Schmitt A, Spaeter S (2023) Providing pandemic business interruption coverage with double trigger cat bonds. Geneva Papers Risk Insur Issues Practice 48(3):687–713
Silverman S, Li J, Wang C (2022) How does our experience with COVID-19 impact the life insurance industry going forward? Milliman. https://us.milliman.com/en/insight/how-does-covid-19-impact-the-life-insurance-industry-going-forward. Accessed 2 Apr 2023
Smetters KA (2020) Stay-at-home orders and second waves: a graphical exposition. Geneva Risk Insur Rev 45(2):94–103
Smitham E, Glassman A (2021) The next pandemic could come soon and be deadlier. Center for Global Development. https://www.cgdev.org/blog/the-next-pandemic-could-come-soon-and-be-deadlier. Accessed 1 Feb 2023
SOA (2007) Potential impact of pandemic influenza on the U.S. life insurance industry. https://www.soa.org/globalassets/assets/files/research/projects/resrch-li-rep-pan-life.pdf. Accessed 26 July 2023
Sompo International (2021) Sompo Global Risk Solutions to offer epidemic and pandemic protection. https://www.sompo-intl.com/media-center/sompo-global-risk-solutions-to-offer-epidemic-and-pandemic-protection/. Accessed 14 May 2023
Spaeter S (2021) How to reconcile pandemic business interruption risk with insurance coverage. Working Papers of BETA 2021-18, Bureau d'Economie Théorique et Appliquée, UDS, Strasbourg
Stefan C, Kiggell C, DeGrauw A (2023a) The role of disaster risk finance in tackling pandemic and epidemic risk. Centre for Disaster Protection. https://static1.squarespace.com/static/61542ee0a87a394f7bc17b3a/t/64e6539a89f8cc279f4a4352/1692816286285/The+role+of+disaster+risk+finance+in+tackling+pandemic+and+epidemic+risk+2.pdf. Accessed 24 Aug 2023

Stefan C, Talbot T, Glassman A et al (2023b) The next pandemic: if we can't respond, we're not prepared. Center for Global Development. https://www.cgdev.org/publication/next-pandemic-if-we-cant-respond-were-not-prepared. Accessed 10 Mar 2023
Swiss Re (2017) Cyber: getting to grips with a complex risk. Sigma No 1/2007. https://www.swissre.com/institute/research/sigma-research/sigma-2017-01.html. Accessed 26 Apr 2023
Swiss Re (2021a) Vita: ILS and risk transfer - Annual report 2021. https://reports.swissre.com/2021/strategy-in-action/risk-transfer/vita-ils-and-risk-transfer. Accessed 20 Apr 2023
Swiss Re (2021b) World insurance: the recovery gains pace. https://www.swissre.com/institute/research/sigma-research/sigma-2021-03.html. Accessed 03 Aug 2023
The Economist (2023) The pandemic's true death toll. https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates. Accessed 31 July 2023
The Independent Panel for Pandemic Preparedness and Response (2021) COVID-19: Make it the last pandemic. https://theindependentpanel.org/wp-content/uploads/2021/05/COVID-19-Make-it-the-Last-Pandemic_final.pdf. Accessed 6 May 2023
The Insurer (2023) FCA: UK Covid-19 BI payouts top £1.51bn. https://www.theinsurer.com/analysis/fca-uk-covid-19-bi-payouts-top-1-51bn/. Accessed 3 Apr 2023
Turner H (2020) Wimbledon's pandemic insurance coverage results in $141m payout. PropertyCasualty360. https://www.propertycasualty360.com/2020/04/10/wimbledons-pandemic-insurance-coverage-results-in-141m-payout/?slreturn=20230330094327. Accessed 29 Apr 2024
United Kingdom Government Department of Health and Social Care (2011) UK influenza pandemic preparedness strategy 2011. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/213717/dh_131040.pdf. Accessed 6 May 2023
United Nations Development Programme (2017) A socio-economic impact assessment of the Zika virus in Latin America and the Caribbean: with a focus on Brazil, Colombia and Suriname. Research report. United Nations Development Programme. https://global.undp.acsitefactory.com/sites/g/files/zskgke326/files/publications/UNDP-RBLAC-Zika-English-WEB.pdf. Accessed 9 May 2023
University of Pennsylvania (2023a) Covid coverage litigation tracker. https://cclt.law.upenn.edu/. Accessed 11 July 2023
University of Pennsylvania (2023b) Trial court rulings on the merits in business interruption cases. https://cclt.law.upenn.edu/judicial-rulings/. Accessed 13 Apr 2023
Weintraub K (2023) As COVID turns 3, experts worry where the next pandemic will come from - and if we'll be ready. USA TODAY. https://eu.usatoday.com/story/news/health/2023/01/01/covid-anniversary-next-pandemic-expert-concern/10847848002/. Accessed 09 May 2023
Wells K (2023) European re/insurer solvency ratios continue post-pandemic recovery: JP Morgan. Reinsurance News. https://www.reinsurancene.ws/european-re-insurer-solvency-ratios-continue-post-pandemic-recovery-jp-morgan/. Accessed 4 Apr 2023
Wilder-Smith A, Osman S (2020) Public health emergencies of international concern: a historic overview. J Travel Med 27(8):taaa227
Willis Towers Watson (2020) Scenario analysis of the COVID-19 pandemic: Analysis of key classes in the US and UK P&C insurance industries. https://www.wtwco.com/-/media/wtw/insights/2020/05/scenario_analysis_of_the_covid-19_pandemic-final.pdf?modified=20200528153942. Accessed 24 Apr 2023
Wolfrom L (2022) Could insurance provide an alternative to fiscal support in crisis response? OECD Working Papers on Fiscal Federalism 40
World Bank (2014) The economic impact of the 2014 Ebola epidemic: Short- and medium-term estimates for West Africa. https://elibrary.worldbank.org/doi/abs/10.1596/978-1-4648-0438-0. Accessed 9 May 2023
World Bank (2020) Fact sheet: Pandemic Emergency Financing Facility. https://www.worldbank.org/en/topic/pandemics/brief/fact-sheet-pandemic-emergency-financing-facility. Accessed 8 Apr 2023
World Bank (2022) Putting pandemics behind us: Investing in one health to reduce risks of emerging infectious diseases. https://openknowledge.worldbank.org/server/api/core/bitstreams/956a58be-ddd8-572f-8aac-df5ab453d7b2/content. Accessed 6 May 2023
World Health Organization (2015) Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003. Accessed 3 Apr 2023
World Health Organization (2017) Pandemic influenza risk management: a WHO guide to inform and harmonize national and international pandemic preparedness and response. https://apps.who.int/iris/bitstream/handle/10665/259893/WHO-WHE-IHM-GIP-2017.1-eng.pdf;jsessionid=FF0E44DE342CCEF9F0A31E1EFB14C8E8?sequence=1. Accessed 5 May 2023
World Health Organization (2019) Influenza update N° 357. https://cdn.who.int/media/docs/default-source/influenza/influenza-updates/2019/2019_12_20_surveillance_update_357.pdf. Accessed 7 May 2023
World Health Organization (2020a) Influenza update N° 376. https://cdn.who.int/media/docs/default-source/influenza/influenza-updates/2020/2020_09_14_surveillance_update_376.pdf. Accessed 7 May 2023

World Health Organization (2020b) Statement on the second meeting of the International Health Regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov). Accessed 14 May 2023

World Health Organization (2020c) WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 14 May 2023

World Health Organization (2023a) Bi-weekly influenza situation update. https://cdn.who.int/media/docs/default-source/wpro---documents/emergency/surveillance/seasonal-influenza/influenza-20230426.pdf. Accessed 7 May 2023

World Health Organization (2023b) Disease outbreak news. https://www.who.int/emergencies/disease-outbreak-news. Accessed 1 Aug 2023

World Health Organization (2023c) MERS outbreaks. https://www.emro.who.int/health-topics/mers-cov/mers-outbreaks.html. Accessed 3 Apr 2023

World Health Organization (2023d) WHO coronavirus (COVID-19) dashboard. https://covid19.who.int/?mapFilter=deaths. Accessed 26 July 2023

World Health Organization (2023e) WHO to identify pathogens that could cause future outbreaks and pandemics. https://www.who.int/news/item/21-11-2022-who-to-identify-pathogens-that-could-cause-future-outbreaks-and-pandemics. Accessed 22 Aug 2023

World Health Organization (2023f) World health statistics 2023: monitoring health for the SDGs, sustainable development goals. https://www.who.int/publications/i/item/9789240074323. Accessed 26 July 2023

World Health Organization (2024) WHO Director-General declares mpox outbreak a public health emergency of international concern. https://www.who.int/news/item/14-08-2024-who-director-general-declares-mpox-outbreak-a-public-health-emergency-of-international-concern. Accessed 7 Oct 2024

Yang YT, Chen B (2022) Imposing a health insurance surcharge on the unvaccinated. Am J Manag Care 28(7)

Yanovskiy M, Socol Y (2022) Are lockdowns effective in managing pandemics? Int J Environ Res Public Health 19(15):9295

Ziady H (2020) It's too late for this pandemic. But everyone wants insurance against the next one. CNN Business. https://edition.cnn.com/2020/03/19/business/pandemic-insurance-coronavirus/index.html. Accessed 9 Apr 2023

Part IV

Life Insurance and Social Insurance

Life Insurance Financial Products

Nadine Gatzert and Hato Schmeiser

1 Introduction: The Worldwide Life Insurance Market 2001–2021

The main drivers of the life insurance market and innovative products can be identified by looking at the history and reasons for the growth and decline of premium volumes in recent years.[1] The following numbers in the text refer to Fig. 1. During this 20-year period, the global life insurance premium volume of USD1.439 trillion in 2001 grew to USD2.998 trillion in 2021.

Looking at the situation in 2001, life insurers in developed countries were particularly challenged by stock market losses and the economic slowdown, which also led to a sharp decline in demand for unit-linked life insurance products (Swiss Re 2002). However, the lower sales of unit-linked products were partially offset by an increase in the number of contracts sold with guaranteed returns and private annuities, so that overall life insurance premiums declined by only 2.7%. In 2002, premiums increased by 3.0%, mainly due to higher sales in the emerging markets (+12.7%), while the industrialized countries grew by +1.9% after the severe losses of 2001.

As in previous years, growth was mainly driven by the deterioration of social security systems and the resulting increase in demand for pensions and annuity products (Swiss Re 2003). However, insurers still had to write down the value of corporate bonds substantially due to lower ratings, which put pressure on their balance sheets.

After improvements in 2002, premiums in the developed world fell in real terms by −1.7% in 2003, mainly due to adverse developments in the United States and the United Kingdom, while emerging markets again recorded growth of +6.6%. Reasons for the reluctance of customers to

[1] The global life insurance market can also be well examined based on the heterogeneity in insurance penetration and density (see, e.g., Ma and Pope 2008).

N. Gatzert (✉)
Chair for Insurance Economics and Risk Management, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
e-mail: nadine.gatzert@fau.de

H. Schmeiser
Chair for Risk Management and Insurance, University of St. Gallen, St. Gallen, Switzerland
e-mail: hato.schmeiser@unisg.ch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_10

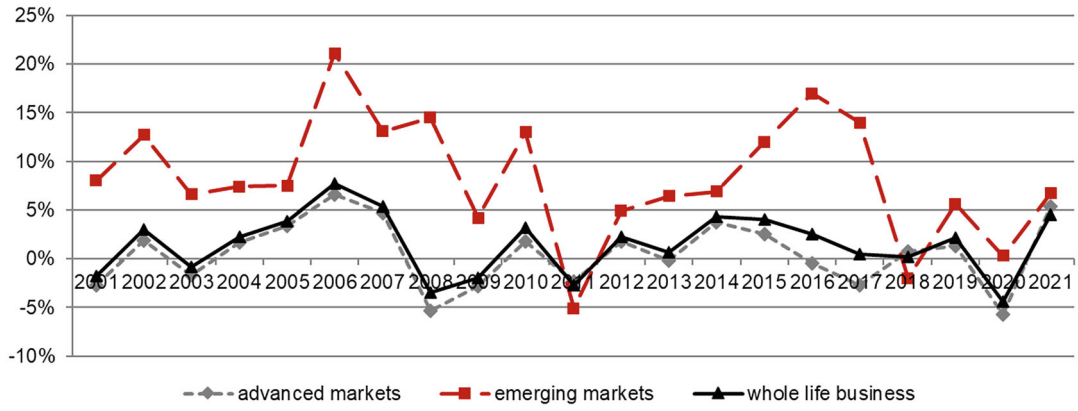


Fig. 1 Real premium growth in the global life insurance industry from 2000 to 2021 (Swiss Re 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022)

purchase traditional savings products included the reduction in surplus participation rates due to losses on investments (falling stock markets and low interest rates). This affected the UK life insurance industry in particular, as with-profit policy payouts and terminal bonuses, which are optional and can generally be reduced by the insurer, if necessary, were reduced. At the same time, customers were still not sufficiently convinced by unit-linked products due to recent financial market experiences and doubts about the outlook for future capital market developments (Swiss Re 2004). Term life and pension products continued to sell well, partly due to increased risk awareness and demographic developments in general.

From 2004 to 2007, life insurance business considerably recovered and growth rates in the life insurance industry accelerated from +2.3% to +3.9%, up to +7.7% in 2006, the highest growth rate since 2000, and to +5.4% in 2007 (Swiss Re 2005, 2006, 2007, 2008). This favorable development was mainly driven by pension reforms in Western Europe, tax advantages for pension products and higher profitability. In addition, a stronger capital base, together with positive macroeconomic indicators and a strongly rising stock market, supported this development, which had a particularly positive impact on unit-linked business. Swiss Re (2007) points to the housing market boom and successes in bancassurance as other key drivers. In some countries, however, specific adverse developments limited growth rates. For example, the German market experienced lower rates in 2005 due to tax reforms that reduced the tax advantages of traditional life insurance products, while growth in the US market was hampered by higher short-term interest rates that made investing in life insurance less attractive. On the other hand, demographic trends, and the aging of the population, as well as government incentives to shift from public to private pension schemes, had a positive impact, leading to a shift from traditional life insurance to annuities and a trend toward single premiums.

When the second financial crisis of the decade hit the markets in 2008, and fears of another recession emerged, sales of unit-linked policies in particular fell, leading to overall premium growth of −3.5% (Swiss Re 2009). In the US market, for example, unit-linked variable annuity business fell by a double-digit percentage (Swiss Re 2009, p. 10). In addition, capital market losses and high costs for investment guarantees embedded in life insurance contracts negatively impacted life insurers' profitability and solvency capital base, highlighting the importance of appropriate pricing and risk management of embedded guarantees in traditional and innovative life insurance contracts. The overall market recovered in 2009 as a result of measures taken by central banks and other institutions to stabilize the credit market and support the economy (Swiss Re 2010). However, growth

rates still fell by 2% as the United States (−15%) and United Kingdom (−12%) markets continued to react strongly, recording double-digit declines in life insurance business, while in Germany and France, sales of traditional life insurance policies with guarantees increased because of increased uncertainty in the financial markets following the crisis and low interest rates. In 2010, a real growth rate of +3.2% was achieved worldwide, in line with the recovery of the global economy and the consequent increase in demand, comprising an increase of +1.8% in developed countries and +13% in emerging markets (Swiss Re 2011). As in the previous year, premium volumes declined in the United States and the United Kingdom, but at a slower rate. The life insurance industry continued to recover from the 2008 financial crisis, also driven by lower surrender and lapse rates and higher investment returns. The global life market declined again in 2011 by −2.7% (developed markets −2.3%, emerging markets −5.1%) (Swiss Re 2012), but with strong differences between countries. Western Europe, with a "resurgent sovereign debt crisis" and "contagion to large peripheral European economies such as Spain and Italy" (p. 5), showed the sharpest decline in premiums at −9.8%, while the United States, Japan, and several Asian markets showed growth. The strong decline in emerging markets was driven by China and India due to new regulatory restrictions on distribution channels (Swiss Re 2012, p. 3). In addition, the low interest rate environment resulting from expansionary monetary policies is cited as a reason for the pressure on insurers' investment returns (p. 6). Low interest rates and uncertainty about the Euro were also a factor in 2012, when overall global premium growth of 2.3% was largely supported by emerging markets and advanced Asia, as well as strong group annuity sales in the United States (Swiss Re 2013, pp. 3, 8). Growth slowed again in advanced markets in 2013, before rebounding to 3.8% in 2014 and 2.5% in 2015, before slowing to −0.5% in 2016. In contrast, strong growth was seen particularly in emerging markets, up to 12% in 2015 and as much as 17% in 2016 (Swiss Re 2014, 2015, 2016, 2017).

The negative trend in advanced markets was also observed in 2017 (−2.7%), despite generally rising GDPs due to low interest rates reducing the attractiveness of classic life products with guarantees, while emerging markets still grew by 14%, driven mainly by China (Swiss Re 2018). This changed to −2% in 2018 due to unfavorable developments in Latin America, Africa and China, before showing premium growth of 5.6% in 2019, with advanced markets growing moderately by 0.8% and 1.3% respectively (Swiss Re 2019, 2020). In 2021, markets globally recover from the COVID-19 crisis with premium growth of 4.5% (5.4% in advanced markets, 6.7% in emerging markets) (Swiss Re 2022), after a decline of 4.4% in 2020 due to lower demand for savings products, which account for 81% of the global life portfolio (Swiss Re 2021, p. 2). In particular, Swiss Re (2021, p. 1) highlights "positive paradigm shifts for insurance" resulting from a sharp increase in risk awareness on the part of clients, which will support demand for risk protection products from individuals (although this may vary by country). Swiss Re (2022, p. 13) overall sees a strong resilience of the insurance sector, as COVID-19 did not lead to as sharp a decline in premium growth as the financial crisis of 2008/2009. Looking at the environment in recent years, on the one hand, periods of uncertainty and heightened risk awareness can increase demand for insurance protection, and higher interest rates offer opportunities for reinvestment. On the other hand, sharply rising inflation rates, driven by "high levels of fiscal stimulus pumped into the economy at the height of the COVID-19 crisis," as well as global supply chain disruptions and "soaring energy and commodity prices" following the start of the war in Ukraine, may lead to higher premiums, and equity markets as well as higher credit risks may weigh on the asset side of insurers' balance sheets (Swiss Re 2022, p. 2; 13).

Global life insurance premium volume in 2021 will be around USD2,998 billion (data taken from Swiss Re 2022). The per capita premium is about USD382 worldwide. The premium volume for the Americas is around USD740 billion, for Europe, the Middle East and Africa around USD1,128 billion and for the Asia-Pacific market around USD1,129 billion.

Developments over the last 20 years highlight the main drivers of growth in the life insurance industry, in particular macroeconomic factors, and financial market conditions (uncertainty in equity markets and generally low interest rates). In addition, other main drivers are regulation, including distribution channels, and tax policy regarding the tax advantages of life insurance products, government policies on public and private pension schemes, with a shift from defined benefit to defined contribution plans, and demographic trends, with an aging population in general, which drive the demand for different types of life insurance products. This also implies the need to offer appropriate products that take account of the changing situation. In particular unit-linked or hybrid policies with innovative (investment) guarantee concepts that take account of policyholders' fears of capital market uncertainty and their potential mistrust of traditional pure unit-linked products. In this context, products that combine traditional and unit-linked designs will become increasingly important, allowing policyholders to participate in positive market developments while ensuring downside protection through a guarantee that is priced and secured by the insurer. Overall, sustainability aspects are becoming increasingly important, for example, by allowing policyholders to choose a sustainable investment fund underlying a unit-linked product (see, e.g., Gatzert and Kraus 2023). Therefore, a major focus of the remainder of this chapter will be on the presentation of different product types, ranging from traditional to newer financial products, as well as the modeling and valuation of such innovative product types with different variations of guarantees from the perspective of policyholders and insurers.

2 From Traditional to New Life Insurance Financial Products

The following section provides an overview of the transition from traditional life insurance products, which are still of high relevance for European and some Asian markets (especially China and India), to new life insurance financial products, which are often more investment-oriented for the accumulation and decumulation phases and are playing an increasing role in most markets.

2.1 Life Insurance with Focus on Creating a Bequest in the Event of the Policyholder's Premature Death

Term, Whole, and Universal Life Insurance

Term life insurance contracts are of relevance in all markets and pay a death benefit, the so-called face value, to the beneficiary of the contract in case the insured person dies during a contractually fixed contract term. Premiums can be a single upfront or a regular payment (see Bowers et al. 1997). The aim of term life insurance policies is to provide coverage for financial responsibilities of the policyholder such as mortgages or other debts, daily living belongings, or education for dependent family members. Besides this basic form, various variations of term life insurance contracts can be found in insurance practice. For instance, the death benefit may increase or decrease in time, partly depending on a participating mechanism (see also next subsection), or the policy may be annually renewable for an ex-ante fixed premium level. In addition, in case of whole or universal life insurance contracts, as typically offered in the United States, the contracts run lifelong and thus provide a payoff with certainty when the insured person dies.[2] Term or whole life insurance products are

[2] Regarding the introduction of universal life products as a successor of whole life, we refer to Doll (1999), who also discusses driving environmental factors for demand such as increasing inflation, rising taxes, changes in the workforce, and rising interest rates.

typically attached to other types of insurance contracts as well, for example, unit-linked or traditional participating life insurance contracts.

2.2 Life Insurance Products Used Primarily for Asset Accumulation

Traditional Participating Life Insurance

Traditional participating life insurance policies (with-profit policies) include a term life insurance policy, as described in the previous section, and a savings component. The aim of the savings process is usually to improve personal retirement provision. The death benefit paid to a beneficiary (heirs or creditor) only comes into play if the policyholder dies before the end of the policy term.

The premiums paid into the policy can therefore be split into two components. In the following, we will only focus on the savings component and the corresponding premium and payout. The savings component is generally characterized by a guaranteed minimum interest rate paid on the savings premium and participation in the company's surplus calculated according to a predefined surplus distribution mechanism. In addition, an (optional) terminal bonus payment may be provided.

The insurer invests the premiums of all policyholders together with the contributions of the shareholders in the capital market, so that it is not possible to identify which part of the assets belongs to an individual policyholder. The traditional contract design is mainly determined by three key contract parameters. The first is a guaranteed minimum annual interest rate, g, which must be earned on the policyholder's reserves. In some countries, this minimum rate is set by law and changes periodically depending on capital market conditions. In German-speaking countries, for example, this rate is guaranteed for the life of the policy, which represents a substantial value to the policyholder and a significant risk to the insurer if it is not adequately priced and hedged. The value of these guarantees is particularly evident under Solvency II, where significant risk capital must be held depending on the interest rate assumptions. The guaranteed interest rate in traditional products can therefore be varied by adjusting the guarantee after a certain period during the life of the contract according to the average interest rate level (as is done in France, for example) or by linking the guarantee directly to an average yield on government bonds.

The second parameter is the annual surplus participation rate, which is generally regulated as well to ensure that policyholders receive an adequate share of insurer's investment earnings (in book values), as these are mainly generated based on the policyholders' savings and excess premiums. The part of the surplus not credited to the policyholder's account is available for the equityholders of the insurer, who have to be compensated for the provided capital used to secure the investment guarantee. In times of positive market developments, policyholders thus participate in the insurer's investment returns that are higher than the guaranteed interest rate. One major problem in this regard is the complexity and opaqueness of the surplus participation scheme that is generally not really transparent and may not be well-defined in mathematical terms. Often, the surplus distribution is based on a smoothing scheme that reduces the volatility of investment returns by buffering the surplus in a buffer account, which increases the stability of the surplus payment through actuarial smoothing but also via accounting smoothing through building and resolving hidden reserves on the asset side (e.g., Maurer et al. 2013). In the case of so-called cliquet-style guarantees, as soon as the annual surplus participation is annually credited to the policyholder's account, it becomes part of the guarantee and thus at least earns the guaranteed interest rates in the following years (see, e.g., Grosen and Jørgensen 2000). As an alternative to these cliquet-style guarantees, participating life insurance contracts may feature a point-to-point guarantee, where only a minimum payment at maturity is guaranteed (see, e.g., Grosen and Jørgensen 2002). In general, particularly the type of smoothing scheme used can have a considerable impact on the contract value (see, e.g., Guillén et al. 2006). The last parameter,

which has an influence on the liabilities, is the terminal bonus distribution. It can optionally be added to the policyholder's account at maturity, for example, depending on the initial contribution rate (see, e.g., Gatzert and Kling 2007).

Participating policies further typically feature numerous implicit options, including settlement options (lump sum, fixed income, etc.), premium payment options, the surrender option, the flexible expiration option, and many more (see Trieschmann et al. 2005).[3] In many cases, closed-form solutions for the payoff to the client cannot be derived due to high complexity and opacity, and hence, standard option pricing formulas cannot be used. Therefore, simulation techniques are used in insurance practice. For example, the combination of interest rate guarantees with different premium payment options under stochastic assets and a stochastic term structure in participating life insurance contracts has been evaluated by Chang and Schmeiser (2022), who investigate two aspects: (a) an approximation of the optimal exercise strategy from the policyholder's point of view and (b) the potential risk of the provider offering such guarantees facing liquidity risk.

Unit-Linked Products

Like participating life insurance contracts, unit-linked policies generally consist of a combination of a term life insurance contract and a savings part. However, in contrast to the traditional products, the benefits paid out to the policyholders from the savings component depend on the development of some specific underlying investment fund. Premiums are thus invested individually, and the product is more transparent than participating life insurance contracts. Hence, in general, the savings part of a unit-linked product is identical to an investment in a traditional mutual fund. Unit-linked products are common in Europe and are increasingly relevant also in Asian markets (e.g., India and China, see Svoboda 2022).

Policyholders can often choose from a selection of investment funds and sometimes even individual stocks, with sustainability playing an increasingly important role.[4] Due to the large number of investment possibilities available, individual preferences of policyholders can be taken into account better as compared to traditional participating life insurance contract. The savings premium is calculated by deducting the term premium from the total gross premium (excluding expenses) and is invested in units of the underlying investment at the prevailing price on the payment date, according to the policyholder's choice. As with any investment in a mutual fund, the account value (market value) is given by the number of units purchased multiplied by the price of one unit. Unit-linked products are often extended by offering various types of investment guarantees and may also be based on constant proportion portfolio insurance (CPPI) strategies (see Black and Jones 1987; O'Brien 1988; Black and Perold 1992).

To illustrate the modeling of such policies, we focus on a unit-linked product with an interest rate guarantee and a lookback guarantee (see Gatzert and Schmeiser 2009). Let us assume that the unit price of the conventional underlying fund S_t at time t with a constant drift and standard deviation can be modeled using a geometric Brownian motion under the objective measure $\mathbb{P}$

$$dS_t = S_t\left(\mu dt + \sigma dW_t\right),$$

[3] The modeling and valuation of embedded options from the insurer's perspective under different assumptions regarding premium payment and guarantee forms can be found in, for example, Schmeiser and Wagner (2011).

[4] See, for example, Becker et al. (2022) on ESG transparency, Gutsche and Ziegler (2019) on the willingness to pay of private investors for sustainable investments, and Gatzert and Kraus (2023) regarding customer choices in the context of (sustainable) funds underlying a unit-linked life insurance contract.

with $S_0 = S(0)$, a constant drift μ, a constant volatility σ, and a standard $\mathbb{P}$-Brownian motion (W_t), and T is the maturity of the contract. The solution of this stochastic differential equation is given by (see, e.g., Bjork 2009)

$$S_t = S_{t-1} \cdot e^{(\mu-\sigma^2/2)+\sigma(W_t-W_{t-1})} = S_{t-1} \cdot e^{(\mu-\sigma^2/2)+\sigma Z_t},$$

where Z_t are independent and standard normally distributed random variables. Under the risk-neutral pricing measure $\mathbb{Q}$, the drift changes to the risk-free rate. The development of the total fund value F for regular premium payments P at time t can be derived from

$$F_t = (F_{t-1} + P)\,\frac{S_t}{S_{t-1}} \text{ with } F_0 = 0.$$

Hence, we assume constant premium payments P at time $t_0 = 0, t_1, \ldots, t_{N-1}$ for a contract term of T years, where $t_N = T$.

In case of an *interest rate guarantee*, the premium is compounded with a guaranteed rate of g until maturity, yielding a guaranteed payment G_T at maturity given by

$$G_T = P \cdot \sum_{j=0}^{N-1} e^{g(T-t_j)}.$$

For $g = 0$, at least the sum of premiums paid into the contract is guaranteed. At maturity, the investor thus receives the terminal payoff L_T, which consists of the value of the investment in the underlying fund or at least the guaranteed payment G_T, that is,

$$L_T = \max(F_T, G_T) = F_T + \max(G_T - F_T, 0),$$

which corresponds to the underlying fund value at maturity plus a put option on this value with strike price G_T.

The fund with the *lookback guarantee*, in contrast, pays the highest value of the unit price of the underlying fund S_t that has been achieved during the term of the policy term, multiplied with the number of shares that the policyholder acquired during the contract term. Thus, at maturity T, the terminal payoff depends on the previous $N-1$ unit prices and can be written as (see Gerber and Shiu 2003)[5]:

$$L_T = P \cdot \sum_{j=0}^{N-1} \frac{\max\limits_{j\in\{0,\ldots,N-1\}} S_{t_j}}{S_{t_j}}.$$

Therefore, if the unit price of the underlying fund remains constant throughout the duration of the contract, the policyholder would receive at least the sum of the premiums paid into the contract, which corresponds to the case of $g = 0$ in the case of the interest guarantee.

In the case of a conventional fund, the additional cost of these guarantees must be charged by the insurer on top of the savings premiums and depends heavily on the riskiness of the underlying investment fund. The guarantee can be valued in a risk-neutral manner by calculating the expected

[5] The fund values are locked in at discrete points in time j, the last time in $(N-1)$, that is, one period before the end of the contract's maturity. Since the contract pays off in $N\,(=T)$, the last unit-price of the mutual funds is not part of the lookback guarantee.

value of the final payout less the value of the premiums paid into the contract under the risk-neutral measure and discounting at the risk-free interest rate r, implying[6]

$$\Pi_0 = E^{\mathbb{Q}}\left(e^{-rT} L_T\right).$$

The market price of the contract is typically higher than Π_0 because of transaction costs (see, e.g., Guillén et al. 2014; Gatzert 2019), and it is also impacted by parameter/model risk (Orozco-Garcia and Schmeiser 2015). In addition, how premiums and specifically guarantee costs in unit-linked products are paid—monthly, upfront, as a percentage of the fund value—does not only have an impact on the performance (see, e.g., Gatzert 2013), but also influences consumer choice and the willingness to pay for guarantees (Huber et al. 2015).[7]

Universal Life

Universal life insurance policies are lifelong contracts and essentially consist of a death benefit contract (without a fixed-term, i.e., whole life insurance) with a fixed face value, while the insurance company invests the remainder of the insurance premium (after deducting the risk premium for the death benefit protection) in the capital market to build up reserves. Unlike traditional participating life insurance or whole life insurance, universal life products are highly flexible in that they allow the amount of the death benefit and the amount and/or timing of premium payments to be varied during the term of the contract. Mortality and expense charges are deducted periodically. The remaining premium is then credited to a cash value account (the reserve). It is the policyholder's responsibility to ensure that the account value is sufficient to cover the costs of each period. The insurance company pays a current market rate of interest on the cash value account (Cherin and Hutchins 1987). In addition, upon surrender, the policyholder would receive the cash value account less a surrender charge. The latter is reduced over time to encourage policy retention (Promislow 2015). Although these products offer a high degree of flexibility, they are currently not common in Europe, partly due to tax advantages that depend, among other factors, on the payment of regular-level premiums over a certain time period.

Dynamic Hybrid Products

Dynamic hybrid contracts are particularly present in the German market and aim to combine the ideas of traditional products (the conventional premium reserve) with the upside potential of unit-linked policies (see Kochanski and Karnarski 2011). They thus offer a degree of security but also the potential to generate higher returns than traditional insurance products. Dynamic hybrid contracts are characterized by a regular (often daily or at least monthly) rebalancing process between a risk-free and a risky investment. The strategy is generally procyclical, as funds are shifted to the risk-free asset when stock prices are falling, while the asset allocation is shifted to the risky asset when stock prices are rising. This process is intended to ensure that all embedded investment guarantees can be met. In 2-fund hybrid products, the accumulation benefits are fully or partially guaranteed by allocating part of the portfolio to the conventional premium reserve. The remainder is then invested in risky assets such as an equity fund or a guarantee fund, which ensures that the investment fund does not fall below a critical level (typically 80%). Unlike static hybrid products, in the dynamic 2-fund approach, the

[6] The question if fair pricing is possible in the case of participating life insurance portfolios has been considered in Hieber et al. (2019) and Orozco-Garcia and Schmeiser (2019).

[7] The potential benefit of such investment guarantees from the policyholder's perspective has been analyzed in Døskeland and Nordahl (2008), Russ and Schelling (2018), Luca et al. (2023). In particular, even if contracts are priced "fair", policyholders' willingness to pay may depend on different contract features (see, e.g., Gatzert et al. 2012a; Branger et al. 2010).

portfolio is split between the premium reserve and the risky assets (instead of the regular premium payment being split between the two funds) and depends on the performance of the latter. Dynamic 3-fund hybrid policies are meant to provide the potential for even higher returns through a more efficient construction of the guarantee. In these products, the first fund is a guarantee fund or may even be structured as a high watermark guarantee fund that involves a lookback guarantee, thus locking in the highest fund value over time as a guarantee. As a result, less money needs to be allocated to the insurer's conventional premium reserve which is the second fund and delivers returns independent of the development in the stock markets. Finally, the third fund consists of risky equity investments. Again, the distribution among the three funds depends on the general market environment. Some dynamic hybrid products contain the option to adjust the guarantee during the term of the contract to lock in capital gains. Furthermore, within the risky part of the product, many insurance companies offer their policyholders the possibility to choose among different investment funds or to switch assets without a charge at certain reference dates (Kochanski and Karnarski 2011). In this way, the individual preferences of the policyholder can be taken into account.

2.3 Annuity-Type Products That Aim to Provide Lifelong Income

In addition to the accumulation phase, the decumulation phase is particularly important for ensuring a sufficiently high standard of living during retirement. In this context, innovative proposals in the academic literature often include tontine-like structures,[8] use actuarial and accounting smoothing schemes, defer annuitization to higher ages, or combine it with additional components such as long-term care options (see Bär and Gatzert 2022, for a review of these products).

Immediate and Deferred Annuities, Fixed-Term, and Lifelong Annuities

In the case of immediate annuities, policyholders pay a single premium and start receiving annuity payments immediately. This is often the case with payouts from a life insurance product, for example. Alternatively, policyholders can purchase a deferred annuity, where single or annual premiums are paid during the accumulation (or savings) phase. As explained in the case of life insurance products in the previous subsection, the amount to be annuitized at the end of the accumulation phase depends on the type of contract and thus on the investment returns, interest guarantees and/or surplus participation and smoothing schemes used during the savings phase. The accumulated amount can then be converted into a lifelong annuity where the insurer covers the longevity risk, or a fixed-term annuity. Again, the amount of the annuity payment depends heavily on the type of annuity, the underlying investment fund, and any smoothing mechanism, which can have a significant impact on the retirement benefit. A special case is the Advanced Life Deferred Annuity (ALDA) or "longevity insurance annuity," which further defers annuitization to a higher age, for example, 85 years (see Milevsky 2005; Pitacco 2016), thus offering longevity insurance but at a lower price (see Bär and Gatzert 2022).

Traditional (With-Profit) Annuities

Traditional deferred annuities, also known as with-profit annuities or participating payout annuities, are similar in the accumulation phase to traditional participating life insurance contracts in that

[8] For a presentation of tontines and pooled annuity funds, where funds are redistributed in case of death among the living pool members based on various mechanisms, we exemplarily refer to Milevsky and Salisbury (2015) and Bernhardt and Donnelly (2019), as well as to Bär and Gatzert (2022) for an overview and a discussion of benefits and challenges. Furthermore, a unit-linked tontine has been developed in Chen et al. (2022).

there is no individual investment for each policyholder, but the premiums are invested collectively in the insurer's assets for the whole pool of contracts. In addition, the amount of surplus credited to policyholders' individual accounts depends on the insurer's investment return, the type of guarantee included in the contract, the size of the bonus reserves, and competition. There may also be bonus smoothing between the portfolio of annuities and a portfolio of traditional participating life insurance contracts (e.g., Jørgensen and Linnemann 2011). As in the case of traditional life insurance, the use of actuarial and accounting smoothing allows for a reduction in payout volatility and an increase in stability (see, e.g., Maurer et al. 2013; Bär and Gatzert 2022).

During the payout phase, typically a certain pension amount is guaranteed to the policyholder, which is entitled to surplus depending on the contract characteristics. *Fixed (level) annuities* are guaranteed in their absolute size and not adjusted over time (thus without accounting for inflation), while *increasing annuities* grow by a contractually defined percentage in each period, which can, for example, be inflation-linked. In Europe, annuities are often subject to participation in the insurer's surplus that is generated based on the risk, costs, and investment results. The surplus can thereby be used to increase the guaranteed annuity for the remainder of the contract term by treating the bonus amount as a single premium for a new contract with the same time to maturity as the original one (see Bohnert et al. 2015). The additional annuity amount is thereby calculated based on actuarial calculation principles. Alternatively, a surplus may also be used as a direct payment that increases the annuity in the sense of a one-time surplus payment.

Unit-Linked Annuities

As with unit-linked life insurance, the amount available for annuitization at the end of the accumulation phase of unit-linked annuities depends on the performance of an investment fund, which is individually observable and can typically be influenced by the policyholder by choosing the riskiness of the underlying fund through the portion invested in risky assets. The amount of the annuity is periodically adjusted based on the current value of the fund using actuarial methods based on interest rate and mortality assumptions. Alternatively, depending on the specific contract design, the insurer assumes a specific growth rate for the underlying investment fund, based on which the initial annuity amount is calculated. If the actual fund return is below or above this rate, the annuity is reduced or increased. In many cases, the underlying fund may also be a guarantee fund, which is intended to ensure that the maximum loss is limited to a certain percentage.[9] In addition, unit-linked products can be equipped with additional guarantee as in case of life insurance.

Variable Annuities

Variable annuities can be characterized as unit-linked life insurance contracts with investment guarantees as described previously but structured as annuity products which, in exchange for single or periodic premiums, allow the policyholder to benefit from the upside potential of the underlying investment funds and to be partially protected if the investment falls in value.[10] They are well-established, especially in the United States and Japan, but also in South Korea and increasingly in Taiwan, as in Asian markets, investment-linked products are becoming increasingly relevant despite currently still generally low market shares (Svoboda 2022).

Variable annuities can be designed as deferred or immediate annuities. In general, the customer can choose from a variety of underlying investment funds. Some contracts additionally allow changing the underlying funds at predefined points in time. In many cases, variable annuities include a minimum death benefit (i.e., if the policyholder dies during the accumulation phase, at least the premiums paid

[9] See, for example, www.annuities-online.com/AnnuityOverview.htm

[10] See EIOPA-11/031, available via www.eiopa.europa.eu

into the contract will be paid to the beneficiaries) and provide flexibility in the use of some of the contract's assets before the contract matures. Embedded options in variable annuities can therefore be divided into the two general groups of living benefits and death benefits, with five different types of these two forms being observed in the market (for the following, see Junker and Ramezani 2010; for additional information, see Milevsky and Posner 2001; Ledlie et al. 2008):

- *Guaranteed Minimum Death Benefit (GMDB)*: A predefined minimum amount is paid to the contract's beneficiary if the policyholder dies. Important designs in this context include:
  (a) Return of premiums: In the case of the policyholder's death, the maximum of the premiums paid into the contract and the account value (adjusted for withdrawals) are paid out to the beneficiary.
  (b) Roll-up: The payment to the beneficiary equals the premiums paid into the contract (adjusted for withdrawals) compounded with a guaranteed interest rate (so-called roll-up rate);
  (c) Ratchet: The beneficiary receives the highest value at the contract anniversary dates in case of the policyholder's death, adjusted for withdrawals;
  (d) Maximum out of (b) and (c): The policy's beneficiary receives the greater of the annual ratchet or the roll-up amount;
  (e) Reset: The death benefit will be adjusted in accordance with the account value (adjusted for withdrawals) at contract anniversary dates. In contrast to c), the death benefit may also decrease over time.
- *Guaranteed Minimum Accumulation Benefit (GMAB)*: After a predefined number of years, a certain amount of money is guaranteed to the policyholder, independent of the development of the underlying investment fund.
- *Guaranteed Minimum Withdrawal Benefit (GMWB)*: In this case, the customer is allowed to withdraw a predefined percentage of the account value given at the beginning of the payout phase, independent of the development of the underlying investment fund. The contract owner may also receive provisions—depending on the development of the underlying—that lock in any growth in the contract, thus extending withdrawals and increasing the benefit amount. Sometimes, the percentage of the total investment that one can withdraw is increased if the customer does not make withdrawals during the first years of the contract. For instance, a GMWB with an annual reset leads to a certain bonus (in percent) for each year in which no withdrawal has been made.
- *Guaranteed Minimum Income Benefit (GMIB)*: The policyholder receives a guarantee regarding the annuity payouts provided by the seller of the contract.
- *Guaranteed Lifetime Withdrawal Benefit (GLWB)*: This guarantee provides a smaller percentage of the account value compared to *GMWB*, starting from the beginning of the annuity payout phase until the policyholder's death. Here, the provider typically covers a considerable longevity risk, too. At certain points in time (usually once every five or ten years), the issuer compares the annuity's current account value to the original account value used to determine the minimum guaranteed withdrawal. If the actual account value is larger, the issuer applies the withdrawal percentage to the current account value, thus increasing the minimum guaranteed withdrawal amount.

Additional options, as mentioned for GMDB under (a) to (e), can be used in different forms in the context of the guarantee forms described above. In general, several key factors influence the price and risk of the various guarantees and options typically offered in variable annuity contracts: The volatility of the underlying investment funds (in general, guarantee prices increase with increasing volatility), the market interest rate (in general, option prices increase with decreasing market interest rates); the duration of the contract; the exercise behavior of the policyholder (in particular in regard

to surrender or withdrawals); the evolution of the mortality rates of the policyholders in the portfolio; the general model set up to value the embedded options (e.g., assumptions about distributions); the contract size and the payment method (single or periodic premium payments).

Only some part of the risk associated with variable annuities (e.g., equity risk, interest rate risk, and currency risk) is hedgeable via static or dynamic strategies for the product provider. An overview and a particular strategy for the hedging of interest rate risk is given by Augustyniak and Boudreault (2017).

Equity-Indexed Annuities

(a) *Fixed Indexed Annuities (FIAs)*

(Deferred) FIAs are linked to the performance of a specified index, particularly in the accumulation phase. Equity-indexed annuities were first introduced in the United States in 1995 (see Tiong 2000). The index-linked interest is determined according to a formula that considers changes in a stock index, such as the S&P 500 or a basket of stocks or mutual funds (see, e.g., Hardy 2003; FINRA 2010; NAIC 2011; Tiong 2000). They typically exhibit a minimum guarantee on the premium, for example, 90% of the premium paid, plus a specified annual interest rate (not in case of surrender) in the United States. In Germany, we often find a money-back guarantee, where the policyholder receives at least the sum of premiums paid into the contract. The additional interest credited to the policy value is determined periodically according to a formula that considers the performance of the underlying index as well as further contract features such as the participation rate, interest cap and indexing method. The policyholder's funds generally remain invested in the insurer's collective general assets and are thus subject to smoothing; see, for example, Alexandrova et al. (2017) for a description and analysis of a German version of this product introduced in 2007. At the beginning of each year, policyholders can thus choose between the equity index participation (as a form of surplus participation) or the surplus interest rate that traditional life insurance products receive, which is defined and communicated at the beginning of each year. If index participation is chosen, the insurer uses the budget resulting from the known bonus rate to hedge the index-linked rate.

The participation rate determines how much of the increase in the index is used to increase the policy value. For example, if the index gain is 10% and the participation rate is 80%, then 8% is credited to the policy. The index-linked interest rate is reduced in case a spread, margin or asset fee is imposed by the insurer, which may be implemented instead or in addition to the participation rate (FINRA 2010). There may also be a cap and/or a floor on the annual interest credited to the policy by defining an upper or lower limit. Thus, if the cap is 6% in the above example, the interest rate would be 6% instead of 8%, while a floor of, for example, 0% would ensure that the index-linked interest rate would not become negative. Even if there is no floor, the minimum guaranteed interest must always be met by the insurer. Depending on the contract design, participation rate, cap, and floor may be adjusted by the company after a certain period, and companies may also guarantee a minimum participation rate during the whole contract term or fix a certain range—that is, a maximum and minimum participation rate (NAIC 2011).

Another contract feature is the indexing method. In case of the point-to-point method, the index-linked interest is determined based on the change in the index during a specified period or at the end of the accumulation period (see NAIC 2011). In the case of an annual ratchet (reset), the index-linked interest is calculated annually by comparing the index value at the end of the policy year with the value at the beginning of the year. The annual interest is then used to increase the value of the policy. As the interest rate is reset annually, the interest rate is locked in, and the policy value cannot fall

even if the index falls in the next period (see Tiong 2000).[11] A high watermark (lookback) feature calculates the change in the index at specified points in time during the contract term, for example, at anniversary dates, as compared to inception of the contract. At the end of the accumulation phase, the index-linked interest is determined based on the highest value achieved over the contract term. As the high watermark and the ratchet feature are very valuable, they are often combined with lower participation rates (see Tiong 2000). Hence, there is typically a tradeoff between the different features.

The index may thereby be calculated by averaging the daily or monthly index value instead of using the actual value, which may reduce the amount of interest earned on the index. In addition, in case simple interest is paid, there is no cliquet-style interest rate effect as the interest is not subject to additional interest. Finally, dividends are typically excluded from the index value. In general, the policyholder may be allowed to choose from several broad indices. Further variations of equity-indexed annuities include, among others, flexibility regarding additional premium payments during the contract term, partial withdrawals at the end of the accumulation phase and the decumulation phase, a guaranteed annuity option, or a flexible change of the duration of the accumulation phase during a limited period.

(b) *Registered Index-Linked Annuities (RILAs)*

RILA investment products are increasingly popular in the US market: According to LIMRA (2022), the premium volume was around USD39 billion in the year 2021—compared to around USD87 billion invested in variable annuities in the same year. These products have elements of a fixed indexed annuity (FIA) and a traditional variable annuity (TVA). Like TVAs, RILAs do not offer a full guarantee of the initial premium to provide the opportunity for higher returns. Like FIAs, the performance of the contract is typically based on the performance of a traded equity index over a series of crediting periods, with some form of downside protection for each period (see Moenig 2022). The forms of downside protection over the crediting period are varied: in most cases, a buffer is introduced to cover a tranche of negative returns, or a floor is agreed to cover returns below a predetermined limit (see Moenig 2022). This downside protection is financed by (i) a profit participation for the policyholder and (ii) "a spread, a participation percentage, a trigger rate/step rate or an explicit cap" (Mezger 2023), which limits the upside potential and thus finances the downside protection. RILA products offer a high degree of flexibility as the level of protection/upside limitation can be reset at the beginning of each new crediting period, which is in contrast to general FIA, for example.

As shown in Moenig (2022), RILAs can be offered at little risk to the provider if the guarantees given to policyholders over a crediting period can be statically hedged with options traded in liquid markets. As this is the case for most of the indices and crediting periods offered by RILAs, providers face few of the many problems associated with the dynamic hedging of TVA guarantees that cost many annuity providers dearly in the wake of the great financial crises in 2008/2009 (see Ellul et al. 2022).

For policyholders, RILAs represent an interesting addition to the range of financial guarantees offered in the deferred annuity market, with a risk/return profile somewhere between an FIA (less risky) and a TVA (riskier). As Mezger (2023) shows, classical economic theory can typically not explain the demand for either of these guarantees by investors with any level of risk aversion (other than the tax advantages associated with a deferred annuity). However, behavioral approaches have shown how "loss aversion," "narrow framing," and "subjective beliefs" can influence the demand for these guarantees (see Mezger 2023). Given that RILAs are structured as a sequence of crediting periods, much like FIAs, a combination of loss aversion and narrow framing, referred to by Benartzi

[11] A more detailed analysis of different types of ratchets (simple or compounding the returns monthly or annually) in equity-indexed annuities can be found in Hardy (2004).

and Thaler (1995) as "myopic loss aversion," is a good starting point for looking at their increasing popularity (see Calvet et al. (2023) for an empirical analysis of this effect in case of products with FIA-like guarantees). Mezger (2023) explores the addition of subjective beliefs to this theoretical framework and finds that individuals who neglect extreme but unlikely index returns show a strong preference for RILAs over FIAs and TVAs, suggesting that this dimension also plays an important role in understanding the demand for different financial guarantees (individuals who overweight the importance of such events find FIAs and TVAs more attractive, depending on whether they are optimistic or pessimistic).

Formula-Based Smoothed Investment-Linked Annuities

Another innovative financial product in life insurance is a product class called "formula-based smoothed investment-linked annuity," which was first introduced by a Danish life insurance company in 2002 in a product line called TimePension. It has been described, for example, in Guillén et al. (2006) and Jørgensen and Linnemann (2011), where the latter focus specifically on the payout phase and the possible adjustments of pensions over time. Jørgensen and Linnemann (2011) also compare the new product with traditional with-profit annuities with a bonus payment and a market-based unit-linked product. For all three products, the amount to be annuitized at the end of the accumulation phase depends on investment returns and other contract determinants, including guarantees and bonus features. The amount of the first annuity payment is determined based on the accumulated value. Jørgensen and Linnemann (2011) layout that TimePension is intended to combine "the best of two worlds": The individuality and transparency of a unit-linked scheme (including a higher proportion invested in riskier assets, e.g., 50% in equities) and the idea of smoothing returns, as in traditional with-profit annuities, which makes it possible to stabilize pension benefits over time. This is done by splitting the value of the policyholder's (individual) investment fund into two accounts, namely the policyholder's (pension benefit) account and an individual smoothening ("equalization") account, which belongs to the policyholder and the insurer according to a certain predefined ratio and serves as a buffer for smoothing investment returns. In contrast to traditional with-profit products, the smoothing mechanism is more transparent and mathematically well-defined, still representing an element of collectivity. Firstly, the minimum interest rate credited to the policyholder's account is determined monthly based on, for example, a weighted average of the yield to maturity of leading Danish government bonds. Secondly, the additional interest credited from the smoothing account to the policyholder's account is determined by a certain fraction of the smoothing account, after a review of the balance between the two accounts. This amount is therefore higher when the smoothing account is high and lower in years when it is low or even negative. The latter also means that the interest can be negative (Guillén et al. 2006, p. 233), which, according to Life and Pensions (2009), allows for better exploitation of upswings in the stock market, as traditional products are "always too late." A follow-up version of the product implements a ratchet on the policyholder's account value to ensure non-negative growth. In the decumulation phase, the annuity payment is deducted at each point in time (monthly or annually) based on the policyholder's current account value (book value), which serves as a single premium for a "new" contract with a shorter maturity and higher age of the insured. In addition, annuity payments reduce the fund value. During the payout period, the customer can also be guaranteed an annual minimum income from an annuity certain or life annuity, which is determined at retirement. In addition, a money-back feature is included during the payout phase (see Life and Pensions 2009), which allows beneficiaries to be paid the cash value of the policy if the policyholder dies during the guaranteed payout period. The guarantees in the policies represent a risk to the insurer that can be hedged by a simple vanilla option-based strategy using put and call options (see Life and Pensions 2009). In particular, a classic zero-cost collar can be used as a macro hedging strategy.

Substandard Annuities

While previous pension products have focused on financial innovations and standard annuities, future products may additionally put emphasis on biometric aspects, where the annual annuity depends on the health status of the insured. For example, as outlined in Gatzert et al. (2012b), there are three types of *substandard annuities*.[12] *Enhanced annuities* are offered to individuals with a slightly reduced life expectancy and offer higher pensions depending on environmental factors (e.g., postal code), lifestyle factors (e.g., smoking, marital status, and occupation), and disease factors (diabetics, high blood pressure, overweight). *Impaired (life) annuities* specifically refer to more severe health impairments including heart attack, cancer, and multiple sclerosis, among others, while *care annuities* (immediate needs annuities with long-term care insurance) are focused on seriously impaired persons who already have started to incur long-term-care costs. The risk classification process for substandard annuities requires special underwriting of applicants and can be advantageous to insurers in terms of profitability. Nevertheless, except for the well-established UK market, these products are still not very common.[13] In the context of demographic trends and increasing demand for annuities in the coming years, these products, with their broad characterization and different types of specifications, may represent an opportunity for insurers to gain market share by offering individual annuities depending on the life expectancy (and care needs) of the insured, thereby ensuring that underwriting is carried out in a thorough manner to reduce underwriting risk, that is, reducing the likelihood of assigning policyholders to the wrong risk class.

2.4 Central Differences of Life Insurance Products and Key Features of Innovative Life Insurance Financial Products

The previous two subsections illustrated that "new" life insurance financial products generally aim to combine the best features from traditional and unit-linked products and to reduce their disadvantages with respect to pricing and risk management as well as their risk-return profiles for the policyholders.

An important feature of traditional participating life insurance is the long-term interest rate guarantee that is typically included. While these guarantees are generally favorable from the policyholder's perspective, they are highly valuable and can pose a considerable risk to the insurer in the event of inadequate reserving or risk management. In addition, participating life insurance contracts are typically combined with various forms of complex types of options. Furthermore, premiums are not invested individually but by the insurer for the entire pool of contracts. As a result, it is not clear what part of the assets belongs to the individual policyholder. The participation feature implies a complex surplus distribution mechanism, often dependent on legal requirements, and involves a smoothing scheme where the insurer's returns (risk, cost and investment returns) are buffered and thus less volatile. One further main difference to unit-linked products is that in case of traditional life insurance, typically book values are taken as the basis of computation, which enables the insurer to influence the underlying and its volatility to some extent and which in turn also has an impact on the surplus participation. The use of book values and the application of accounting smoothing can be beneficial for customers in an adverse capital market environment, since the contracts participate in the insurer's hidden reserves that have been accumulated in the past. In an adverse capital market, insurance companies typically have to dissolve hidden reserves to be able

[12] See, for example, Ainslie (2001, p. 16), Brown and Scahill (2010, pp. 5–6), and Cooperstein et al. (2004, pp. 14–15), Gatzert and Klotzki (2016).

[13] See Gatzert and Klotzki (2016), who also study drivers of demand as well as barriers to supply of substandard annuities.

to provide the guaranteed interest rate and an adequate surplus. However, from a customer's point of view, not much transparency is provided in these kinds of products regarding the underlying, the complex surplus mechanisms based on book values, and the prices of embedded options. In addition, participating policies do not offer much flexibility, for example in terms of withdrawing parts of the investment before maturity. However, despite their opacity and complexity, they offer very stable and secure returns due to the use of smoothing techniques and the pooling of contracts, which reduces volatility but limits upside potential. In addition, as risk management is mainly done through pooling and smoothing on a book value basis, model risk is reduced as there is generally no need to implement complex dynamic hedging programs.

In contrast to traditional participating life insurance contracts, unit-linked products can be more transparent, as premiums are invested individually in an underlying mutual fund that is individually observable, and they can also offer the customer a choice between different (sustainable) underlying funds with different risk-return profiles. In addition, a clear modular structure with additional contract components such as additional guarantees and term life insurance is possible. The higher the proportion of risky assets, the higher the upside potential, but this comes at the price of more volatile returns and unstable annuities. One major problem in traditional products is thus the long-term guarantee and the insufficient transparency, which is what some of the innovative products discussed aim to circumvent by integrating an individual underlying fund and by adjusting the guarantees. However, transparency is not increased for all products, as VAs for instance may lack transparency as pointed out by Jing and Bauer (2020). Furthermore, the way of combining traditional and unit-linked aspects differs considerably. Dynamic hybrid products apply the financial concept of iCPPI (individual constant proportion portfolio insurance) by shifting the policy value from the conventional premium reserve (the traditional element) as the riskless asset to a guarantee fund (and possibly a third fund) as the risky asset, depending on the capital market development, to generate the guarantee without additional major hedging measures. In this setting, dynamic hybrid contracts profit from long-term interest earned by the long-term invested premium reserve, even though capital can be withdrawn in the short term. This basically implies a transfer from policyholders having traditional contracts that are pooled toward policyholders with dynamic hybrid products. An advantage is the higher upside potential and the reduced model risk as the guarantee is ensured by means of the investment strategy.

A variant of equity-indexed annuities introduces a traditional element by offering policyholders an annual choice between receiving the safe return, that is, the interest credited to traditional policies consisting of a guaranteed interest rate and an additional surplus generated for the policyholders as a collective or participating in an equity index according to a simple and clearly defined formula. The formula ensures that policyholders have at least a money-back guarantee and also a higher upside potential than in the case of traditional contracts but less volatility and less upside potential than in the case of a classical unit-linked policy.

Danish formula-based smoothed investment-linked annuities, on the other hand, combine traditional and unit-linked aspects in that premiums are invested individually in a mutual fund. The minimum interest rate is linked to an average government bond yield to maturity, which reduces the insurer's risk. The product includes a buffer account, adapted from traditional participating policies to smooth investment returns over time, using a smoothing algorithm to determine the additional annual interest. The smoothing account allows the product to offer more stable returns than unit-linked products, while offering higher expected returns than traditional products. Risk management is partly done through the smoothing account, but also requires hedging, which can be well done due to the well-defined and transparent formula. An important advantage of variable annuities is their modular structure and the flexibility in the type and level of guarantee that consumers can choose for their contracts. However, the guarantees are backed by dynamic hedging programs that are prone to model and basis risk.

Therefore, a key aspect of future innovative products will be the type of guarantee and how it is secured, for example, based on dynamic hedging programs, smoothing algorithms and/or the investment strategy. In addition to downside protection, flexibility is very important, for example, access to a portion of the underlying investment funds prior to the maturity of the contract. In addition, at contract inception, many life insurance products include a guaranteed annuity payment after the accumulation phase. However, the various forms of guarantees described above come at a considerable cost to the client and therefore generally reduce the upside potential of the investment. The choice of product, the embedded options, and the structure of the underlying clearly depend on the customer's individual preferences. Besides, the associated cost of risk management measures will be crucial.

3 Outlook on Future Life Insurance Financial Products

As far as the future of life insurance contracts is concerned, the main issue is likely to be increasing pressure on the cost side of products. Customers and regulators will demand greater price transparency for life insurance products, a process that has already begun in the European Union and the United States, as evidenced by the rise of transparent and rather low-cost RILAs—at the expense of traditional VA policies—in recent years. Such a development could also put price pressure on embedded options and, at least in Europe, lead to substantial advantages for (modular) unit-linked insurance products with flexible and transparently priced investment guarantees. In this context, the investment strategy is particularly important to ensure that the guarantees promised to customers can be met, depending on the performance of the financial markets. Based on transparent guarantee pricing, customers can then decide whether they want to include an investment guarantee in their life insurance product. However, in addition to the cost of risk management—even if it is transparent—the model risk associated with hedging and risk management programs in general will be an important issue in contract design from the provider's viewpoint. In the context of pensions, this has already led to a strong global shift from defined benefit to defined contribution, where guarantees play a subordinate role and various concepts such as pension pools, pooled annuities or tontines are used.

Over the past 10–15 years, new generations of product classes have been introduced in various markets that attempt to address these issues, including RILAs, variants of equity-indexed annuities and formula-based smoothed investment-linked annuities. One key aspect of the innovation is the aim to combine the best features of the traditional and unit-linked worlds by combining the stable and secure returns of traditional life products with the transparent and individualized upside potential of unit-linked products. In addition, costs of guarantees are lower and tend to be less prone to model risk than, for example, in case of variable annuities, which rely on complex dynamic hedging or asset allocation programs. In addition, the annual adjustment of the minimum policy interest rate on a monthly basis, using, a weighted average of the yield to maturity of leading government bonds, significantly reduces the risk to the insurer associated with the guarantee.

Along the entire value chain, digitalization in the areas of distribution, internal processes and claims management will play an increasingly important role—especially when it comes to transparency and customer experience. An example of a transparent product in Europe is the Tontine Trust in Ireland, a start-up offering app-based annuities based on a tontine mechanism. Internationally, several insurtechs and insurers are working with the intention of offering simple and digitized life insurance products, but with a focus on term life products rather than life insurance financial products, for example, in terms of (automated and online/app-based) underwriting using big data analytics and

more.[14] Moreover, it is already common in China to compare and pay for financial services contracts via the "super apps" WeChat or Alipay, and other Asian markets are following suit,[15] with India also showing strong developments in terms of digitization and digital platforms (Swiss Re 2022, p. 16).

In addition, the EU's Sustainable Finance Disclosure Regulation (SFDR), which came into force in 2021, and the Insurance Distribution Directive (IDD), which came into force in August 2022, will introduce greater transparency regarding the sustainability attributes of products. While the SFDR requires financial (life insurance) products to be classified into three categories according to their sustainability objectives,[16] the IDD requires intermediaries and insurance agents selling insurance-based investment products to ask their clients about their sustainability preferences (see Gatzert and Kraus 2023).

Finally, in addition to the type and level of guarantees, transaction costs and risk management issues in general, the regulatory environment and tax treatment, as well as government subsidization of pension products, will play a central role in the development of new financial products in the life insurance sector. Thus, the further development of innovative products will to some extent be country-specific, but the consumers' need for guarantees and stability, individuality, flexibility, and transparency seems to us to be the common driver for future innovation.

References

Ainslie R (2001) Annuities for impaired elderly lives, vol 5. Risk Insights, pp 15–19

Alexandrova M, Bohnert A, Gatzert N, Russ J (2017) Equity-linked life insurance based on traditional products: the case of select products. Eur Actuar J 7(2):379–404

Andrews DL (2018) InsurTech: the next disruptor to the insurance industry, vol 18. Society of Actuaries, Predictive Analytics & Futurism News, pp 22–25

Augustyniak M, Boudreault M (2017) Mitigating interest rate risk in variable annuities: an analysis of hedging effectiveness under model risk. N Am Actuar J 21(4):502–525

Bär M, Gatzert N (2022) Products and strategies for the decumulation of wealth during retirement: insights from the literature. N Am Actuar J. https://doi.org/10.1080/10920277.2022.2078374

Becker MG, Martin F, Walter A (2022) The power of ESG transparency: the effect of the new SFDR sustainability labels on mutual funds and individual investors. Financ Res Lett 47(B). https://doi.org/10.1016/j.frl.2022.102708

Benartzi S, Thaler RH (1995) Myopic loss aversion and the equity premium puzzle. Q J Econ 110(1):73–92

Bernhardt T, Donnelly C (2019) Modern tontine with bequest: innovation in pooled annuity products. Insur Math Econ 86:168–188

Bjork T (2009) Arbitrage theory in continuous time, 3rd edn. Oxford University Press, Oxford

Black F, Jones R (1987) Simplifying portfolio insurance. J Portf Manag 13(3):48–51

[14] See, for example, Andrews (2018), and https://insurtechdigital.com/insurtech/top-five-digital-life-insurers-of-2022, access 2023/08/29. Note that this is also more common for non-life products, for example in case of the pure smartphone-based U.S. insurer Lemonade.

[15] https://www.the-digital-insurer.com/gen-re-blog/tdi-life-and-health-digital-insurance-landscape-in-asia/, access 2023/08/29. See also Huang (2018) on the distribution of insurance products in China using WeChat for marketing.

[16] Article 8 financial (life insurance) products (according to the SFDR) *promote* sustainability aspects but without explicit focus; Article 9 products *explicitly pursue* a sustainability goal by contributing to at least one of the six environmental goals of the EU Taxonomy (Directive (EU) 2020/852); otherwise, Article 6 products have to explicitly state that they do not consider sustainability aspects as required in Article 7 of the EU Taxonomy (see, e.g., Gatzert and Kraus 2023). First empirical and experimental insight suggests that sustainability attributes of products including a sustainable investment strategy play a role for customers with a higher willingness to pay or customer preferences (Becker et al. 2022; Gutsche and Ziegler 2019). In the context of unit-linked life insurance products, Gatzert and Kraus (2023) use a choice-based conjoint analysis for a German sample and find that while private investors do value sustainable product attributes in a unit-linked life insurance contract (Article 8/9 funds according to the EU taxonomy; sustainable investment strategy), the risk-return profile of the fund and especially the ongoing costs are more important. In addition, sustainability attributes play a weaker role in terms of mean willingness to pay in a unit-linked life insurance context than in a pure fund savings plan setting.

Black F, Perold AF (1992) Theory of constant proportion portfolio insurance. J Econ Dyn Control 16(3–4):403–426
Bohnert A, Gatzert N, Jorgensen PL (2015) On the management of life insurance company risk by strategic choice of product mix, investment strategy and surplus appropriation schemes. Insur Math Econ 60(2015):83–97
Bowers N, Gerber H, Hickman J, Jones D, Nesbitt C (1997) Actuarial mathematics. The Society of Actuaries, Illinois
Branger N, Mahayni A, Schneider JC (2010) On the optimal design of insurance contracts with guarantees. Insur Math Econ 46(3):485–492
Brown RL, Scahill PL (2010) Issues in the Issuance of Enhanced Annuities, Social and Economic Dimensions of an Aging Population Program, Research Paper No. 265, McMaster University, Available at https://www.soa.org/globalassets/assets/library/journals/actuarial-practice-forum/2007/october/apf-2007-10-brown-scahill.pdf (download on 10/28/2024)
Calvet LE, Celerier C, Sodini P, Vallee B (2023) Can security design foster household risk-taking? J Financ 78(4):1917–1966
Chang H, Schmeiser H (2022) Life insurance surrender and liquidity risks. Quant Financ 22(4):761–776
Chen A, Nguyen T, Sehner T (2022) Unit-linked tontine: utility-based design, pricing and performance. Risks 10(4):78
Cherin A, Hutchins R (1987) The rate of return on universal life insurance. J Risk Insur 54(4):691–711
Cooperstein SP, Jessen J, Sell SJ (2004) Retirement Income Solutions: Payout Annuities, SOA Spring Meeting, San Antonio, TX, Record, 30(2), available at https://www.soa.org/globalassets/assets/library/proceedings/record-of-the-society-of-actuaries/2000-09/2004/june/rsa04v30n230pd.pdf (download on10/28/2024)
Doll DC (1999) A brief history of universal life. Society of Actuaries
Døskeland TM, Nordahl HA (2008) Optimal pension insurance design. J Bank Financ 32(3):382–392
Ellul A, Jotikasthira C, Kartasheva A, Lundblad CT, Wagner W (2022) Insurers as asset managers and systemic risk. Rev Financ Stud 35(12):5483–5534
FINRA (2010) Equity-indexed annuities—a complex choice, FINRA investor alert. Available at http://www.finra.org/investors/protectyourself/investoralerts/annuitiesandinsurance/p010614 (Download on 27 Sept 2023)
Gatzert N (2013) On the relevance of premium payment schemes for the performance of mutual funds with investment guarantees. J Risk Finance 14(5):436–452
Gatzert N (2019) An analysis of transaction costs in participating life insurance under mean-variance preferences. Insur Math Econ 85:185–197
Gatzert N, Kling A (2007) Analysis of participating life insurance contracts: a unification approach. J Risk Insur 74(3):547–570
Gatzert N, Klotzki U (2016) Enhanced annuities: drivers and barriers of supply and demand. Geneva Pap R I-Iss P 41:53–77
Gatzert N, Kraus (2023) Do sustainability attributes play a role for individuals' decisions regarding unit-linked life insurance? Working paper, Friedrich-Alexander-Universität Erlangen-Nürnberg
Gatzert N, Schmeiser H (2009) Pricing and performance of mutual funds: lookback versus interest rate guarantees. J Risk Res 11(4):31–49
Gatzert N, Holzmüller I, Schmeiser H (2012a) Creating customer value in participating life insurance. J Risk Insur 79(3):645–670
Gatzert N, Hoermann G, Schmeiser H (2012b) Optimal risk classification with an application to substandard annuities. N Am Actuar J 16(4):462–486
Gerber HU, Shiu ESW (2003) Pricing lookback options and dynamic guarantees. N Am Actuar J 7(4):94–95
Grosen A, Jørgensen PL (2000) Fair valuation of life insurance liabilities: the impact of interest rate guarantees, surrender options, and bonus policies. Insur Math Econ 26(1):37–57
Grosen A, Jørgensen PL (2002) Life insurance liabilities at market value: an analysis of insolvency risk, bonus policy, and regulatory intervention rules in a barrier option framework. J Risk Insur 69(1):63–91
Guillén M, Jørgensen PL, Nielsen JP (2006) Return smoothing mechanisms in life and pension insurance: path-dependent contingent claims. Insur Math Econ 38(2):229–252
Guillén M, Jarner SF, Nielsen JP, Pérez-Marín AM (2014) Risk-adjusted impact of administrative costs on the distribution of terminal wealth for long-term investment. Sci World J ID:521074
Gutsche G, Ziegler A (2019) Which private investors are willing to pay for sustainable investments? Empirical evidence from stated choice experiments. J Bank Financ 102:193–214
Hardy M (2003) Investment guarantees: modeling and risk management for equity-linked life insurance. (Wiley Finance Series). John Wiley & Sons
Hardy M (2004) Ratchet equity indexed annuities. Working paper presented at the 14th annual international AFIR colloquium
Hieber P, Natolski J, Werner R (2019) Fair valuation of cliquet-style return guarantees in (homogeneous and) heterogeneous life insurance portfolios. Scand Actuar J 6:478–507
Huang W (2018) The development of wechat marketing and distribution of insurance products in China. Society of Actuaries, December 2018

Huber C, Gatzert N, Schmeiser H (2015) How do price presentation effects influence consumer choice? The case of life insurance product. J Risk Insur 82(2):401–432
Jing X, Bauer D (2020) Accommodation or obfuscation? Financial innovation in the variable annuities market. Working paper
Jørgensen PL, Linnemann P (2011) A comparison of three different pension savings products with special emphasis on the payout phase. Ann Actuar Sci 6(1):137–152
Junker L, Ramezani S (2010) Variable annuities in Europe after the crisis: blockbuster or niche product? McKinsey working papers on. Risk No. 23
Kochanski M, Karnarski B (2011) Solvency capital requirements for hybrid products. Working paper of University of Ulm, Faculty of Mathematics and Economics
Ledlie MC, Corry DP, Finkelstein GS, Ritchie AJ, Su K, Wilson DCE (2008) Variable annuities. Br Actuar J 14(2):327–389
Life & Pension (2009) Innovation of the year 2009, Life & Pensions Awards 2009. Laurie Carver Reports. Available on http://www.sebpension.dk/pow/apps/sebpension/pdf/Tidspension/Innovation_of_the_year_2009_original.pdf (Download on 27 Sept 2023)
LIMRA (2022) Fact tank – sales data. Available on https://www.limra.com/en/newsroom/fact-tank (Download on 9 Sept 2023)
Luca D, Schmeiser H, Schreiber F (2023) Investment guarantees in financial products: an analysis of consumer preferences. Geneva Pap Risk Insur 48(4):906–940
Ma YL, Pope N (2008) Foreign share, insurance density, and penetration: an analysis of the international life insurance market. Risk Manag Insur Rev 11(2):327–347
Maurer R, Rogalla R, Siegelin I (2013) Participating payout life annuities: lessons from Germany. ASTIN Bull 43(2):159–187
Mezger M (2023) The appeal of registered index-linked annuities for myopic prospect theory investors. Working paper, Institute of Insurance Economics, University of St. Gallen
Milevsky MA (2005) Real longevity insurance with a deductible: introduction to advanced-life delayed annuities (ALDA). N Am Actuar J 9(4):109–122
Milevsky M, Posner S (2001) The titanic option: valuation of the guaranteed minimum death benefit in variable annuities and mutual funds. J Risk Insur 68(1):93–128
Milevsky M, Salisbury T (2015) Optimal retirement income tontines. Insur Math Econ 64:91–105
Moenig T (2022) It's RILA time: an introduction to registered index-linked annuities. J Risk Insur 89(2):339–369
NAIC (2011) Buyer's guide to equity-indexed annuities, reprinted by the Illinois Department of Insurance. http://www.insurance.illinois.gov/Life_Annuities/equityindex.asp. Download on 9 June 2012
O'Brien TJ (1988) The mechanics of portfolio insurance. J Portfolio Manag Spring 14(3):40–47
Orozco-Garcia C, Schmeiser H (2015) How sensitive is the pricing of look-back and interest rate guarantees when changing the modelling assumptions? Insur Math Econ 65:77–93
Orozco-Garcia C, Schmeiser H (2019) Participating life insurance portfolios: is fair pricing possible? J Risk Insur 86(2):521–560
Pitacco E (2016) Guarantee structures in life annuities: a comparative analysis. Geneva Pap R I-Iss P 41:78–97
Promislow S (2015) Fundamentals of actuarial mathematics, 3rd edn. Wiley
Russ J, Schelling J (2018) Multi cumulative prospect theory and the demand for cliquet-style guarantees. J Risk Insur 85(4):1103–1125
Schmeiser H, Wagner J (2011) A joint valuation of premium payment and surrender options in participating life insurance contracts. Insur Math Econ 49:580–596
Svoboda A (2022) Why are unit-linked life insurances not popular in Asia? J Insur Financ Manag 7(1):61–84
Swiss Re (2002) World insurance in 2001, sigma 6/2002, Zurich
Swiss Re (2003) World insurance in 2002, sigma 8/2003, Zurich
Swiss Re (2004) World insurance in 2003, sigma 3/2004, Zurich
Swiss Re (2005) World insurance in 2004, sigma 2/2005, Zurich
Swiss Re (2006) World insurance in 2005, sigma 5/2006, Zurich
Swiss Re (2007) World insurance in 2006, sigma 4/2007, Zurich
Swiss Re (2008) World insurance in 2007, sigma 3/2008, Zurich
Swiss Re (2009) World insurance in 2008, sigma 3/2009, Zurich
Swiss Re (2010) World insurance in 2009, sigma 2/2010, Zurich
Swiss Re (2011) World insurance in 2010, sigma 2/2011, Zurich
Swiss Re (2012) World insurance in 2011, sigma 3/2012, Zurich
Swiss Re (2013) World insurance in 2012, sigma 3/2013, Zurich
Swiss Re (2014) World insurance in 2013, sigma 3/2014, Zurich
Swiss Re (2015) World insurance in 2014, sigma 4/2015, Zurich
Swiss Re (2016) World insurance in 2015 sigma 3/2016, Zurich

Swiss Re (2017) World insurance in 2016, sigma 3/2017, Zurich
Swiss Re (2018) World insurance in 2017, sigma 3/2018, Zurich
Swiss Re (2019) World insurance: the great pivot east continues, sigma 3/2019, Zurich
Swiss Re (2020) World insurance: riding out the 2020 pandemic storm, sigma 4/2020, Zurich
Swiss Re (2021) World insurance: the recovery gains pace, sigma 3/2021, Zurich
Swiss Re (2022) World insurance: inflation risks front and center, sigma 4/2022, Zurich
Tiong S (2000) Valuing equity-indexed annuities. N Am Actuar J 4(4):149–170
Trieschmann JS, Hoyt RE, Sommer DW (2005) Risk management and insurance, 12th edn. South-Western College Pub, Mason

Private Equity Investments in the Life Insurance Industry

Anastasia Kartasheva

1 Introduction

The participation of private equity (PE) investors in the life insurance industry has increased in the last 20 years. The ownership of PE has accelerated, particularly after the global financial crisis (GFC) and the consequent period of low interest rates. Why do PE investors enter the life insurance market? Which segments of the life insurance market do they target? What implication does PE ownership have for the life insurance industry's performance, policyholder protection, and financial stability? In this chapter, we discuss the empirical evidence and emerging academic research addressing these questions, building on the experiences of the European and US life insurance markets. To explore the possible mechanisms through which PE ownership can affect the life insurance industry, we also refer to the vast literature on PE financing.

2 Recent Trends in Participation of Private Equity Investors in the Life Insurance Market

Since the GFC, PE firms have been some of the most active buyers in the life insurance industry in the United States and Europe. In the United States, PE firms such as Apollo Global Management, Ares Private Equity Group, Blackstone Group, and Carlyle Group, and diversified financial services firms such as BlackRock and Guggenheim Partners have invested significant amounts in acquiring insurance companies and existing blocks of insurance policies. According to recent reports by the National Association of Insurance Commissioners (NAIC 2020, 2022), the US insurance regulator, the number of PE-owned insurers has steadily increased over the last decade, resulting in 89 insurers at the year-end 2019, 117 at the year-end 2020, and 132 at the year-end 2021. The aggregate cash and

A. Kartasheva (✉)
School of Finance, University of St. Gallen, and Swiss Finance Institute, St Gallen, Switzerland
e-mail: anastasia.kartasheva@unisg.ch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_11

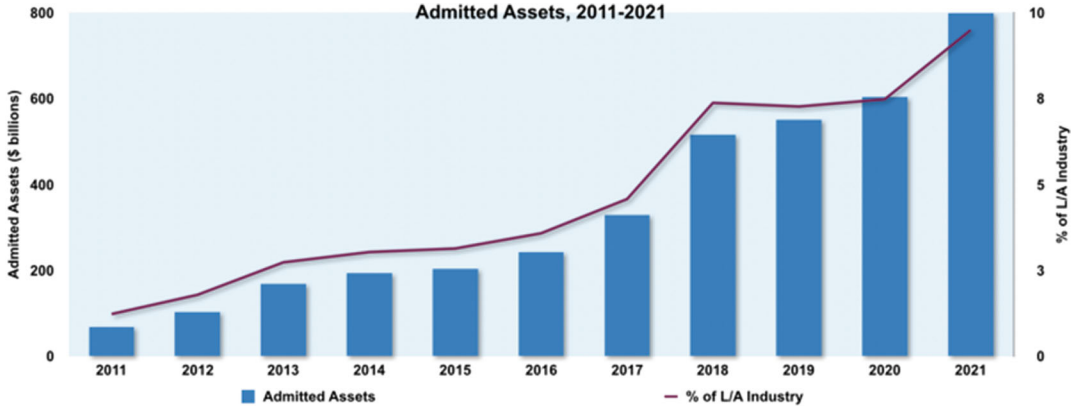


Fig. 1 Private-equity-owned insurer assets in the United States. Source: A.M. Best

invested assets of PE-owned insurers have also steadily increased from $343.5 billion in 2019 to $472 billion in 2021 (NAIC 2022). As illustrated in Fig. 1 and based on the data reported by A.M. Best,[1] the trend of increasing private ownership in insurance has been present for a decade, and in 2021, PE ownership accounted for 10% of the life insurance industry in the United States.

Similar patterns are observed in the European life insurance industry. The European Insurance and Occupational Pension Authority (EIOPA 2022) reports that since 2016, European life insurance merger and acquisition deals have reached €636 billion in terms of transferred insurance liabilities. The UK market was the first European market to experience PE acquisitions, in the early 2000s. It developed as the dominant market for deals between 2016 and 2019, accounting for 77% of all European deals. However, activity in other European countries including Belgium, Germany, Ireland, and the Netherlands has been picking up recently. As a result, the UK deals percentage dropped to 25% by 2020, while transactions in other European markets increased to 30%.

Different types of private market investors are involved in acquisitions in the life and annuity sector. Some acquisitions by PE firms use a typical leverage buyout investment strategy, which we review in Sect. 4. In such cases, the life insurer becomes one of the portfolio companies of a PE fund. However, there are other types of private and even public investors that are outside of the traditional life insurance industry that are actively building up life insurance businesses by acquiring portfolios and entities from traditional insurers. Several privately owned insurers—that is, not listed on the exchange, are also active in the market. Interestingly, the ownership form of the investors can evolve. For example, in the United Kingdom, the pioneering market for PE ownership in insurance in Europe, PE investors have built insurance companies that serve as consolidation platforms to acquire blocks of life insurance policies or life insurance companies. Once the platforms reach scale, they can be sold in the public market by the means of an initial public offering (IPO). We discuss these fluid ownership forms in more detail in Sect. 3.

Two types of life insurance liabilities are targeted by investors. The first type of sales involves closed life insurance books, interchangeably referred to as run-off or legacy books. As the name suggests, these are blocks of life insurance policies that have been sold in the past by the insurer and do not generate new business. The sales of legacy books can be in the form of an entity sale (share

[1] A.M. Best is the rating agency specializing in rating insurance companies. See AM Best

Table 1 Overview of the life insurance divestment market in the United States and Europe

	European Market	US Market
Portfolio Transfers/Asset Deals	35 (29%)	9 (12.5%)
Entity Transfers/Share Deals	87 (71%)	63 (87.5%)
Total	122	72

Source: Capital IQ, own research

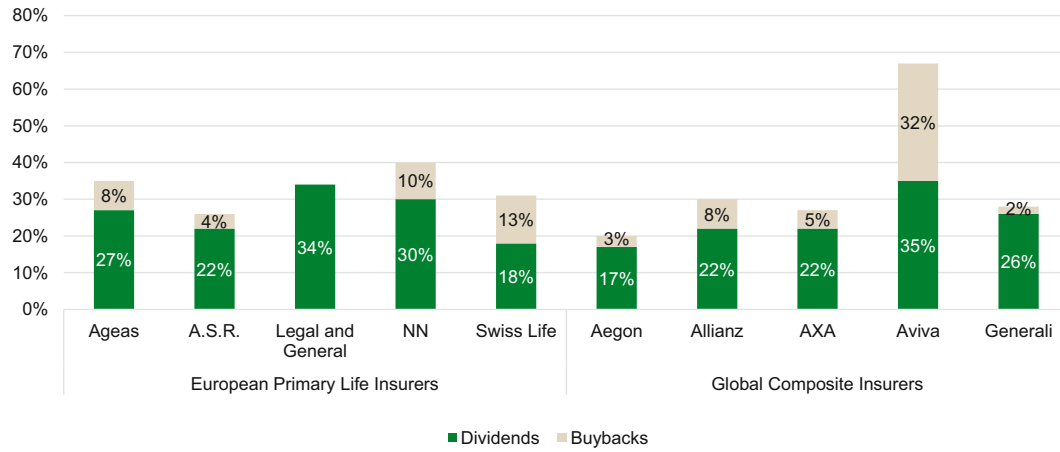


Fig. 2 Cumulative capital return (dividends, special dividends, and share repurchases) in 2018–2022 versus end-January 2023 market capitalization. Source: Companies' financial statements, Bloomberg

deal) or a portfolio transfer (asset deal) where the ownership of the portfolio is transferred from the seller to the buyer. The other type of target, typically in the form of a share deal, is the acquisition of an insurance company that still underwrites new policies and can be utilized as a platform to generate new business. Both types of acquisitions are present in Europe and the United States.

Using various sources, including Capital IQ, media announcements, and company reports and press releases, we have assembled a dataset of European and US life insurance acquired by investors outside the insurance industry. The overview of this market is reported in Table 1. We can observe that in both markets asset deals and share deals are present.

The expansion of PE investors in the life insurance market contrasts with the strategy of the major listed US and European insurers. Besides exiting the guaranteed life insurance business and selling legacy portfolios with financial guarantees to investors outside of the insurance industry, these insurers are divesting from the life insurance industry by paying out significant dividends and buying back the shares. Figure 2 depicts cash payments to shareholders during 2018–2022 for a selected group of major European and global insurers, comprising dividends, special dividends, and share repurchases as per company financial statements. It illustrates that for several major global insurance groups, these payments over the four-year period accounted for 25–30% of their market capitalization in January 2023.

The evolution of the life insurance industry starting from the 1980s needs to be considered against the background of the secular decline in long-term yields, which has accelerated since the GFC and the Euro debt crisis due to EU and US central bank interventions. Persistently low interest rates depress insurers' net investment spread on new business and force them to reinvest the proceeds from

maturing bonds into bonds paying lower coupon rates. Because changes in long-term bond coupon rates tend to co-move with changes in long-term interest rates, lower coupon rates further depress life insurers' overall net investment spread and, in turn, their financial condition. As a result, the net investment spread on new business decreases because the difference between the insurer payoff on its fixed-rate liabilities and the returns it earns on its bond investments decreases. In addition, the net investment spread on the insurers' legacy business also decreases because maturing bonds of relatively shorter duration are reinvested in lower-paying bonds. Consistent with rational interest rate risk management, the life insurance industry's response to the secular decline in long-term yields has been to exit certain business lines, especially savings products with guaranteed minimum returns.

Though the adjustment of the business line focus applies to the life insurance industry in the aggregate, it is notable that not all life insurers do so. Several large incumbents in the US market have grown significantly since the GFC. These insurers have refocused their liabilities to lower cost-deferred annuities and increased their investment in private markets, as articulated in detail in (Foley-Fischer et al. 2023). Life insurance incumbents with private credit origination platforms and deep expertise in private assets have been large players in the private asset credit space for a long time because holding such assets fits asset liability management. However, following the GFC, many more life insurers were pressured to sell their legacy portfolios, which is the likely reason we observe the significant entry of PE firms.

Why has investing in the life insurance business become so attractive to PE and alternative asset management investors, while traditional life insurers are reducing their supply? In the following section, we discuss several supply and demand drivers.

3 Supply and Demand Drivers in the Divestment Market

In this section, we review the factors that contributed to the emergence and growth of the divestment market. On the supply side of the market are the insurance companies, which faced low interest rates and modernized regulatory requirements after the GFC. These structural shifts significantly reduced the profitability of the traditional life insurance business with financial guarantees. Putting legacy portfolios in the run-off and divesting the portfolios or entities with guaranteed life insurance liabilities became an attractive strategy for some insurers. On the demand side of the market are the private market investors, who are attracted by the sticky nature of life insurance liabilities as a source of funding, possibilities to access large amounts of assets to manage and earn asset management fees, and opportunities to improve the operational performance of target life insurers.

3.1 Supply Side: Structural Changes in the Life Insurance Industry

Life insurance liabilities with financial guarantees is the largest class of life insurance products in Europe and in the United States. In Europe, these liabilities accounted for about €4 trillion in 2022, and they are particularly pronounced in Belgium, France, Italy, and the Netherlands.[2] Similarly, in the United States, life insurance policy reserves in the annuities segment of the life insurance market were USD4.3 trillion in 2021, corresponding to 67% of the life insurance market (ACLI 2022).

Life insurers are exposed to interest rate risk. Life insurance products combine biometric risk and savings elements. Policyholders pay premiums upfront for payoffs that will be realized in the future. The downside risk for the life insurer is that interest rates decrease and thus reduce the returns

[2] Data source: ECB. https://sdw.ecb.europa.eu/reports.do?node=1000005760

the insurer earns on its fixed-income investment portfolio. The extent of the exposure to the risk depends largely on the product features, including the schedule of contributing premiums to the policy, the flexibility to withdraw accumulated funds and the fees and tax incentives associated with such withdrawals, and the insurer's guarantees to the policyholder on the performance of the savings portion of the insurance policy.

To some extent, life insurers can hedge their exposure to interest rate risk (Brunetti et al. 2023; Hartley et al. 2017). Asset liability management is an important strategy used by insurers to ensure that funds are available to meet policyholders' demands. It creates natural demand for high credit quality fixed-income assets, such as corporate and government bonds (Koijen Ralph and Yogo 2023). However, creating a portfolio to completely eliminate the exposure to interest rates may be suboptimal or simply infeasible, mainly because the liabilities are long dated and derivatives and long-term assets can be in limited supply. As an imperfect substitute, insurers use approximate hedges, such as duration matching, which help insurers to eliminate small changes in interest rates but leave exposure to large interest rate changes unhedged (Hartley et al. 2017). Furthermore, Domanski et al. (2017) argue that duration matching can be ineffective during sharp changes in interest rates because the synchronized search for duration of life insurers and pension funds acts as an amplification mechanism and results in further downward pressure on interest rates. Sen (2023) shows that the regulatory treatment of interest rate risk of financial guarantees can curtail the incentives of life insurers to reduce their interest rate risk exposure. Verani and Yu (2021) study how the costs of interest rate risks exacerbate the adverse selection problem by raising annuity prices, which in turn changes the duration of annuity liabilities. They estimate that the cost of interest rate management conditional on the effect of adverse selection accounts for about half of annuity markups.

Following the GFC, interest rates decreased significantly and remained low for a long period of time (Fig. 3). Hartley et al. (2017) show that exposure to the decline in interest rates was particularly harmful for life insurers who offered products with financial guarantees and policyholder options such as the variable annuities offered in the United States. In contrast, UK insurers, who significantly reduced their exposure to financial guarantees in the early 2000s, in part in response to the adoption of the risk-based Solvency I regulatory regime, were not significantly exposed to low interest rates in the decade following the GFC.

Life insurers hit by the prolonged period of low interest rates have been adapting to the new environment. Changing the product mix by reducing the share of products with financial guarantees or reducing the generosity of guarantees and their costs for policyholders is one strategy. Koijen Ralph and Yogo (2022a) report that the share of contracts with minimum return guarantees in the US variable annuity market decreased from 36% in 2007 to 20% in 2011; the average annual fee on contracts with minimum return guarantees increased from 2.04% to 2.38% between 2007 and 2009. Koijen Ralph and Yogo (2022a) estimate the model to assess the drivers of the fee increase and show that it is driven to almost equal extents by the option value of the guarantees and the shadow cost of capital, highlighting that financial frictions are as important as option pricing considerations for explaining the fees.

Due to the long-term nature of insurance liabilities, the adjustment of the business mix takes time. New business accounts for only 10–15% of insurers' annual revenues, which can explain the persistence of the underperformance of life insurers with variable annuity portfolios during the GFC and the COVID-19 pandemic (Koijen Ralph and Yogo 2022b).

Another response by insurers willing to restore their investment return is reaching for yields that intensified during the period of low interest rates (Becker and Ivashina 2015). Ellul et al. (2022) show how the interaction of financial guarantees and the regulatory framework that incentivizes insurers to hedge the guarantees shifts insurers' investments into high-risk and illiquid bonds. Since

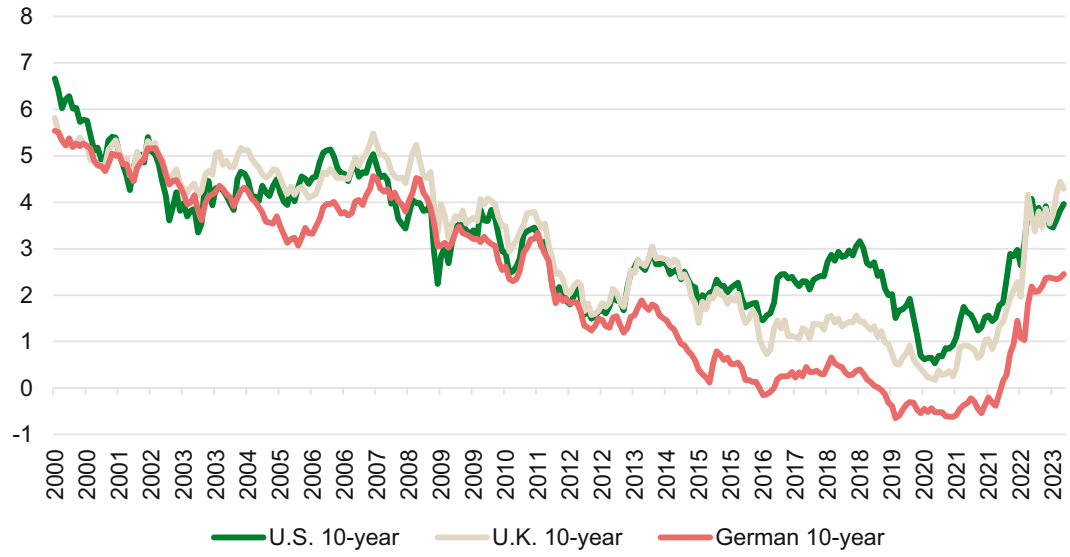


Fig. 3 Ten-year constant maturity government bond yields. Source: Capital IQ

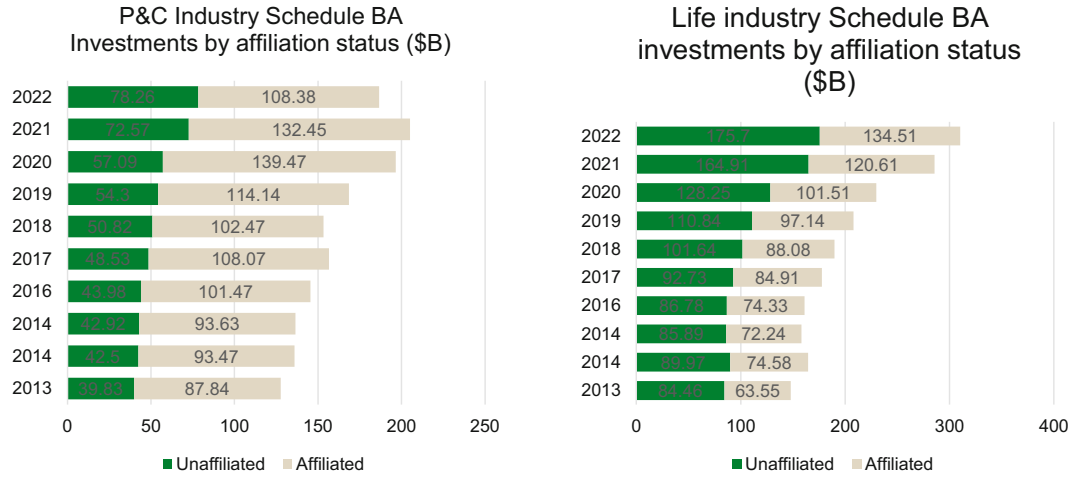


Fig. 4 Aggregate investment in alternatives by US insurers, 2013–2022. Source: Capital IQ

the incentives are synchronized among insurers exposed to financial guarantees, the share exposures can exacerbate fire sales during market downturns.

Interestingly, like other institutional investors hit by the prolonged period of low interest rates, both life and non-life/property-casualty insurance companies have been increasing their participation in the alternative asset classes to reach for yield (Fig. 4), which also includes insurer investment through PE.

Though adjusting a product mix and increasing risk taking in the asset portfolio can temporarily improve performance, the longer-term response to low interest rates by life insurers in Europe and the United States has been to divest from the life insurance business with financial guarantees and to expand the protection and fee-based business. Regulatory modernization in the United States and Europe in terms of the valuation of financial guarantees and the broader risk-based framework of Solvency II implemented in the European Union in 2016 have accelerated the transition.

Of course, such a transition is only possible when the other side of the market is willing to acquire life insurance businesses with financial guarantees and is able to create value in ways that are not feasible for traditional life insurers.

3.2 Demand Side: Types of Private Equity Investors and Their Potential Motives

The buyers of legacy books can be categorized into three types, depending on their ownership structure: (1) PE-fund-owned players, where the PE fund is a closed-end fund with a finite investment horizon further described in Sect. 4 (e.g., Eurovita, Viridium, and GamaLife); (2) privately owned limited insurance companies with permanent capital investments (e.g., Athora, Monument, and Resolution); and (3) publicly owned insurance companies that specialize in closed life insurance books consolidation (e.g., Chesnara and Phoenix). Figure 5 provides more details about the market activity of these different types of investors in Europe.

The proposed categorization is fluid and may evolve depending on the business focus and success of the buyers. In terms of the volume of transactions, the companies with permanent capital (2) and listed companies (3) represent the largest share of the market participants in Europe. The publicly owned companies (3) operate primarily in the UK mature market; these consolidators of life insurance legacy books were originally built with private capital, but once they reached a meaningful size and profitability they were listed in the public market. PE-fund-owned players (1) may enter the market by starting to acquire legacy books with the aspiration to develop their business model and expand to become type (2) and, eventually, become type (3). However, type (1) buyers may also include opportunistic buyers who are willing to commit capital according to the fund-limited investment horizon before then selling the legacy books portfolio to the other two types of buyers, (2) and (3).

Within each of these types, which are distinct in the ownership structure and the investment horizon of the shareholders/owners, there are further important characteristics that determine the buyers' business models and value drivers. Some investors operate with significant leverage akin to leveraged buyout (LBO) structures, while others have low leverage. Some acquirers have financial strength ratings from A.M. Best and other major credit ratings commonly used by traditional insurers, while other acquirers are unrated. In addition, buyers can either operate across multiple geographic markets or focus on one market. Some buyers primarily target smaller transactions—also driven by their limited access to capital—while others increasingly look for larger transactions.

The source of value creation can also differ across buyers. Mature life insurance closed book consolidators generate cost efficiencies and scale benefits driven by their experience and operational and IT infrastructure created by the consolidation platforms. Some buyers are more focused on generating incremental value from the balance sheet by improving risk and capital management and revising their asset allocation. Among these buyers, private companies with permanent capital aim to create long-term businesses with differentiated investment strategies. Finally, in each of the categories, there are buyers who rely on significant leverage and financial engineering to boost their returns.

Some examples illustrate the diversity of buyers' business models. In the type (2) category, international portfolio consolidators, such as Resolution, are built to be large and are focused on acquiring entities or portfolios; they have a worldwide scope and seek economies of scale and diversification. Larger players do not consider portfolios of less than €1 billion. In contrast, local portfolio accumulators, such as Monument Re, target smaller portfolios in selected geographical markets. Because of their limited capital, these buyers are unable to acquire large portfolios. Instead, they buy portfolios at a deep discount and seek opportunities in a pool of portfolios in the €50–300

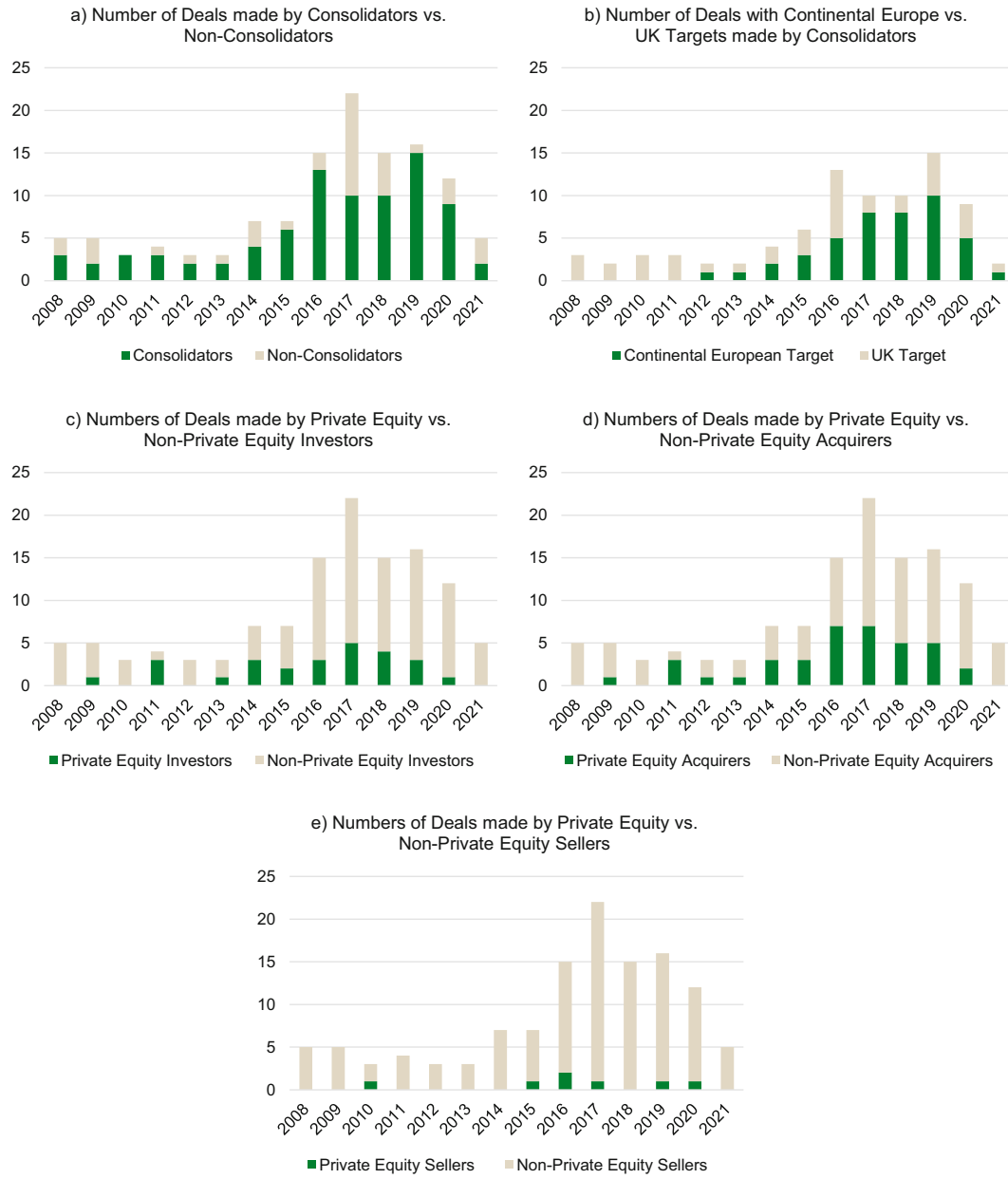


Fig. 5 European investment activity in the legacy book divestment market by investor type for the period 2008–2021 Q1. Source: Capital IQ, media announcements, company reports, own research

million range. Many direct insurers are willing to sell these portfolios to remove their administrative burden.

In the type (1) category, there are large PE-owned insurers focused on one geographical market, for example, Viridium in Germany and Eurovita in Italy, which are controlled by the United Kingdom-based private equity firm Cinven. We discuss the Eurovita case in more detail in Sect. 5.2. Type (1) entities use the common PE fund structure and sometimes aim at fast growth during a limited

investment horizon. In addition, there are PE-backed insurers focused on the acquisition of portfolios across Europe, such as GamaLife in Italy and Portugal and MEDVIDA in Spain.

4 The Features and Impact of Private Equity Ownership

Private equity has several unique institutional features that make it distinct from the traditional corporate structure of insurance companies. Some of these features are relevant to the regulatory issues of policyholder protection and financial stability that appeared recently in policy discussions. For these reasons, we start by describing the PE fund structure and the deal structure.[3] Then, we summarize the extensive literature that analyzes the impact of PE ownership on portfolio companies and aims to identify through what mechanism PE affects their operating performance. Finally, we review the relatively small literature focusing on PE acquisitions of life insurers.

4.1 What Is Private Equity Ownership?

Private equity involves a range of investment strategies including leverage buyouts (LBO), venture capital and growth equity, special situations such as distressed debt, and mezzanine. However, it is the LBO structures that are most relevant to PE involvement in the life insurance industry. Besides the primary market, there is a growing secondary market, where the original PE owners exit by selling the company to another PE fund. As we have mentioned in the previous section, secondary market purchases are not uncommon in the life insurance industry and are used as an exit strategy by some opportunistic investors.

At the core of the structure is the PE fund, which is formed as a limited partnership (Fig. 6). Investors in the fund are typically institutional investors, such as pension funds, asset managers,

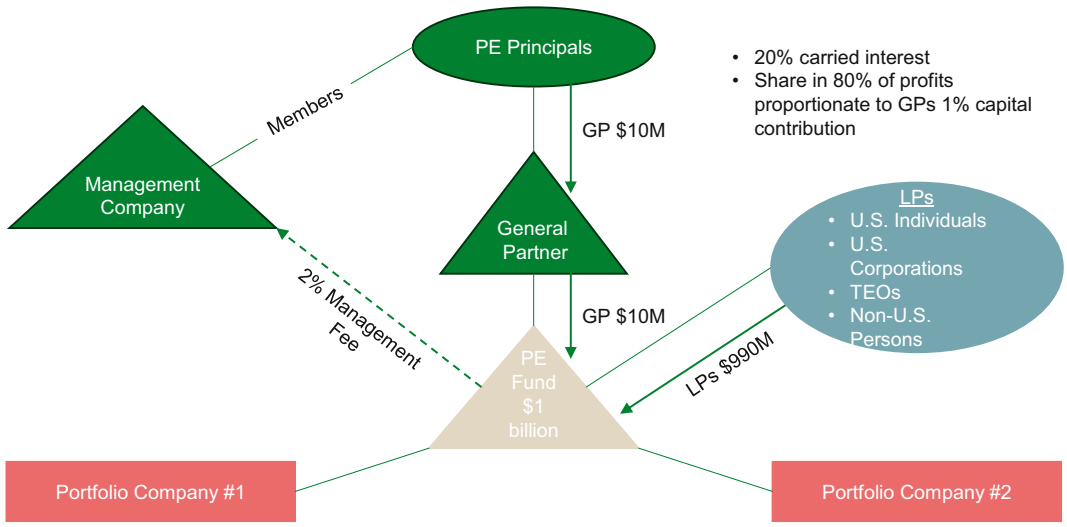


Fig. 6 Structuring a private equity fund. Source: Alternative Investment Program at Wharton

[3] Comprehensive reviews of the operation of the private capital market are provided by Jenkinson et al. (2021) and Ivashina and Lerner (2019).

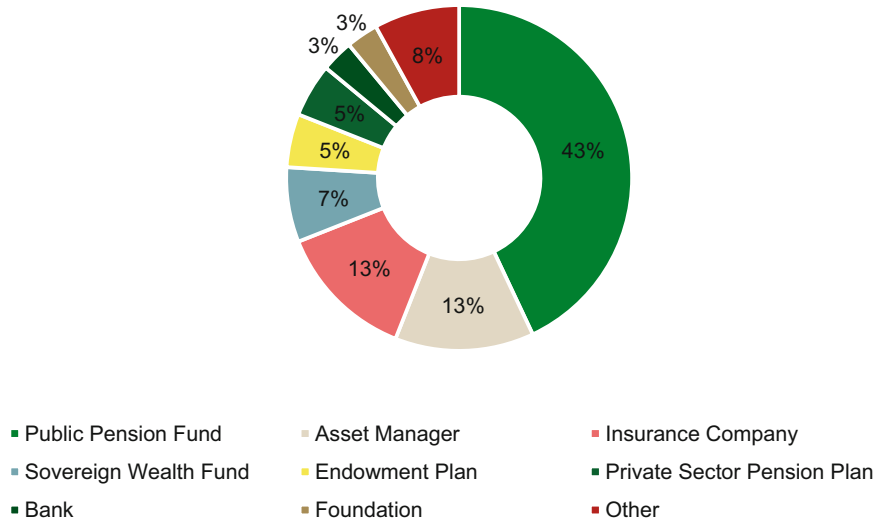


Fig. 7 Limited partners institutional affiliations. Source: Preqin

sovereign wealth funds, and so on (Fig. 7); they are referred to as limited partners (LPs). Besides being protected by limited liability capped by their investment amount, LPs are also passive investors who are not involved in investment decision-making. LPs' limited liability and passive role in investment management are the key features that suit deep-pocket institutional investors such as pension funds by allowing them to invest in private markets without the exposure to downside risk beyond their capital commitment to the fund. At the same time, the PE fund structure implies that the LPs' investment is illiquid and requires LPs to overcommit capital.[4] Despite the illiquidity and rigid fund structure, institutional investors have been attracted by PE investment due to the higher returns that PE funds can deliver, particularly during the period of secular decline in interest rates (Ivashina 2022).[5]

Management and investment decisions of the fund require a general partner (GP). Though in principle the GP has unlimited liability for any claims against the PE fund, typically, GPs are separate legal entities rather than individuals. Another important entity is the management company, which employs investment management professionals and is involved in sourcing and selecting deals and performing due diligence, in return for the 2% management fee. The PE principals, often (and confusingly) also referred to as general partners (GPs), are members of the management company. The PE principals are expected to provide about 1% of their own capital to the fund, while the LPs provide the other 99%, as illustrated in the numerical example in Fig. 6. The PE principals are compensated with 20% carried interest and the share of profits proportionate to the GP's 1% capital contribution.

[4] LPs commit capital to the fund, but the fund is called only when it is required by general partners (GPs) to acquire a new investment or to pay fund expenses. As a result, LPs typically need to over-commit capital, since less than 100% of the capital will be invested at each point in time. The LPs have no control on when the capital is called and when the investments are exited.

[5] The assessment of PE funds' performance is an active research area. Overall, the consensus of the literature is that PE on average outperforms S&P 500 net of fees, and the returns tend to be persistent (Harris et al. 2014; Sensoy et al. 2014). However, Phalippou (2014) shows that the results of the studies strongly depend on the choice of the benchmark index. Since Kaplan and Schoar (2005), the literature relies on public market equivalent to compare the fund return to the relevant public equity market benchmark. Using such a comparison requires risk adjustment, which is addressed in several recent studies (Driessen et al. 2012; Jegadeesh et al. 2015; Korteweg and Nagel 2016; Sorensen et al. 2014).

The ultimate management company (not shown in Fig. 6) was established by the initial founders and manages several funds with the structure depicted in Fig. 6. The ultimate management company does not need to be a partnership, and there are ultimate management companies listed on exchanges, such as Apollo, Blackstone, and KKR. The ultimate management company controls the brand and can launch financing of new funds.

The companies acquired by the fund are referred to as portfolio companies (PCs), and life insurers acquired by PE investors are an example of a portfolio company. In the case of life insurance, as mentioned above, the acquisition can take the form of a block of insurance policies, closed insurance books (CIBs), or the whole legal entity. Sometimes, a portfolio company is a legal entity that consolidates blocks of insurance policies of the acquired companies on a single platform.

GPs can create value for life insurance PCs or blocks of closed-book insurance policies either by boosting the returns of the asset portfolio or by reducing operational costs. The superior asset returns can be obtained due to GPs' access to other funds within the PE firm. By investing in these funds, GPs can earn the illiquidity premium but also increase the liquidity risk of the asset portfolio underlying the insurance policies. In addition, GPs can add expertise in managing the credit risk of the portfolio compared to traditional life insurance investment strategies in high quality fixed-income instruments such as government and investment grade corporate bonds.

By consolidating blocks of insurance policies on the same platform, GPs can reduce the operational costs due to higher incentives to invest in the IT infrastructure to monitor the development of liabilities as well as the ability to improve the predictability of payouts due to diversification among different blocks. In contrast, the management of a closed insurance block by a single insurance company becomes progressively more costly in terms of liquidity and required capital per unit of exposure, as the block size diminishes in the run-off.

LBO deals are highly levered, which boosts the returns. The amount of debt was peaking at 60% prior to the GFC. Since in the LBO the majority stake of the PC is acquired by the fund, the acquired assets can typically serve as collateral for the senior debt. The predictable cash flow to service the debt is thus an important criterion for a company to be a good LBO candidate. The predictability and regularity of the life insurance portfolio payouts is thus one of the reasons why life insurance liabilities fit well for LBO acquisitions.

PE funds have a fixed-term lifespan, usually 10 years. The relatively short-term investment horizon of GPs compared to the long-term nature of the life insurance business model can potentially lead to opportunistic behavior of the GP. However, GPs can also use the time to build a consolidation platform and exit by selling it to the public market. Another relevant form of PE ownership in insurance is a private insurance company—that is, one not listed on the exchange. These types of companies have a longer investment horizon than the 10-year horizon of a PE fund; often, private insurance companies built for the consolidation of life insurance closed books are affiliated with PE firms and thus have similar access to PE funds.

4.2 Value Creation or Wealth Transfer?

The analysis of the question of whether PE investors create or transfer/extract value from the portfolio companies is at the core of PE financing research. Our summary of the literature is based on a recent comprehensive review of more than 300 articles on the PE and venture capital financing literature by Tykvová (2018), and several more recent papers. Though the scope of Tykvová's (2018) review covers a wider range of topics, including the relationship between GPs and LPs and PE performance measurement, it is remarkable that about 50% of the reviewed papers deal with the various aspects of value creation or value transfer issues. These papers analyze the impact of PE ownership on

PCs in terms of operating performance, financial support, and exit strategies. In this subsection, we summarize the mechanisms that have been analyzed in the PE literature and discuss their relevance to the life insurance industry.

The general conclusion of the literature is that PE acquirers succeed in improving companies' operating performance. Acharya et al. (2013) show positive abnormal performance of PE-initiated deals, controlling for leverage and sector returns. Boucly et al. (2011) find that PE acquirers foster not only the profits of PCs but also their growth, by relaxing targets' credit constraints. Other studies (Aslan and Kumar 2011; Chen et al. 2014; Cohn et al. 2014; Davis Steven et al. 2014) also demonstrate a positive impact of PE acquisition on PCs.

One mechanism through which PE funds can affect the performance of PCs is by taking an active role in monitoring and advising PCs, which leads to more effective corporate governance. Cornelli et al. (2013) show that governance reforms increase the effectiveness of board monitoring by avoiding firing a CEO for bad luck or in response to external shocks. Chen et al. (2014) find that compared to non-PE acquirers, PE acquirers are more likely to place directors with finance- or industry-relevant experience on companies' boards. As a result, targets of PE investors experience higher short-term and long-term performance. Furthermore, Bloom et al. (2015) and Edgerton (2012) show that PE investors tend to improve management practices.

However, improvement in the operating performance of PCs can derive from various sources, such as increase in revenues, increase in prices, reduction of the input costs, tax optimization, or increase in risk taking and leverage. Furthermore, better operating performance of PCs can come at a cost to other stakeholders, such as employees and customers. Therefore, distinguishing between value creation and value transfer requires determining the source of higher operating performance.

Evidence on the sources of improvement is mixed. Analyzing a large sample of target firms in various industries, Davis Steven et al. (2014) find that PE buyouts lead to modest job losses but large increases in gross job creation and destruction. Also, the source of total factor productivity gains occurs mainly through the accelerated exit of less productive firms and the greater entry of highly productive ones. Analysis of the impact on target firms' consumer products by Fracassi et al. (2020) shows that while manufacturer increases revenues, growth is driven by increasing the volume rather than raising the price. Firms increase the variety of products or innovate into new products and expand geographically. By contrast, Eaton et al. (2020) provide evidence from the higher education sector that PE-owned schools excel at obtaining government aid. At the same time, these schools increase tuition and per-student debt, provide less education inputs, and have lower graduation and repayment rates. Furthermore, their graduates have lower earnings. In the hotel sector, Spaenjers and Steiner (2021) document that operating efficiency improvements come at the expense of lower guest satisfaction. For nursing homes, Appelbaum and Batt (2020), Gupta et al. (2021), and Liu (2021) also argue in favor of value transfer by documenting the negative effects of PE buyouts such as an increase in patients' short-term mortality, increased market power, and an increase in the total spending of the privately insured. However, Gandhi et al. (2020a, b) explain that nursing homes' staffing changes are heterogeneous depending on market concentration. Gao et al. (2021) show that nursing homes' skilled-worker ratios improve, while patient satisfaction scores do not decline.

LBOs rely heavily on debt, which boosts the observed performance but can also increase the risk of financial distress and bankruptcy. However, several studies support the opposite view that the PE approach to debt financing eases and improves the terms of access to capital. Tykvová and Borell (2012) show that the distress risk indeed increases after the buyout. However, higher risk does not translate into higher bankruptcy rates than comparable non-buyout companies. In fact, PC bankruptcy rates are even lower when companies are backed by experienced PE funds, suggesting their higher ability to manage distress risk. Also, Demiroglu and Christopher (2010) and Ivashina and Sun (2011)

suggest that PE investors have better access to loans and obtain loans on more favorable terms because of the bank relationships and reputation that PE investors form through repeated interactions with banks. Bernstein et al. (2019) found that during the 2008 GFC, PE-backed companies decreased investments less than their peers and were able to raise more debt and have higher asset growth. They argue that the ability of PE investors to actively manage PCs during the crisis helped to alleviate the financial shocks.

Tax benefits are another aspect of debt financing typical to LBOs, and the improvements in performance may simply be obtained from tax savings of debt. As suggested by Shleifer and Summers (1988), tax savings of PE deals are the redistribution from the government to PE investors, and thus they transfer rents rather than create social value. In favor of this view, Guo et al. (2011) find that higher debt levels boost tax shields, which raise PE returns and represent a transfer from taxpayers.

Short-termism is another potential issue of PE ownership. Short-termism can be driven by the structure of the PE funds, which are expected to realize returns and distribute them to LPs within a finite investment horizon. Therefore, PE funds may have incentives to forgo long-term growth to boost short-term performance. Lerner et al. (2011) analyze one aspect of long-run performance—investment in innovation as measured by patenting activity. Their analysis does not support the idea that PE investors sacrifice long-term investment. On the contrary, PE-backed companies' patents are more cited and have become more concentrated in important areas of companies' innovative portfolios.

One empirical challenge in assessing the impact of PE investors on the performance of targets is that the matching between the PE investors and their PCs is not random. Hence, the observed positive outcomes summarized in this section may be the result of the ability of PE principals to select targets with high potential (selection effect) rather than to add value after the acquisition (treatment effect). The PE literature suggests that both effects are present. Acharya et al. (2013) and Chen et al. (2014) find that abnormal performance depends on the matching of the PE partner background with the deal type (treatment effect). Bernstein and Sheen (2016) find that PE ownership improves restaurants' operations rather than select better-run establishments (treatment effect). Tykvová and Borell (2012) show that PE investors select companies that are less financially distressed (selection effect).

4.3 What Is the Impact of PE Ownership on Life Insurance?

Which of the mechanisms discussed in Sect. 4.2 are relevant for PE acquisitions in the life insurance industry? Though PE ownership in life insurance has been steadily increasing since the GFC in the United States and Europe, and even prior to the GFC in the United Kingdom, very few academic studies provide insights into the role of PE ownership in insurance. In this section, we summarize the recent research and discuss whether these studies affirm value creation or value transfer.

In the context of the US market, Kirti and Sarin (2023) analyze the acquisition of subsidiaries of life insurance companies (but not life insurance closed books) during the period 2005–2014, which amounts to 57 acquisitions in their sample. Furthermore, the sample consists of acquisitions by a broader range of private investors than conventional PE funds and also includes hedge funds, business development companies, and other private investors. The target insurers in the sample tend to be smaller and less capitalized than non-target insurers. In the share deal segment of the market targeted by these private investors, the targets are concentrated in the fixed-annuity market (32% of the life insurance industry in 2009) rather than life insurance (24%) or variable annuities (43%). Over the same period of the study, the market share of PE-backed life insurance companies in the fixed-annuity market grew from 3% to 20%.

Kirti and Sarin (2023) document that upon the acquisition of life insurance subsidiaries, PE owners implement quick and significant changes to their asset portfolios. PE-backed insurers substantially

reduce their holdings of highly rated corporate bonds and increase holdings of poorly rated private label asset-backed securities (ABSs), increasing their holdings of ABSs by two-thirds of the industry average.

The aggressive investment in ABSs by PE-backed insurers is possible because the NAIC risk-based capital (RBC) charges are historically based on public corporate bond default experience. As a result, prior to the GFC, an A-rated subprime mortgage-backed security (MBS) had the same RBC charge as an A-rated US corporate bond. During the GFC, widespread rating downgrades on MBSs meant that life insurers experienced a massive increase in required capital amid the crisis. To reduce the sudden abrupt increase in regulatory capital needs of life insurers, the NAIC disconnected the RBC charge on MBSs and commercial mortgage-backed securities (CMBSs) from corporate bond ratings. Hanley and Nikolova (2021) show that these changes prompted insurers to increase their MBS and CMBS holdings and reallocate their assets into the lower credit quality/riskier holdings. Furthermore, Becker et al. (2022) analyze the effect of the reform on US life insurers' MBS holdings and find that insurers retain more downgraded MBS assets affected by the reform compared to other downgraded assets.[6]

A by-product of the NAIC reform of capital charges for MBSs and CMBSs is that it created a capital arbitrage for other ABSs, such as collateralized loan obligations (CLOs), that have low book value relative to par value. As discussed in detail in Foley-Fischer et al. (2023), CLO arbitrage allows a cut of two-thirds to the RBC charge associated with holding leverage loans directly and allows securitization risk retention arbitrage. The reduction of capital charges is a boon for life insurance companies that invest aggressively in CLOs, including PE-owned. We extend the discussion of CLO investments by life insurers in the United States in Sect. 5.3.

Kirti and Sarin (2023) show that PE-backed insurers exploit the regulatory change to decrease capital requirements by aggressively purchasing higher-yield ABSs, benefiting from the favorable regulatory treatment within days of PE acquisition. Compared to the capital charges that would be required pre-reform, PE-backed insurers' capital charges are 20% lower, and for some insurers, the savings of regulatory capital are up to 50%.

Beyond capital regulation arbitrage, it is also possible that PE firms have superior investment skills and identify underpriced bonds. However, the newly issued bonds that are acquired by the PE-owned insurers do not exhibit any significant additional return relative to bonds of all life insurers. Thus, there is no evidence that PE firms have superior investment skills. Nonetheless, the analysis of expected future losses using the annualized default rates based on bond ratings suggests that PE-owned insurers increase their current profits at the cost of greater potential future losses.

Another possible channel for delivering higher returns is increasing investment in alternative assets. Kirti and Sarin (2023) document that PE-backed insurers increase their alternative investment shares by 88%, or by 4.5%, relative to an insurance industry average of 5%. Also, a higher share of affiliated investments indicates that the PE-owned insurers use insurance premiums of fixed-annuity contracts to acquire ownership stakes in other PCs of the PE firm.

It is possible that insurers pass on some benefits of the regulatory capital savings and higher returns to their policyholders, by offering higher guarantee rates on the fixed-annuity contracts and increasing the market shares. The supporting evidence on this front, according to LIMRA reports (LIMRA 2008, 2013, 2017), is that the share of PE-backed insurers in the market increased from 3% in 2009 to about 20% in 2017. Kirti and Sarin (2023) compare the pricing of annuity contracts on PE-owned and non-PE-owned insurers and find that the guaranteed rates offered by PE-owned firms are above the 75th percentile for non-PE-backed firms.

In sum, Kirti and Sarin's (2023) analyses suggest that in the sample of 57 PE acquisitions of insurers active in the fixed-annuity market, PE ownership generates value for shareholders, and some

[6] We acknowledge and thank an anonymous referee for outlining the institutional details related to NAIC capital charges.

benefits are passed on to policyholders in the form of higher guaranteed rates. At the same time, these immediate benefits simultaneously raise potential future losses.

In the segment of the US variable annuity market, which became loss-making after the GFC, Barbu (2023) analyzes another strategy utilized by PE acquirers of life insurance companies to reduce the losses of legacy variable annuity contracts—consumer exchanges. Named after Article 1035 in the US Tax Code, a "1035 exchange" allows a policyholder to transfer funds from an existing variable annuity to another variable annuity issued by an insurance company (an original or a new one) on a tax-free basis. While the account value carries over in an exchange, the new variable annuity contract can alter or remove completely any favorable terms or benefits of the original contract. Barbu (2023) reports that over the last decade, more than $429 billion of legacy variable annuities were removed by having policyholders give up their existing policies for new, less generous policies.

While 1035 exchanges are used by various life insurance companies, the terms of the exchange are more aggressive for PE-owned life insurance companies. Barbu's (2023) interpretation is that PE-owned insurers have a lower reputation cost of exploiting the policyholders because they do not face the need to generate new business in the variable annuity segment of the market. The replacement intensity for the contracts of the PE-owned insurers was reported as 1.18% higher than that of the non-PE-owned insurers. In addition, the difference between PE-owned and non-PE-owned contracts originated among the most generous contracts. Furthermore, for those PE-owned insurers that acquired closed insurance books, the intensity was even higher, by 2.3%.

Foley-Fischer et al. (2023) document the post-GFC expansion of the life insurance industry into CLO lending. Interestingly, large, traditional life insurers themselves are very active participants in the CLO market, while the insurers controlled by PE account for about one-third of the participants in this market.

Overall, the recent literature has explored only a few mechanisms through which PE ownership has impacted life insurance companies selling traditional guarantees, and their policyholders. It appears that low interest rates could have been the main driver of interest. As conditions change, the appeal of life insurance books to external investors may deteriorate for opportunistic buyers, and some risks may materialize (see Sect. 5.2 for the recent example). As discussed by Axelson et al. (2013), variation in economy-wide credit conditions is the main determinant of leverage in buyouts.

5 Implications of Private Equity Ownership for Policyholder Protection and Financial Stability

The growing PE ownership in insurance has raised concerns for regulators in recent years. According to the IAIS (2022), supervisors need to deepen their understanding of the motives of PE firms' investment in the insurance sector and the additional risks associated with higher allocation to more illiquid and volatile assets by PE-owned insurers. The supervisors also noted that the complexity and opacity of the ownership structure hinders the identification and monitoring of the risks that PE ownership imposes on policyholder protection and financial stability. In this section, we present a summary of the regulatory concerns. In addition, we illustrate the potential risks of PE ownership using two cases: the case of Eurovita, an Italian PE-owned life insurer that was put into administration by the regulator in January 2023, and the case of the growing concentration of CLO investments in the life insurance industry in the United States, including PE-owned insurers.

5.1 Regulatory Concerns

The common mandate of insurance supervisors is to ensure policyholder protection and the financial stability of the insurance sector. To fulfill this mandate, divestment strategies of life insurers and related growing ownership by PE and other alternative asset managers in the sector requires assessments of the acquirers' ability to pay policyholder claims and to provide the services for extended periods of time consistent with the long-term nature of insurers' commitments of life insurance and annuities policies. Furthermore, the risks of the acquirers including the use of leverage, exposure to market shocks, and organizational complexity need to be weighed against the potential value added in terms of operational efficiency and asset management expertise. The current markets in the United States and Europe seem to be composed of a variety of acquirers employing highly diverse business strategies across these dimensions. Although some acquirers bring long-term capital to the life insurance industry, others are opportunistic, with short-term holding horizons and heavy use of leverage to generate returns. Further empirical research will be useful for clarifying the alignment of the acquirers' strategies with the regulatory goals.

One important aspect of the current regulatory discussions is the form of acquisition. Most divestments in the life insurance market occur in the form of the sale of the legal entity or of the portfolios of insurance contracts to another legal entity. As discussed in Kartasheva (2023), these transactions imply a complete transfer of all assets and liabilities as well as the risks of the legacy book from the insurer to the acquirer. Additionally, the acquirer becomes the ultimate obligor to the policyholders. Kartasheva (2023) analyzes an alternative strategy for divestment: funded reinsurance. This functions like a standard quota share reinsurance in the non-life insurance market. A reinsurer is paid a premium in the form of assets of the life insurer. In exchange, the reinsurer is responsible for its quota share of benefit payments to policyholders, which are reinsurance recoverable on the balance sheet of the insurer. Furthermore, the reinsurer is entitled to the investment returns that it generates by investing in the assets. While funded reinsurance naturally exposes the insurer to the counterparty risks, one benefit from the regulatory perspective is that the original insurer remains the ultimate obligor to the policyholders and has incentives to select a high credit quality reinsure to offload the risks of the legacy book.

5.2 The Case of Eurovita in Europe

The recent and still unraveling failure of Eurovita, an Italian life insurance company, illustrates the potential risks of LBO acquisitions of life insurers by PE funds. Eurovita is owned by a European PE firm, Cinven. On January 31, 2023, the Italian insurance regulator IVASS announced that Eurovita had been put into administration.[7] Furthermore, in response to the run of policyholders seeking to redeem their policies, on February 6, 2023, IVASS suspended the policyholders' right to surrender insurance and capital redemptions contracts, which was further extended on March 30, 2023.[8] The impact of the failure is significant, affecting approximately 680,000 life policies, with more than €2 billion of premiums and approximately €17 billion of gross reserves. Eurovita distributed its enhanced product offering through a diversified network of approximately 150 agents, 11,000 financial advisors, and 2500 bank branches.[9]

[7] See IVASS - Provision no. 0024821 of 31 January 2023

[8] See IVASS - Provision no. 0029903 of 6 February 2023 and IVASS - Order No. 0075800 of 30 March 2023

[9] See https://www.cinven.com/where/investments/eurovita/

The entry of Cinven into the Italian market started with the acquisition of ERGO Italia, a subsidiary of the German insurance group ERGO. The life business ERGO Previdenza managed approximately 600,000 life policies, mainly traditional products, with €5.1 billion gross reserves and a net income of €38 million. The acquisition aimed to implement a strategy of creating a market-leading life insurance consolidation platform in Italy, citing the consolidation opportunities within the Italian life insurance market as well as the intention of international insurers operating in Italy to divest their legacy books.

The current entity, Eurovita, is the result of multiple transactions. Eurovita Assicurazioni was purchased by Cinven on April 2, 2017, from JC Flowers & Co. Over the next five years, Cinven purchased Italian businesses/portfolios from multiple entities, including Skandia and PrimericaLife (part of Prudential US).[10] It grew rapidly over the next five years, distributing its products through banks.

The recent rapid and substantial increase in yields from 2021 implies that life insurers with better-matched portfolios face the challenge of recognizing significant losses on fixed-income portfolios and increased capital requirements because of lapse risk. According to the financial stability report of the Bank of Italy (Bank of Italy 2023), the average solvency ratio of Italian insurance companies decreased from 257 to 249 between June and December 2022. In life insurance, the decline was driven by the increase in the capital requirements due to the higher exposure to interest rate risk and lapse risk—that is, the risk that the policyholder surrenders the policies to reinvest in other investment products with more attractive terms (Fig. 8). In March 2023, the aggregate net unrealized losses in the Italian insurance sector amounted to €31 billion, an improvement from €52 billion losses at the end of 2022. These losses mostly impacted life insurance companies (Fig. 9).

The buildup of net unrealized losses on portfolio investment was the cause of the deteriorating capital adequacy of Eurovita (Bank of Italy 2023). In 2020, Eurovita posted a net loss of €28 million. According to the Eurovita 2021 financial report, its solvency ratio at the end of 2021 dropped below the required 150%. In 2021 and 2022, Cinven attempted to sell Eurovita. However, the sale process was suspended in 2022 as bids did not meet the target of €600–700 million set by Cinven.

In September 2022, IVASS directed Eurovita to raise its capital reserves by nearly €250 million and seek to sell the entity. The owner, Cinven, did not provide the required capital and sought to sell the entity. Their disengagement to inject additional capital and spreading of the news that the controlling shareholder was unwilling to strengthen the company's capital position triggered rapid and significant growth in surrenders (Bank of Italy 2023), endangering the insurer's liquidity position. Citing the need to safeguard financial stability, IVASS had to suspend redemptions and appoint an interim commissioner to reorganize the company. Cinven eventually agreed to provide €100 million, although Eurovita needed €400 million in 2023.

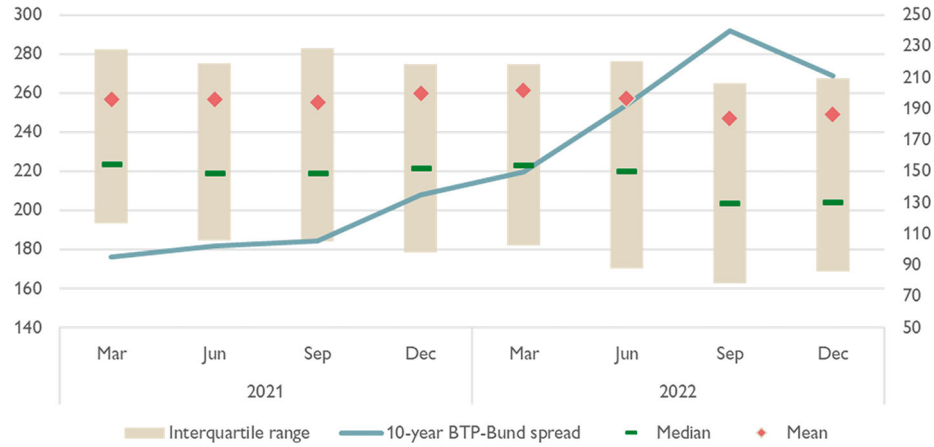
According to Reuters, the Italian authorities and financial firms are considering rescuing Eurovita by splitting it among the leading Italian insurers, including Generali, UnipolSAI, Intesa Sanpaolo, Allianz, and Poste Italiane. Insurers that have been called upon to help suggested that banks that have been distributing Eurovita's products (and collecting the fees) should also participate in the rescue.[11]

The case of Eurovita indicates at least two issues that arise from the PE-fund structure ownership of life insurance companies. The first issue is that the PE owner of the fund faces conflicting incentives of delivering returns to the LPs who invested in the fund and protecting the interests of the policyholders of the life insurance companies. It appears that in the case of Eurovita, the former incentive dominated. Indeed, since Eurovita is one of many investments of Cinven in Europe and internationally, the reputational costs of dropping the failing life insurer may be lower than the negative impact of the

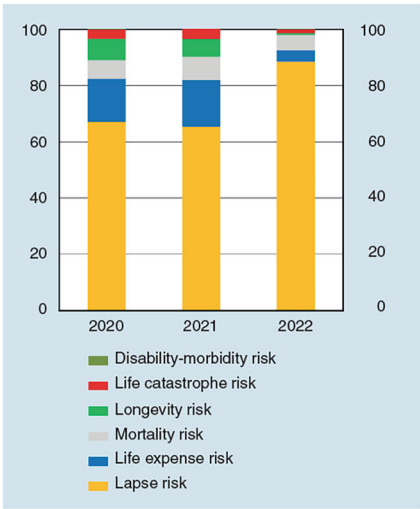
[10] See Cinven press releases: Eurovita | Investments | Where | Cinven - Focused European Integrated World-class

[11] See Reuters: Italy discusses break-up rescue for Eurovita - sources | Reuters

(a) Solvency ratio and BTP-Bund Spread (percent and basis points)



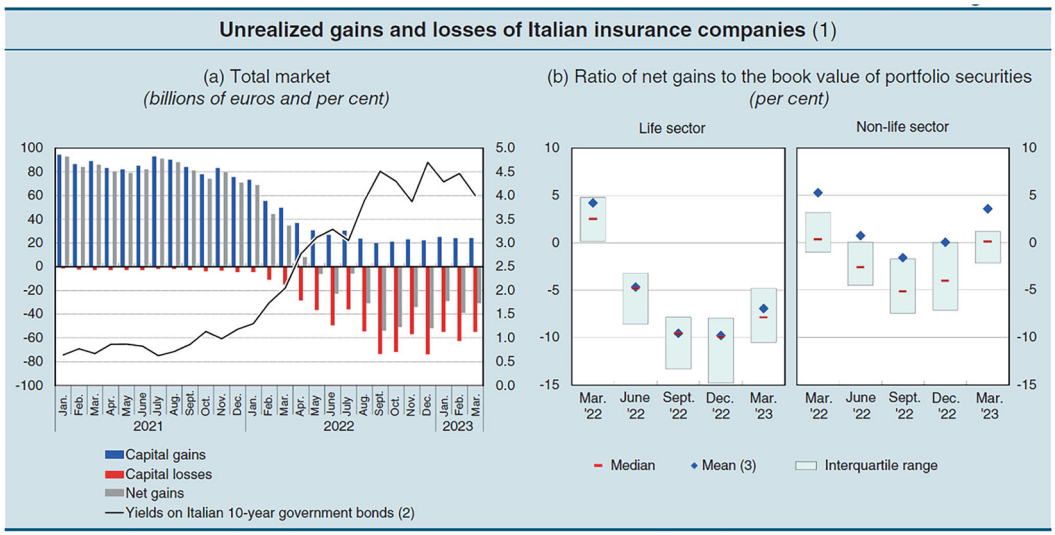
(b) Solvency capital requirement for life underwriting risks



Note: In panel (a), the solvency ratio is calculated as the ratio of eligible own funds held for coverage of the solvency capital requirement established under Solvency II. The regulations require a ratio of 100 percent or more. The data are taken from the quarterly Solvency II supervisory reports based on the quantitative reporting templates. The mean is calculated as the weighted average with weights equal to the solvency capital requirement. The BTP-Bund spread refers to the end of each period and is depicted in the right-hand scale. In panel (b), the data only consider those companies (76 undertakings representing 60 percent of total assets) that use the standard formula to calculate the solvency capital requirement (SCR). The standard method used for calculating the spread risk does not set capital requirements for exposures for an EU state that is denominated and funded in the domestic currency.

Source: IVASS and Bank of Italy.

Fig. 8 Solvency ratio and its drivers in Italy in 2021–2022. (**a**) Solvency ratio and BTP-Bund Spread (percent and basis points). (**b**) Solvency capital requirement for life underwriting risks



Note: (1) Unrealized gains and losses are the difference between the market value and the book value of portfolio securities. (2) Right-hand scale. End-of-period data. (3) Weighted average with weights equal to the book value of portfolio securities.

Source: IVASS and Bank of Italy.

Fig. 9 Unrealized gains and losses of Italian insurance companies

additional capital injection requested by the regulator to fund performance and the ability of the PE firm to attract LPs in the future.

The second important component is that the disengagement of the shareholder to provide capital imposed an externality on other companies in the life insurance sector. According to the Bank of Italy (2023), the ratio of surrenders in the life insurance sector in Italy continued to rise, especially in the early months of 2023. The surrenders were particularly prominent for insurance companies that were distributing via the banking channel or through financial advisors. One can speculate that the systemic risk imposed on the insurance sector by a run on a PE-owned insurer may be stronger the higher is the share of PE-owned life insurers in the market.

Addressing these challenges is further complicated by the fact that the owner of the life insurer and other significant participants involved in providing financing of Eurovita are outside of the regulatory perimeter of the Italian supervisory authorities. As with the rescue of failing banks, moral suasion is often a soft regulatory tool that is used in reorganizations of financial institutions. The regulator's ability to impose moral suasion is lower for entities that are outside of the Italian supervisors' regulatory perimeter, resulting in lower reputational costs.

5.3 The Case of Investment in CLOs in the United States

After the GFC, PE-owned insurers became significantly involved in the booming collateralized loan obligations (CLOs) market in the United States. The CLO market emerged, at least in part, in response to the significant reforms of the banking sector in the United States and other countries in the context of the Basel III framework after the GFC. The post-GFC regulations offer banks

incentives to hold safe assets and increase the costs of holding risky loans (BIS 2018). In response, using their expertise in originating corporate loans but unwilling to hold risky loans on the balance sheet due to higher capital requirements, banks started to sell loans to syndicates of investors who would use securitization to issue liabilities backed by a pool of below-investment grade leveraged loans—collateralized loan obligations (CLOs) (Foley-Fischer et al. 2020). Interestingly, alongside banks, US life insurers have become active participants in the CLO market by originating risky loans, holding them, and securitizing them in CLOs. According to Foley-Fischer et al. (2023), by 2020, US life insurers managed about a quarter of all CLOs in the US market, and PE firms controlled about one-third of these life insurers. In this section, we summarize the extensive analysis of the life insurers' expansion into the CLO market developed by Foley-Fischer et al. (2023) and discuss the key drivers of these developments and their implications for financial stability.

Foley-Fischer et al. (2023) label the CLO business of life insurers as shadow banking business, for the reasons we explain below. It involves a triangular organizational structure. The first type of entity in the structure is a United States-domiciled life insurance company that offers fixed-annuity contracts and grows either through writing new business or by acquiring legacy blocks of annuity business through the purchase of insurance companies, funded reinsurance transactions, or pension buyouts. Large traditional insurers active in CLO business can also tap the wholesale funding market by issuing nontraditional insurance liabilities to institutional investors, such as funding agreement-backed securities (FABS) and Federal Home Loan Bank (FHLB) advances collateralized by funding agreements.

The second type of entity is capital reinsurers located in Bermuda, who assume mortality and morbidity risks included in the annuity liabilities. As fixed annuities are savings instruments with little mortality risk, the balance sheet of the captive reinsurer is rather small. At the same time, Bermuda-domiciled captive reinsurers permit tax arbitrage, which allows insurers to dramatically decrease the cost of fixed-annuity funding for life insurers.

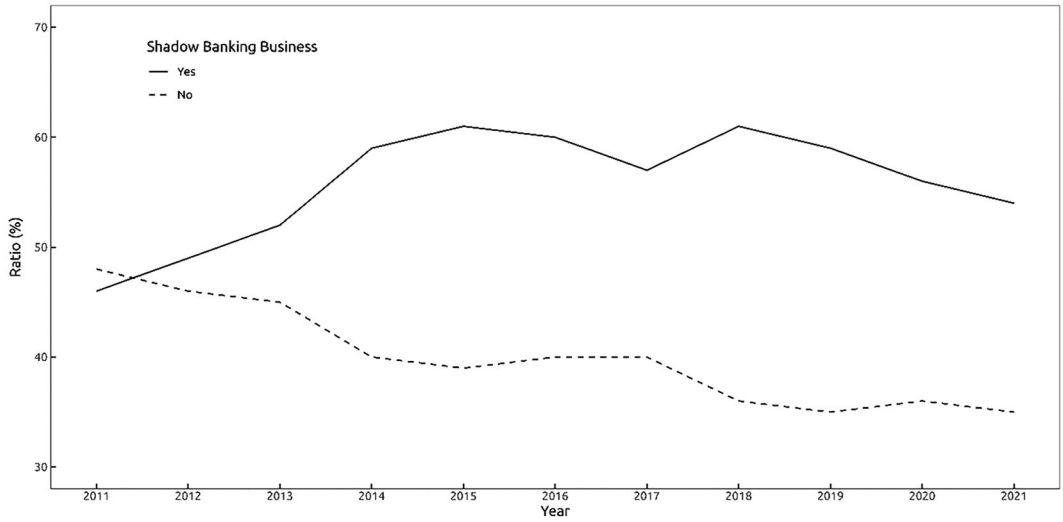
The third type of entity is an asset manager that manages all the assets of a United States-domiciled life insurer. Some large US life insurers have developed in-house asset managers, while others have acquired an existing asset manager, or both.

Foley-Fischer et al. (2023) constructed a unique, comprehensive dataset of insurance groups with shadow banking businesses. They document that there was a significant growth in insurance groups that contain shadow banking business after 2010. As shown in Fig. 10, insurers with shadow banking business account for more than half of total life insurance industry assets. Comparing insurers with and without shadow banking business, they find that the former is more leveraged than the latter.

PE-owned life insurers are also part of the trend. Figure 11 shows the rapid growth of life insurers controlled by PE firms in the United States. By 2021, PE-owned insurers controlled nearly $500 billion in assets, or roughly 10% of the US life insurance industry. On one hand, PE firms that have expertise in private credit and private assets are attracted by the CLO market development. On the other hand, the life and pension industry offers a large source of stable funding to PE firms. However, Foley-Fischer et al. (2023) argue that PE firms are not the key drivers of the CLO business in life insurance, which is dominated by the large US incumbents.

The growth of the CLOs business may potentially pose systemic risk and financial stability concerns originating from the life insurance sector, and, since the US PE-owned life insurers are outside of the regulatory perimeter of the Federal Reserve Board or the NAIC, their exposure to risks originating from US life insurers engaged in shadow banking business could potentially amplify the financial stability risks in the event of a corporate market downturn.

The vulnerability of the large incumbent US life insurers to run risk occurs due to their ability to issue shadow-banking-style liabilities, such as institutional annuities. These are primarily funding agreements that can be privately placed, used to obtain advances from FHLBs, structured



Note: The lines show the percentage of general account assets funded by deferred fixed and institutional annuities. The upward trend in the solid line relative to the dashed line shows the growth in these annuities for life insurers with shadow banking business.

Source: Foley-Fischer et al. (2023)

Fig. 10 Growth in fixed annuities of life insurers with shadow banking business

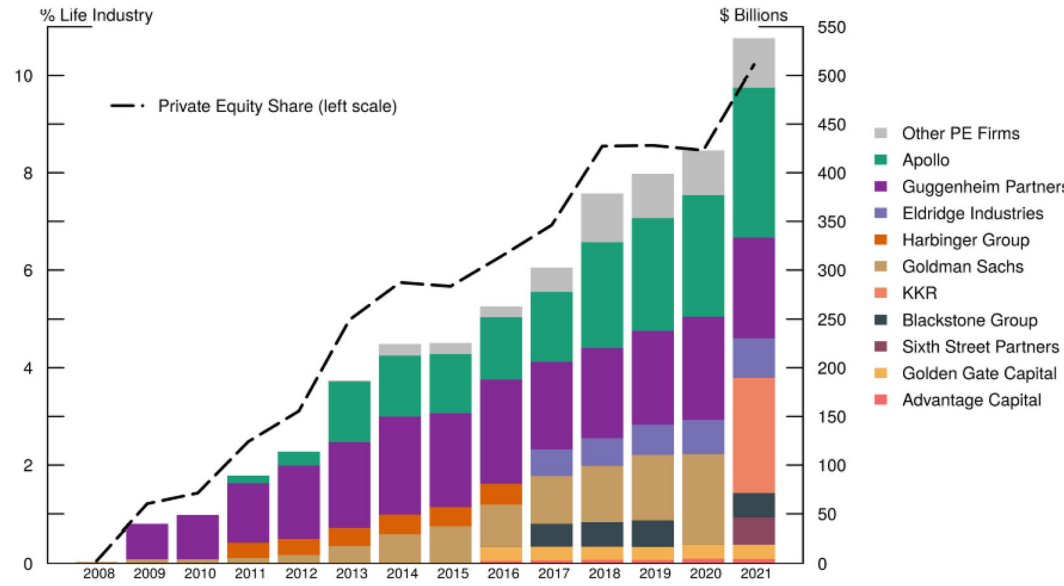


Fig. 11 US life insurer general account assets controlled by private equity firms

as guaranteed interest contracts, and securitized into FABS (Foley-Fischer et al. 2020). As these agreements are on the general account of insurers, funding agreements are senior to debt and thus reduce the cost of funding. Furthermore, these contracts are not classified as debt by regulators and

Table 2 Insurers with shadow banking business hold a disproportionately large amount of the risky tranches of CLOs issued by insurers

Dependent variable:	$CLO_holding_share_{ikt}$			
	(1)	(2)	(3)	(4)
$High_yield_CLO_k$	0.036***	0.014	0.014	0.033*
	(0.013)	(0.013)	(0.016)	(0.016)
$High_yield_CLO_k$ X $Shadow_bank_i$		0.101***	0.099**	
		(0.038)	(0.038)	
$PE\text{-}owned_{it}$				0.093
				(0.078)
$High_yield_CLO_k$ X $PE\text{-}owned_{it}$				0.073
				(0.053)
SE clustering	Firm	Firm	Firm year	Firm year
Fixed effects	Firm	Firm	Firm year	Firm year
Observations	4815	4815	4815	4905
Adjusted R^2	0.213	0.219	0.229	0.227

Source: Foley-Fischer et al. (2023)

Note: The dependent variable in all specifications ($CLO_holding_share_{ikt}$) is the fraction of all class k CLO tranches held by insurer i at the end of year t that was issued by any insurer. The independent variables are all indicator variables. $CLO_holding_share_{ikt} \in \{0, 1\}$ takes the value 1 if tranche k is not senior. $Shadow_bank_i \in \{0, 1\}$ takes the value 1 if insurer i has a shadow banking business. $PE - owned_{it} \in \{0, 1\}$ takes the value 1 if insurer i is owned by private equity at the end of year t. $*p < .1$; $**p < .05$; $***p < .01$

therefore do not count as leverage under the traditional measures. In addition, they are tax-exempt in Bermuda.

The incentives to issue liabilities in the form of funding agreements come from the ability to obtain the cheapest source of funding for life insurers with investment grade credit ratings. For example, life insurers can access wholesale funding markets by using FABS to convert funding agreements into short-term structured products. FABS also permit life insurers to transform an insurance liability into corporate debt.

Wholesale funding activities render life insurers vulnerable to runs. The exposure is similar to the GFC, where short-term institutional investors ran on certain short-term funding markets and also ran on short-term FABS issued by several other large life insurers. In some cases, insurers required substantial government assistance to mitigate the systemic impact of the runs.

Foley-Fischer et al. (2023) also document that US insurers invest heavily in CLOs. The life insurance industry holdings of CLOs have increased exponentially, reaching about $194 billion by the end of 2020, mostly concentrated in large property-casualty and life insurance groups. US insurers are attracted to CLO investments for various reasons, including the fact that CLOs offer floating rates that allow insurers to avoid having to lock into low-return long-term fixed-income investments during periods of low interest rates. However, it is the insurers with shadow banking business that hold disproportionally large amounts of the risky CLOs issued by insurers. Table 2 reports the results of the regression analysis by Foley-Fischer et al. (2023) that aimed to investigate whether insurers with shadow banking services hold a disproportionally higher share of insurer-originated risky CLO tranches. It shows that these insurers hold a 4.6 percentage points (or 19.2%) larger share of insurer-originated tranches. In addition, columns (3) and (4) explore whether PE-owned insurers are selecting the riskier tranches. The results are insignificant, suggesting that it is the shadow banking business of life insurers and not the PE ownership that drives the disproportionate holdings of risky CLO tranches.

At the same time, the participation of PE-owned insurers in the CLO market can amplify the systemic risk during a period of market downturn. For insurers with significant shadow banking business, the severe decline in the value of loans backing the CLOs can significantly erode their equity. Consequently, it might trigger a run of liquidity-sensitive institutional investors, such as FABS investors or FHLBs, and create a severe liquidity crisis for the life insurance industry. Using the COVID-19 shock, Foley-Fischer et al. (2023) estimate that insurers with larger shadow banking businesses experienced disproportionally higher stock price volatility during the pandemic. Given the size of the shadow banking business in the US life insurance industry, their synchronized failure may pose a severe burden for the US federal government.

6 Conclusion and Avenues for Further Research

Structural shifts in the life insurance industry have prompted insurers to look for investors outside of the insurance industry. Acquisitions by PE investors and alternative asset managers have allowed life insurers in Europe and the United States to reduce their exposure to traditional life insurance policies with financial guarantees. As in other sectors that PE investors target, PE ownership in insurance can improve efficiency and bring new expertise, particularly in the management of alternative assets. However, it may also introduce new risks or improve operational efficiency at the cost of other stakeholders. The recent scarce academic research in this area indicates that both elements are at play. Further analyses on the motives for entry by PE investors, the impact of the divestment on the seller and the policyholders, and the overall impact of PE participation in the life insurance industry along with traditional insurers presents an interesting and policy-relevant area of future research. We would like to highlight the following themes.

The impact of the form of transactions, namely portfolio transfer, legal entity sale, or funded reinsurance, on the stakeholders and policyholders is rather unclear. However, as discussed in Kartasheva (2023), the ability of transitional insurers to divest to private investors and eliminate exposure to their previous liabilities reduces the insurers' incentives to find a buyer with a prudent attitude toward policyholders. Portfolio transfer and legal entity same transactions can also be complex. However, the empirical assessment of the outcomes of various divestment forms is a promising avenue for further research.

Another relevant topic for future research is the role of the regulatory environment for the development of PE ownership in life insurance in the United States and Europe. The arbitrage documented in Foley-Fischer et al. (2023) appears extreme relative to the EU's Solvency II regime. Yet, PE acquisitions are booming in the EU, suggesting that the primary driver of PE acquisitions of life insurers is not a regulatory or tax arbitrage.

Finally, we know little about which PE funds choose to enter the life insurance business. From the regulatory perspective, it will be important to assess whether PE firms with lower liquidity are the ones that select to acquire life insurers with sticky and stable liabilities. These companies would be more exposed to insolvency during periods of market downturn.

Acknowledgments I am grateful to anonymous referees for insightful comments and suggestions, and to Nathan Foley-Fischer, Tereza Tykvová, and Stéphane Verani for useful input and discussions. I would like to thank Marcel Frei, Jessica Meyer, and Carolin Schmitt for their excellent research assistance.

References

Acharya VV, Gottschalg OF, Hahn M, Kehoe C (2013) Corporate governance and value creation: evidence from private equity. Rev Financ Stud 26:368–402

Appelbaum E, Batt R (2020) Private equity buyouts in healthcare: who wins, who loses? (March 15, 2020). Institute for New Economic Thinking Working Paper Series No. 118. Available at: https://ssrn.com/abstract=3593887. Accessed 10 Oct 2023

American Council of Life Insurers (2022) Life insurers fact book. Available at: 2022lifeinsurersfactbook_v2.pdf (acli.com). Accessed 18 Oct 2024

Aslan H, Kumar P (2011) Lemons or cherries: growth opportunities and market temptations in going public and private. J Financ Quant Anal 46:489–526

Axelson U, Jenkinson T, Strömberg P, Weisbach MS (2013) Borrow cheap, buy high? The determinants of Leverage and pricing in buyouts. J Financ 68:2223–2267

Bank for International Settlements (BIS) (2018) Structural changes in banking after the crisis. CGFS paper No. 60. Available at: https://www.bis.org/publ/cgfs60.htm. Accessed 22 Oct 2023

Bank of Italy (2023) Financial stability report No.1. Available at: https://www.bancaditalia.it/pubblicazioni/rapporto-stabilita/2023-1/index.html. Accessed 10 Oct 2023

Barbu A (2023, September 19) Ex-Post Loss Sharing in Consumer Financial Markets. Available at SSRN: https://ssrn.com/abstract=4079524

Becker B, Ivashina V (2015) Reaching for yield in the bond market. J Finance 70(5):1863–1902

Becker B, Marcus MO, Farzad S (2022) Regulatory forbearance in the US insurance industry: the effects of removing capital requirements for an asset class. Rev Financ Stud 35(12):5438–5482

Bernstein S, Sheen A (2016) The operational consequences of private equity buyouts: evidence from the restaurant industry. Rev Financ Stud 29(9):2387–2418

Bernstein S, Lerner J, Mezzanotti F (2019) Private equity and financial fragility during the crisis. Rev Financ Stud 32(4):1309–1373

Bloom N, Sadun R, Van Reenen J (2015) Do private equity owned firms have better management practices? Am Econ Rev 105:442–446

Boucly Q, Sraer D, Thesmar D (2011) Growth LBOs. J Financ Econ 102:432–453

Brunetti C, Foley-Fisher N, Verani S (2023) Measuring interest rate risk management by financial institutions. Finance and economics discussion series 2023-067. Board of Governors of the Federal Reserve System, Washington. https://doi.org/10.17016/FEDS.2023.067

Chen G, Kang JK, Kim JM, Na HS (2014) Sources of value gains in minority equity investments by private equity funds: evidence from block share acquisitions. J Corp Financ 29:449–474

Cohn JB, Mills LF, Towery EM (2014) The evolution of capital structure and operating performance after leveraged buyouts: Evidence from US corporate tax returns. J Financ Econ 111:469–494

Cornelli F, Kominek ZW, Ljungqvist A (2013) Monitoring managers: does it matter? J Financ 68:431–481

Davis Steven J, Haltiwanger J, Handley K et al (2014) Private equity, jobs, and productivity. Am Econ Rev 104(12):3956–3990

Demiroglu C, Christopher MJ (2010) The information content of bank loan covenants. Rev Financ Stud 23(10):3700–3737

Domanski D, Shin HS, Sushko V (2017) The hunt for duration: not waving but drowning? IMF Econ Rev 65:113–153. https://doi.org/10.1057/s41308-016-0026-9

Driessen J, Lin TC, Phalippou L (2012) A new method to estimate risk and return of nontraded assets from cash flows: the case of private equity funds. J Financ Quant Anal 47:511–535

Eaton C, Howell ST, Yannelis C (2020) When investor incentives and consumer interests diverge: private equity in higher education. Rev Financ Stud 33(9):4024–4060

Edgerton J (2012) Agency problems in public firms: evidence from corporate jets in leveraged buyouts. J Financ 67:2187–2213

EIOPA (2022) Financial stability report, December. Available at: Financial Stability Report December 2022 (europa.eu). Accessed 10 Oct 2023

Ellul A, Jotikasthira C, Kartasheva A, Lundblad C, Wagner W (2022) Insurers as asset managers and systemic risk. Rev Financ Stud 35(12):5483–5534

Foley-Fischer N, Narajabad B, Verani S (2020) Self-fulfilling runs: evidence from the US life insurance industry. J Polit Econ 128(9):3520–3569

Foley-Fischer N, Heinrich N, Verani S (2023) Are life insurers the new shadow banks?. Available at: https://ssrn.com/abstract=3534847. Accessed 10 Oct 2023

Fracassi C, Previtero A, Sheen A (2020) Barbarians at the store? Private equity, products, and consumers. NBER Working Paper No. w27435
Gandhi A, Song Y, Upadrashta P (2020a) Private equity, consumers, and competition: evidence from the nursing home industry. Available at: https://ssrn.com/abstract=3626558. Accessed 10 Oct 2023
Gandhi A, Song Y, Upadrashta P (2020b) Have private equity owned nursing homes fared worse under COVID-19? Available at: https://ssrn.com/abstract=3682892. Accessed 10 Oct 2023
Gao J, Sevilir M, Kim YS (2021) Private equity in the hospital industry. European Corporate Governance Institute – Finance Working Paper No. 787/2021. Available at: https://ssrn.com/abstract=3924517. Accessed 10 Oct 2023
Guo S, Hotchkiss ES, Song W (2011) Do buyouts (still) create value? J Financ 66:479–517
Gupta A, Howell ST, Yannelis C, Gupta A (2021) Does private equity investment in healthcare benefit patients? Evidence from nursing homes. Working papers 2021–20. Becker Friedman Institute for Research in Economics
Hanley KW, Nikolova S (2021) Rethinking the use of credit ratings in capital regulations: evidence from the insurance industry. Rev Corp Fin 10(2):347–401
Harris RS, Jenkinson T, Kaplan SN (2014) Private equity performance: what do we know? J Financ 69:1851–1882
Hartley D, Paulson A, Rosen RJ (2017) Measuring interest rate risk in the life insurance sector: the US and the UK. In: Hufeld F, Koijen RSJ, Thimann C (eds) The economics, regulation, and systemic risk of insurance markets. Oxford University Press, Oxford, pp 124–150
IAIS (2022) Global insurance market report, December. International Association of Insurance Supervisors. Available at: GIMAR-2022.pdf (iaisweb.org). Accessed 16 Oct 2023
Ivashina V (2022) When the tailwind stops: private equity industry in the new interest rate environment. LTI report 1. CEPR Press
Ivashina V, Lerner J (2019) Patient Capital. Princeton University Press
Ivashina V, Sun Z (2011) Institutional demand pressure and the cost of corporate loans. J Financ Econ 99(3): 500–522
Jegadeesh N, Kraussl R, Pollet JM (2015) Risk and expected returns of private equity investments: evidence based on market prices. Rev Financ Stud 28:3269–3302
Jenkinson T, Kim H, Weisbach M (2021) Buyouts: a primer. NBER working paper No. 29502. Available at: w29502.pdf (nber.org)
Kaplan SN, Schoar A (2005) Private equity performance: returns, persistence, and capital flows. J Financ 60:1791–1823
Kartasheva A (2023) Funded reinsurance as a divestment strategy for life insurance books with financial guarantees. Working paper. University of St Gallen
Kirti D, Sarin N (2023) What private equity does differently: evidence from life insurance. Rev Financ Stud. https://doi.org/10.1093/rfs/hhad055
Koijen Ralph SJ, Yogo M (2022a) The fragility of market risk insurance. J Financ 77(2):815–862
Koijen Ralph SJ, Yogo M (2022b) Global life insurers during a low interest rate environment. AEA Papers Proc 112:503–508
Koijen Ralph SJ, Yogo M (2023) Understanding the ownership structure of corporate bonds. Am Econ Rev Insights 5(1):73–92
Korteweg A, Nagel S (2016) Risk-adjusting the returns to venture capital. J Financ 71:1437–1470
Lerner J, Sorensen M, Stromberg P (2011) Private equity and long-run investment: the case of innovation. J Financ 66:445–477
LIMRA (2008) LIMRA secure retirement institute US individual annuities sales survey
LIMRA (2013) LIMRA secure retirement institute US individual annuities sales survey
LIMRA (2017) LIMRA secure retirement institute US individual annuities sales survey
Liu T (2021) Bargaining with private equity: Implications for hospital prices and patient welfare. Available at: https://ssrn.com/abstract=3896410. Accessed 10 Oct 2023
NAIC (2020) Private equity (PE) owned US insurers as of year-end 2019. Capital markets special report. Available at: capital-markets-special-report-private-equity-owned.pdf (naic.org). Accessed 10 Oct 2023
NAIC (2022) Private equity (PE) owned US insurers investments decrease as of year-end 2021: Number of PE-owned US insurers increases. Capital Markets Special Report. Available at: PE-Owned YE2021 (naic.org). Accessed 10 Oct 2023
Phalippou L (2014) Performance of buyout funds revisited? Rev Financ 18:189–218
Sen I (2023) Regulatory limits to risk management. Rev Financ Stud 36(6):2175–2223
Sensoy BA, Wang Y, Weisbach MS (2014) Limited partner performance and the maturing of the private equity industry. J Financ Econ 112:320–343
Shleifer A, Summers L (1988) Breach of trust in hostile takeovers. In: Auerbach AJ (ed) Corporate takeovers: causes and consequences. University of Chicago Press, Chicago
Sorensen M, Wang N, Yang J (2014) Valuing private equity. Rev Financ Stud 27:1977–2021

Spaenjers C, Steiner EM (2021) The value of specialization in private equity: evidence from the hotel industry. HEC Paris Research Paper No. FIN-2020-1410. Available at: https://ssrn.com/abstract=3738265. Accessed 10 Oct 2023
Tykvová T (2018) Venture capital and private equity financing: an overview of recent literature and an agenda for future research. J Bus Econ 88:325–362
Tykvová T, Borell M (2012) Do private equity owners increase risk of financial distress and bankruptcy? J Corp Financ 18:138–150
Verani S, Yu P (2021) What's wrong with annuity markets? Working paper. Federal Reserve Board

Insurance: Private or Social?

Peter Zweifel

1 Introduction

At present, private and social insurance jointly claim more than 40% of a worker's pay in industrial countries, with Canada and the United States markedly below but other OECD countries above this benchmark (see Table 1 and footnote). Although measured as benefits in the case of social insurance (SI henceforth) and premiums paid for life and health and life coverage in the case of private insurance (PI) for lack of consistent international data, there is no doubt that PI and SI together constitute a major expenditure item in today's household budgets.[1] Admittedly, personal lines of PI comprise more than just health and life; yet it is evident from Table 1 that even in the United Kingdom, one of the major markets for personal PI, the SI component is about twice as important as the PI component, measured relative to GDP. In the United States, it is even about six times as important (SI figures for 2020 and 2021 were not yet available at the time of writing). Both the high values of SI and PI combined and the preponderance of SI are mainly the result of a remarkable expansion of SI. In Germany, for instance, SI relative to GDP was already at a comparatively high value of 22.1% in 1980 but continued to increase to 26.6% until the year 2000, stabilizing since then. The PI value was at 2.4% in 1990 (the first year with comparable data), nine times lower than the SI figure of 21.7%, a ratio that also obtains for 2019. Detail not reported in Table 1 but provided by OECD shows that the expansion of SI was led by health, with pensions second; accordingly, the term "social insurance" as used in this Chapter is more comprehensive than "social security" in US language.

[1] PI is offered by companies that are typically profit-seeking and charge a premium for the coverage of a risk. Examples are health, life, auto, property, and liability insurance as well as several types of business insurance (which are not discussed here because they do not constitute substitutes to SI. SI also covers some of the same risks (in addition to unemployment and old age) but is typically mandatory, financed by income-related contributions, and written by a public entity.

P. Zweifel (✉)
University of Zurich, Zurich, Switzerland
e-mail: peter.zweifel@uzh.ch

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_12

Table 1 Social (SI) and private (PI) insurance in some OECD countries relative to GDP and to labor income YL, in%

Country		1980	1990	2000	2010	2019	2020	2021	% of YL[b]
Canada	SIPI	13.2	17.5	15.7	17.8	18.0[a]3.1	n.a.3.5	n.a.3.3	31.7
France	SIPI	20.8n.a.	24.93.4	27.76.7	32.47.4	31.06.0	n.a.5.1	n.a.6.1	55.5
Germany	SIPI	22.1n.a.	21.72.4	26.63.0	27.13.4	25.92.6	n.a.2.8	n.a.2.6	42.8
Italy	SIPI	18.0n.a.	20.00.6	23.33.5	27.78.4	28.26.2	n.a.5.3	n.a.6.9	51.6
Japan	SIPI	10.4n.a.	11.37.0	16.58.1	22.37.5	22.86.9	n.a.5.8	n.a.6.1	44.6
United Kingdom	SIPI	10.5n.a.	16.86.8	18.612.8	23.810.4	20.68.0	n.a.8.8	n.a.6.9	42.9
United States	SIPI	13.2n.a.	13.54.1	14.54.5	19.82.9	18.72.9	n.a.3.0	n.a.2.6	32.4

Sources: OECD Social Expenditure Database, several editions; Sigma of Swiss Re, several editions
[a]2018 for SI
[b](SI+PI)/YL, as of 2019. Labor's share is put at two-thirds of GDP; thus one arrives at a combined PI & SI share in workers' pay of 42.7% (=25.2 + 3.4)/0.67) for Canada
SI: Includes benefits (e.g., for housing) that are part of public welfare
PI: Premiums paid for life and health insurance, estimated from graphs published by Sigma 4/1992, 6/2001, 2/2010, and 4/2022 (Table IX)

Much of the economics literature has justified the existence and even the preponderance of SI by its efficiency-enhancing properties, arguing that SI can mend or at least mitigate market failures of private insurance (PI). However, early on, Feldstein (1995) estimated that the privatization of US pensions would increase economic welfare by $1.5 per Dollar of net social security wealth.

This contribution seeks to shed light on the relative importance of PI and SI and its change over time in four ways. First, it reviews the efficiency reasons supporting the view that there is a need for SI, possibly complemented by PI. The famous Rothschild and Stiglitz (1976) model can be combined with the "market for lemons" dynamics described by Akerlof (1970) to predict the possibility of a "death spiral" for PI markets in the presence of insurers' lack of information about consumers' true risk status. Second, some empirical evidence concerning the alleged "death spiral" and insurer behavior in the face of their lack of information is examined. Third, the alternative public choice view emphasizing the interests of citizens and voters in wealth redistribution through SI is expounded. Fourth, normative issues are addressed by asking whether and in which lines of insurance the future division of labor between PI and SI could be improved.

2 The Efficiency View of Social Insurance

Theoretical explanations of the existence of SI usually refer to two main shortcomings of private insurance markets due to asymmetric information, moral hazard, and adverse selection.

2.1 Moral Hazard as a Failure of Both Private and Social Insurance

The observable consequence of moral hazard is a positive correlation between the degree of insurance coverage and the probability of loss (*ex ante* moral hazard), and the amount of loss (*ex post* moral hazard), as argued by Chiappori and Salanié (2000).[2] To the extent that the loading contained in insurance premiums increases along with expected loss (e.g., because of increasing cost of administration and risk bearing), the true cost of insurance rises.[3] According to Arnott and Stiglitz (1991), this is but one of four market failures caused by moral hazard.

[2] The distinction between the two types of moral hazard will not be pursued further because it is not relevant in the present context.

However, SI is subject to the same moral hazard effects as PI unless one is willing to assume that institutions of SI can observe insureds' behavior. In the case of health, accidents, disability, and old age, this is clearly not true. Whereas private insurers usually tailor the parameters of their contracts to individual behavior reflecting (the control of) moral hazard (e.g., by granting rebates for no claims), SI is strongly bound to the solidarity principle which requires equal *ex ante* benefits for equal (rates of) contributions. Therefore, the argument that SI has an efficiency advantage over PI due to its lower loading holds with regard to the cost of administration at best. But administrative expense may well reflect effort at controlling moral hazard (Danzon 1992). For instance, it is lower in Canadian health insurance (which continues to rely on fee-for-service) than in US health insurance (which has created Managed Care (MC) alternatives to combat moral hazard). In Switzerland, competitive social health insurers offer conventional fee-for-service plans with a choice of deductible in excess of the legal minimum (CHF 300 annually, 1 CHF = 1.20 USD in 2006) and MC-type plans in return for a lowered premium. After controlling for the difference in health status in individual data provided by a major insurer, Trottmann et al. (2012) find that the Individual Practice (IPA) variant of MC was associated with a greater reduction in healthcare expenditure than even the conventional plan with the highest deductible (≥CHF 1000). While managing an IPA certainly is more expensive than administering a range of deductibles,[4] the precise cost difference is unknown, obviating a comparison in terms of cost-effectiveness.

One might argue that compared to PI, SI can better count on physicians as agents for the verification of health claims (Arrow 1963). However, Dionne and St-Michel (1991) find evidence suggesting that physicians helped workers to benefit from an increase in the generosity of the public workplace accident scheme of Quebec. Eliason et al. (2018) also found that physicians failed to act as agents in the interest of SI. In Sweden, the replacement ratio of social disability insurance was lowered in 1991 from 90% to 65% when absence from work was less than 3 days, to 80% when it was 4–90 days, and unchanged when it was more than 90 days. Since the benefits of one long absence clearly exceeded those associated with several short ones, this reform discouraged early return to the workplace. This unintended effect could have been counteracted by the physicians, but it was in a mere 8% of consultations involving absence from work that they refused to issue a medical certificate. Therefore, SI was once again not able to enlist the support of physicians as intermediaries to rein in moral hazards.

The one remaining component of the loading is acquisition expense, which is negligible in the case of SI because SI constitutes a monopoly. Its cost advantage therefore has to be weighed against the imposed uniformity of promised benefits, which entails an efficiency loss as soon as preferences with regard to insurance differ within the population, for which there is substantial evidence (Cohen and Einav 2007; Cutler et al. 2011; Ericson et al. 2012).

Conclusion 1: Although moral hazard is widely regarded as a cause of market failure of PI markets, the ability of SI to rein it in appears to be no greater. In particular, physicians fail to act as agents controlling moral hazard in accident and health insurance, while the lower acquisition cost of SI has to be weighed against the efficiency loss due to its neglect of preference heterogeneity.

2.2 Adverse Selection: The Crucial Market Failure of Private Insurance?

Ever since Rothschild and Stiglitz (1976), the crucial drawback of PI is seen to originate from the other consequence of asymmetric information, adverse selection. Insurers do not (and never will) know the true type of consumers, reflected by the probability of loss. Therefore, PI has to fall back

[3] The true price of actuarially fair insurance is zero on expectation because the premium flows back to the consumer in the guise of loss payments on average.

[4] Private communication with Maria Trottmann.

on a pooling contract whose premium reflects the average probability in the population. Since high risks have a higher demand for insurance than low ones at the given premium, adverse selection also induces a positive correlation between the amount of coverage and the probability of loss, very much the same as in the case of moral hazard (Chiappori and Salanié 2000). Given that the trade-off between premium and coverage is valued differently by high and low risks, a competitor can always challenge an incumbent by offering a contract with less coverage but a lower premium that appeals to the low but not the high risks. The incumbent is stuck with its high risks and forced to increase its pooling premium to maintain financial equilibrium. This only strengthens the incentive of low risks to leave the incumbent, triggering a "death spiral" which may even wipe out the entire PI market because the challenger necessarily writes a pooling contract which is not sustainable either.

2.3 Empirical Evidence Concerning Adverse Selection

Since much of the evidence concerning adverse selection comes from the United States and revolves around private health insurance, two periods are distinguished in this section. Prior to the Affordable Care Act of 2010 (ACA, aka Obamacare), health insurance coverage was mainly negotiated by employers, who prevented health insurers from engaging in risk selection in a major way by imposing open enrollment. After 2010, so-called ACA exchanges were created, which are characterized by individual rather than group contracts, creating much more scope for risk selection.

Evidence Prior to the Affordable Care Act

An early analysis of term life insurance premiums by Cawley and Philipson (1999) finds that unit premiums decrease rather than increase with the amount of coverage (compensating for the positive correlation between coverage and probability of loss). Moreover, low risks hold more coverage than high ones, contradicting the notion of a competing insurer siphoning off low risks with a low-premium, low-coverage alternative. Moreover, this strategy may miss its target[5]: Using data from a Portuguese auto insurer covering 3 years, Przemysław et al. (2019) found that it poached customers from competitors by offering them a higher discount on the premium, although these customers generated 21% higher cost[6]. It appears that insurers weigh additional premium income against the cost of adverse selection (which usually accrues with a lag), causing them to behave in a way different from what is predicted by theory.

In the case of US health insurance, Einav et al. (2010) are able to estimate the effect of adverse selection on premiums and market volume while excluding moral hazard. In 2004, Alcoa let the presidents of some 40 business units set their employees' contribution to health insurance. Since they did not have access to information about past healthcare expenditures, the contributions did not reflect moral hazard. The authors find that the insurer's marginal and average costs increase significantly with the premium, pointing to adverse selection. They estimate a demand curve to determine the efficient equilibrium point where price equals marginal cost. The divergence caused by the break-even condition given adverse selection, price = average cost, results in a "Harberger triangle," indicating an annual welfare loss of some $10 per employee. Relative to the total welfare attainable from purchasing health insurance (the aggregated excess of marginal willingness-to-pay over marginal cost), this loss amounts to some 3%. For all the theoretical emphasis on adverse selection as a source of inefficiency of PI, this is a remarkably small figure.

[5] This can also occur in competitive social health insurance because high risks may return to low risk status and vice versa, as found by Beck et al. (2010) in the case of Switzerland.

[6] This reference could be retrieved thanks to Christophe Courbage (Geneva)

Using the same Alcoa dataset, Einav et al. (2013) investigate whether risk selection is also pursued by the insured rather than the insurer as conventionally assumed. From the transitions to contracts with high copayment between 2003 and 2004, they identify low-moral hazard types (and vice versa). Given their choice, these types exhibited $130 lower healthcare expenditure ceteris paribus in 2004. One can draw two conclusions from this finding. First, risk selection does not only entail the threat of adverse selection to insurers but may be welfare-enhancing by allowing consumers to control their moral hazard through their choice of contract. Second, risk adjustment (RA)[7] needs to target not only the behavior of I insurers but also of $N>>I$ consumers, who have an incentive to choose an insurer that can offer a low premium thanks to the cross-subsidy obtained through RA, resulting in risk selection as well. But Tinbergen's (1952) law states that the number of instruments must equal the number of objectives to be achieved, a condition violated by current RA schemes (see Zweifel and Frech 2016 as well as Frech and Zweifel 2017 for details).

Evidence Subsequent to the Affordable Care Act

The ACA exchanges created by the Affordable Care Act of 2010 exchanges are prone to adverse selection, as argued early on in spite of risk adjustment (Silow-Carroll et al. n.d., ca. 2012).[8] Chandler (2015) even saw signs even of a "death spiral," noting that from 2015 to 2016, the preferred provider organization (PPO) variant of the so-called Platinum plans (the most generous type available and hence attractive to the chronically ill) increased premiums by 25.6% on average. At the same time, in just 76.7% of rating areas, at least two of those plans were available, down from 92.5%. However, according to McDermott and Cox (2020), the share of ACA enrollees in counties with two or more participating health insurers began to increase again after 2017, suggesting that a "death spiral" was not occurring in the longer run thanks to decreasing loss ratios (Cox et al. 2019).

Colorado was one of the states adopting ACA exchanges, which became operational at different times across the state, permitting Panhans (2019) to achieve exogeneity of premiums in his estimate of the annual efficiency loss caused by adverse selection. With $312 per enrollee and year, it is far higher than the $10 obtained by Einav et al. (2010) for the pre-ACA era. This difference is to be expected because ACA exchanges predominantly cater to individuals, permitting insurers to discriminate against high-risk types using also attributes not accounted for by RA. This results in reduced levels of enrollment by any participating insurer and likely in the aggregate as well, and hence a relatively high welfare loss due to adverse selection. In contrast, Einav et al. (2010) analyzed a case where Alcoa, as the employer, imposed an open enrollment condition on participating insurers, meaning that they had to accept employees regardless of their risk status.

Using individual-level data from the ACA Exchange of California, Saltzman et al. (2021) studied adverse selection in a market dominated by just four health insurers during 2014–2018. The exogeneity of premiums in the author's structural model derives from the fact that regulations concerning penalties (for abstaining from buying coverage) and subsidies changed especially during the phase-in period. The authors find a high degree of inertia caused by an increase of 44% in average

[7] Risk adjustment is a secondary regulation designed to neutralize competitive insurers´ incentive to engage in risk selection induced by premium regulation such as community rating (Pauly 1984). It makes insurers with a below-average share of high risks pay into the scheme and uses the proceeds to subsidize insurers with an above-average share of high risks.

[8] In 2003, a partial privatization of Medicare (so-called Medicare Advantage and Medicare Part D covering drugs) occurred. However, these changes are minor compared to those introduced by the Affordable Care Act. The main issues are whether consumers or insurers benefit more from them (see, e.g., Cabral et al. (2018), who find that plans participating in Medicare Advantage benefit from advantageous rather than adverse selection), and whether creating a public option in Medicare Part D would increase overall welfare by lowering drug prices thanks to the government's monopsony while reducing pharmaceutical profits (Miller et al. 2019).

premiums associated with a change of insurer, which slows down the risk selection process. The downside is that inertia preserves insurers´ market power.

Findings of this type leave open one important question, whether a competitive market for PI would exist at all in view of the threat of "death spirals" occurring. An early case study by Cutler and Reber (1998) suggests this to be a possibility. When Harvard University employees had to come up with a much higher personal contribution to health insurance, those with a favorable cost record migrated from the contract with comprehensive coverage to a more restrictive but cheaper alternative, while those with an unfavorable record kept their comprehensive policies. Within 2 years, the more generous contracts had to be withdrawn. However, it is not clear that the insurers writing them approached insolvency. Rather, the evidence suggests that they withdrew loss-making contracts to find a new equilibrium, as theoretically described by Wilson (1977) and Miyazaki (1977). This was also the response of Prudential, who in 1981 ceased selling a particular individual health insurance plan in California; but it was only a decade later that the number of policies in force dropped to 5000 (from 270,000), a process which Frech and Smith (2015) dubbed "a truly slow-motion death spiral."

Conclusion 2: At least in the case of group health insurance, the efficiency loss caused by adverse selection seems to be surprisingly small. However, it is likely much larger in the exchanges created by the Affordable Care Act, which mostly cover individuals. Also, it causes insurers to withdraw particular loss-making plans rather than the PI market to unravel in the guise of a "death spiral."

2.4 Separating Contracts and Efficiency Enhancement by Social Insurance

The intuition behind separating contracts is that through appropriate design, consumers of a given risk type can opt for the contract appropriate for them. However, this subjects the low risks to rationing of coverage. Compulsory SI serves to relax this rationing and therefore acts like a complementary element enhancing the efficiency of PI (Dahlby 1981). Since many of the later contributions refer to her analysis, it is described in some detail.

Along the two insurance lines P^{H} and P^{L} in Fig. 1, a competitive insurer breaks even, with costs of administration and risk bearing neglected for simplicity [the insurance line $P^{\mathrm{H}}(\lambda)$ will be explained below]. Premiums therefore just cover expected loss given by $\pi^{\mathrm{H}} \cdot LOSS$ and $\pi^{\mathrm{L}} \cdot LOSS$, respectively. Indifference curves $\overline{EU}^{\mathrm{H}}$ and $\overline{EU}^{\mathrm{L}}$ reflect the subjective trade-offs between additional coverage and higher premium. The slope of the $\overline{EU}^{\mathrm{H}}$ curve must be steeper than that of $\overline{EU}^{\mathrm{L}}$ because the high risks by definition are confronted with a higher probability of loss π^{H} than the low risks; therefore, they are willing to accept a higher extra premium in return for an increase in coverage I.

Generally, the marginal utilities of wealth in the loss and the no-loss state are involved in the definition of such an indifference curve. However, when the wealth levels are equal ($I^* = Loss$ in Fig. 1) the two marginal utilities of wealth are equal as well, causing the indifference curves $\overline{EU}^{\mathrm{H}}$ and $\overline{EU}^{\mathrm{L}}$ to have slopes π^{H} and π^{L}, respectively. With the tangency condition for an optimum satisfied, points H^* and L^* denote the first-best optima in the absence of asymmetric information.

Since a pooling contract is not sustainable in the presence of informational asymmetry, the insurer may launch a pair of separate contracts. A high-premium, full-coverage contract is designed for the high risks, who therefore attain point H^* of Fig. 1. A low-premium contract would attract high risks as well if it were to offer full coverage (according to the indifference curve $\overline{EU}^{\mathrm{H}}$, a high-risk prefers point L^* to H^*). Indeed, benefits pertaining to this second contract cannot be extended beyond point $\tilde{L}$, where the indifference curve $\overline{EU}^{\mathrm{H}}$ intersects the insurance line labeled P^{L}. Evidently, separating contracts can avoid unsustainable pooling, but at a price. They entail a rationing of coverage for the

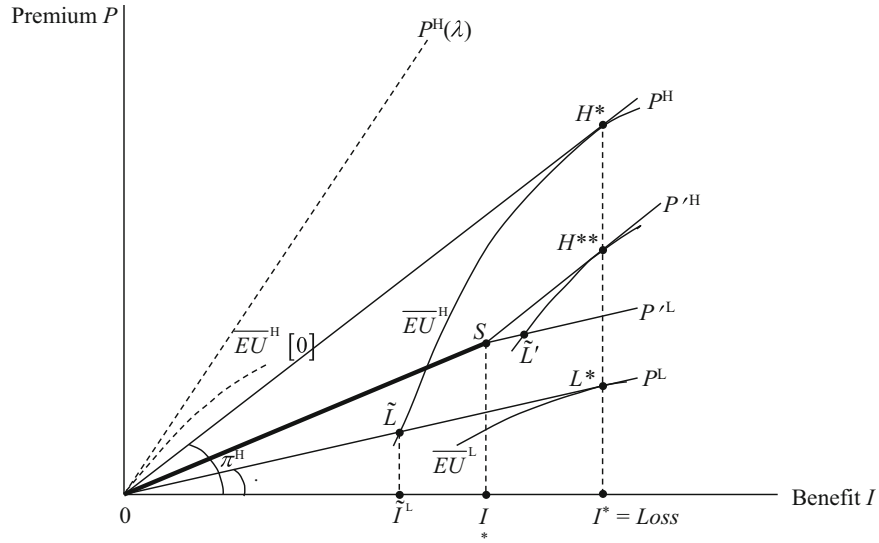


Fig. 1 Potentially Pareto-improving social insurance. (Notes: There are four insurance lines showing the relationship between premium und benefit. The steepest ($P^{H}(\lambda)$) shows the case of the high-risk type characterized by a high loss probability π^{H} as well as a proportional loading λ for the higher variance of losses. The flattest insurance line pertains to the low-risk type with $\pi^{L} < \pi^{H}$. The solid line reflects the pooling of risks imposed by SI with partial coverage. For complementary PI, this shifts the origin from 0 to *S*. Preferences are depicted by indifference curves defined by constant expected utility $\overline{EU}$; they show the subjective trade-off between benefit and premium to be paid. Being exposed to a higher probability of loss, the high-risk type is willing to accept a higher extra premium for more coverage than the low-risk type. However, the loading can be so high that they prefer to go without coverage (dashed lines). In the absence of a loading, the combination of SI and PI would make full coverage at tangency points H^* and H^{**} optimal; but due to asymmetric information, PI cannot offer coverage beyond $\tilde{L}'$ to low-risk types lest high-risk types infiltrate the contract designed for them)

low risks, although they would be willing to pay the necessary premium for the additional coverage (note that L^* lies on a higher-valued indifference curve than $\tilde{L}$).

In this situation, SI can indeed improve the welfare of both risk types. Since it is mandatory, it amounts to a pooling contract; since it has a uniform benefit, its pertinent insurance line ends at point *S* (assuming partial coverage). For complementary PI, point *S* becomes the new origin. If moral hazard effects caused by another insurer providing additional coverage are neglected, the loss probabilities remain unchanged, resulting in the new insurance lines P'^{H} and P'^{L}, respectively. This permits the high risk to attain H^{**}, an improvement of welfare over H^*. The benefit of SI for the low risks is that PI can now offer complementary coverage up to $\tilde{L}'$ (where the high risk's indifference curve through H^{**} intersects the new insurance line P'^{L}). Yet the welfare comparison between points $\tilde{L}'$ and $\tilde{L}$ is ambiguous, depending on the curvature of $\overline{EU}^{H}$. Therefore, SI (of an undefined amount) may but need not be Pareto-improving according to this model.

Coverage provided by mandatory SI can be strictly Pareto-improving, as shown by Besley (1989), under the crucial assumption that SI is able to observe the risk type. However, the amount of coverage provided by PI must fall in order to limit moral hazard effects, an effect that was to be dubbed "crowding-out" later (see Sect. 4). By way of contrast, Blomqvist and Johansson (1997) argue that in the absence of observability, moral hazard effects caused by PI spill over to SI, making a mixed system strictly less efficient than either the fully private alternative or comprehensive SI.

Boadway et al. (2006) examine the contribution of SI to efficiency in the presence of both (*ex post*) moral hazard and adverse selection. Contrary to Besley (1989), the authors also consider the case

where the government can observe neither individuals´ risk type nor productivity while PI companies can observe their productivity. Due to moral hazard, PI imposes a negative externality on SI—but the converse is also true. In its choice of not only SI coverage but also the income tax schedule, the government takes the individuals´ decisions w.r.t. labor supply and healthcare expenditure as well as the rationing constraint imposed on the low-risk types in the separating contract of PI into account. The optimal degree of SI reflects its overall effect on expected utility, extending the work of Dahlby (1981).

Olovsson (2010) focuses on the pay-as-you-go (PYG) feature of old-age provision through social security, in important component of SI. The author decomposes its risk-sharing contribution to welfare into a direct and a general equilibrium effect (which can be traced to a wage increase due to the higher capital stock in the economy). In his overlapping generations (OLG) model, optimal SI needs to satisfy both intertemporal optimality and distributional optimality conditions across the three generations distinguished. These conditions are not satisfied in a laissez-faire economy calibrated to US parameters. In spite of a very high coefficient of relative risk aversion amounting to 12.75 derived from an Epstein and Zin (1989) utility function (which allows to separate risk aversion from time preference), the author finds that safe old-age benefits reduce aggregate welfare by an estimated 10% of per-period consumption in general equilibrium. The reason is that this safety requires high contributions to SI precisely in the face of negative productivity shocks that expose the young to high-income uncertainty while the old have the opportunity to build up their hedging portfolio.

A contribution by Zhao (2017) casts further doubt on the efficiency-enhancing effect of SI by generalizing the theoretical analysis to include its effects not only on labor supply but also on private saving (end hence bequests) as well as on PI (crowding-out again) in a dynamic setting. Focusing on the stationary equilibrium and calibrating the model to reflect parameters applicable to the United States, the author is able to estimate the individual welfare consequences of social insurance (including Medicaid after the 1996 reform, which created a consumption floor). Due to all these side effects of SI, he estimates a welfare gain caused by its elimination amounting to 2.1%, measured in terms of a consumption equivalent variation CEV (0.3% with only Medicaid eliminated, see the author's Table 7). A more refined analysis by Bagchi and Jung (2022) also comes to the conclusion that SI is not welfare-enhancing. The authors develop an OLG model with transition probabilities for labor income and health status (which may be subject to idiosyncratic shocks). They insert the parameters of the components of SI in great detail in their calibration using US data, also distinguishing the retired from the active population. According to Fig. 5, their simulation tracks data provided by the Medical Expenditure Panel Survey MEPS concerning medical expenditure and the share of both PI- and SI-insured by age quite well. Rather than showing the effects of a total elimination of SI, the simulations revolve around a cut by 50% in contribution rates to SI both in the presence and absence of stochastic health shocks. In their presence, individual welfare in terms of CEV increases by 4.73% (even by 5.13% in their absence; authors' Table 4). This gain is highest among the college-educated (6.84% and 6.94%, respectively) and lowest among those without completed high school (3.74% and 3.27%), reflecting the redistributive effect of SI.

Conclusion 3: While early contributions suggested that SI is (almost) certainly efficiency-enhancing, more recent ones are inconclusive or even arrive at a negative verdict.

However, these contributions cannot explain two phenomena: First, the political debate invariably revolves around the inability of high risks to obtain PI rather than the rationing of coverage imposed on the low ones by separating contracts. Second, in industrial countries SI strongly expanded from a few percent of GDP prior to World War II to around 15% by 1980 and some 25% (see Table 1 again).

The first shortcoming can be remedied by assuming risk-specific transaction costs for PI that are reduced by SI (the argument is a modification of Newhouse (1996); see also Zweifel et al. (2021), Chap. 10). In Fig. 1, let these transaction costs give rise to a proportional loading λ so large as

to cause the high risks to go without any PI coverage at all while the low risks continue to pay the actuarially fair premium ($\lambda = 0$). Graphically, the dashed indifference curve $\overline{EU}^{H}$ [0] through the origin indicates a higher expected utility than any point on the dashed insurance line $P^{H}(\lambda)$. In addition, let this loading reflect the price of risk bearing by PI, which is substantial in the case of high risks because, for example, their losses have not only high expected value but also high variance and (positive) skewness. Both are reduced when SI imposes coverage up to point S. For simplicity, assume the loading to become zero, making the insurance line P'^{H} originating from S the relevant one, as in Sect. 2.3. The high risk is now predicted to move from zero coverage to point H^{**} of Fig. 1, entailing an even greater welfare gain than in Dahlby (1981). In this sense, SI can be said to solve the problem of high risks not obtaining coverage. The other shortcoming is addressed in Sects. 3 and 4, respectively.

3 The Optimal Amount of SI

As noted toward the end of Sect. 2.3, the efficiency view of SI leaves the optimal amount of coverage to be provided by SI relative to PI undetermined, and with it, the optimal division of labor between SI and PI. There have been two strands of literature addressing this issue. One is to introduce a benevolent government as the agent who optimizes SI on behalf of a representative citizen. The other recognizes that the implementation of SI ultimately requires a majority support by voters.

3.1 A Benevolent Government Optimizing the Amount of SI

Health insurance constitutes a leading case where the government determines the relative importance of PI and SI, in particular by permitting citizens to purchase supplementary coverage offered by PI. This is the focus of Petretto (1999), whose model comprises three stages. First, the government introduces SI with a degree of coverage α ($0 < \alpha < 1$). Next, consumers select their preferred PI coverage, which is partial only in response to the loading contained in the premium (in Fig. 1, the high risks even choose zero coverage because of a very high loading). This leaves them with some positive net expense in the loss state, which depends on the consumer's decision with regard to healthcare expenditure and labor supply (and hence available income). While the model is couched in health insurance terms, its multiple-layer structure also applies to, for example, unemployment, where SI sets a rate of income replacement. Employers provide PI through a severance package (the premium being a wage rate below the marginal product of labor). In the provision for old age, SI creates a first layer in the guise of a public pension. A second layer comes from PI, which is either purchased individually in the guise of a whole life insurance contract or again through the employer and financed by a wage deduction.

The amount of loss falling on PI is not exogenous anymore (contrary to the exposition in Sect. 2.1) but is now subject to an *ex post* moral hazard effect. In the case of a health loss, this is the extra healthcare expenditure that would not be incurred without insurance coverage. In the case of unemployment, the worker may exert less search effort to find a new job. Similarly, retirees may seek to maximize their remaining life expectancy in response to a pension that is topped up by a PI contract.

Optimization proceeds in three stages. First, the government sets an optimal uniform degree of coverage α by maximizing the sum of citizens' utilities. It is subject to a budget constraint that involves a uniform rate of taxation t on labor income averaged over the loss and the no-loss state. In the second stage, individuals select their PI coverage $1 - k_i$, with k_i denoting the individual

rate of copayment. To simplify the analysis (and contrary to Sect. 2), the loss probabilities are assumed to be known; therefore, SI cannot be credited with relaxing the rationing constraint imposed on low risks by a separating contract. In the third stage, individuals decide about their amount of healthcare expenditure. Solving the model backward, Petretto (1999) obtains for the optimal degree of PI coverage,

$$\frac{1-k_i^*}{k_i^*} = \frac{u'[Loss]_i - \overline{u}_i'(1+\lambda)}{\overline{u}_i(1+\lambda)} \cdot \frac{1}{e_i}, \tag{1}$$

with $e_i > 0$ an elasticity (in absolute value) relating the loss borne by PI to the net cost borne by the insured. This result is quite intuitive. The optimal degree of PI coverage [the right-hand side of Eq. (1)] is higher,

- the higher the degree of risk aversion of individual i. The marginal utility of income in the loss state, $u'[\text{Loss}]_i$ is higher than the weighted average $\overline{u}_i'$ over the loss and the no-loss states, reflecting the concavity of the Von Neumann-Morgenstern risk utility function;
- the lower the proportional loading λ and hence the true price of insurance;
- the smaller the *ex post* moral hazard effect symbolized by the elasticity e_i. In a health context, e_i indicates how much the insured's demand for increases in response to a decrease in the net price of medical care. In an unemployment context, e_i shows how much search effort decreases in response to the lowered opportunity cost of remaining unemployed. Note that this opportunity cost also depends on the amount of coverage α provided by the first layer.

Finally, the optimal value of α is implicitly given by the condition,

$$\begin{aligned}\sum_i^n \text{Cov}\left(u_i'[\text{Loss}]_i, \text{NetLoss}_i\right) + n\left[\overline{u}'[\text{Loss}] \cdot \text{Cov}\left(\overline{u}_i'/\overline{u}', \overline{\text{NetLoss}_i}/\overline{\text{NetLoss}}\right) - 1\right] \cdot \overline{\text{NetLoss}} \\ = \sum_i^n u_i'[\text{Loss}]\left[(1-\alpha^*)\left(1-k_i^*\right) + \alpha^*\right] \frac{\partial \overline{\text{Netloss}_i}}{\partial \alpha}\end{aligned} \tag{2}$$

On the left, the marginal benefit associated with an increase of α has two components.

- The first term symbolizes a social risk-sharing gain. It is high if over the n individuals considered, there is a marked covariance between their marginal utility of wealth in the loss state and the amount of net loss they have to bear because PI coverage is only partial. Note that PI coverage demanded typically decreases when SI coverage expands (this "crowding-out" effect is taken into account in the contribution by Chetty and Saez (2009), to be discussed in Sect. 3.2 below). Generally, the amount of covariance is high if the marginal utility of income is high in the loss state due to strong risk aversion.
- The second component of marginal benefit is the social redistribution gain. Its main term is a (positive) covariance between the individual's relative average marginal utility and his or her relative net loss. The two quantities are relative because they relate the individual to a mean marginal utility in the population $\left(\overline{u}'\right)$ and the mean loss striking the population $\left(\overline{\text{Netloss}}\right)$, respectively. If this covariance is high, the insured has a strong interest in redistribution through SI. This is the case if an individual who suffers from an above-average net loss simultaneously is characterized by an above-average marginal utility of income (which reflects below-average income or wealth). The marginal utility of income prevailing in the population $\overline{u}'$ [Loss] translates this benefit into utility terms.

The sum of the two marginal benefits has to equal marginal cost at the optimum. Marginal cost consists of three multiplicative components.

- The last factor is the basic trigger. It shows how strongly the expected net loss associated with individual i reacts to an expansion of SI. This is the *ex post* moral hazard effect originating with SI (the probability of loss is observable, precluding *ex ante* moral hazard).
- The second factor shows that the degree of SI and PI coverage both act as amplifiers. The main impulse comes from α^* itself. In addition, however, PI covers the remainder $(1 - \alpha^*)$ to the tune of $\left(1 - k_i^*\right)$.
- Finally, the first factor transforms these effects into a utility value using the marginal utility in the loss state because *ex post* moral hazard occurs in the loss state only.

Equations (1) and (2) combined define the optimal division of labor between PI and SI.

3.2 Taking Into Account Moral Hazard, Adverse Selection, and Crowding-Out

In view of the results presented in Sect. 2, a major challenge for determining the optimal relative importance of PI and SI is to take into account informational asymmetry as well as the fact that an expansion of SI does not increase total insurance coverage in step because PI is "crowded out" to some degree. This challenge has been taken up by Chetty and Saez (2009), who use unemployment insurance (UI) as their leading example. While the division of labor between PI and SI is much more debated in the context of health insurance, there is a PI involvement in UI to the extent that employers provide severance pay. Both components of UI are financed by a payroll tax, one explicit (τ) levied by the government, the other, implicit (t_k) levied by the employer, who can observe the wage income W_k of type k. As in the preceding section, there is a benevolent government that seeks to maximize citizens' expected utilities. Through setting τ, it also determines the benefit rate α of the preceding section if the SI scheme is to maintain its financial equilibrium. Chetty and Saez (2009) show that the optimum explicit contribution rate is given by

$$\frac{\tau}{1-\tau} = -\frac{1}{e_{\overline{W},1-\tau}} \frac{\overline{\mathrm{Cov}}\left(\overline{u}_k{}', \overline{W}_k\right)}{\overline{u}' \cdot \overline{W}} + \frac{1}{e_{\overline{W},1-\tau}} \cdot \sum_k s_k \left(1 - r_k\right) e_k \left(t_k^* - t_k\right) \cdot \frac{\overline{u}_k' \cdot \overline{W}_k}{\overline{u}' \cdot \overline{W}}. \tag{3}$$

Note that the left-hand side of Eq. (3) is increasing in τ. It is determined as follows.

- *Moral hazard*: This is reflected by two parameters. The first is the elasticity $e_{\overline{W},1-\tau}$ which indicates how strongly labor income $\overline{W}$ averaged over the loss, and the no-loss state reacts to the amount that the worker retains per Dollar earned and hence to the payroll tax τ levied by SI. The stronger this response, the more the benefits of SI in terms of income smoothing (see below) have to be adjusted downwards. It corresponds to the elasticity e_i in Eq. (1) as in Petretto (1999), who, however, does not yet distinguish between the moral hazard effects induced by the SI and PI layers. This is achieved here by the second parameter e_k, an elasticity indicating the response of $\overline{W}$ to the overall retention per Dollar earned, given by $(1 - \tau)(1 - t_k)$. The stronger this group-specific effect of PI (relative to SI), the more the scaling down of SI due to $e_{\overline{W},1-\tau}$ has to be corrected in favor of SI again.
- *Adverse selection*: This is represented by $\left(t_k^* - t_k\right)$, the difference between the PI contribution rate (levied by the employer) in the absence of the self-selection constraint $\left(t_k^*\right)$and in its presence

(t_k). This difference is positive due to the rationing effect of adverse selection (see Fig. 1 again), indicating a benefit favoring SI.

- *Crowding-out of PI*: This is represented by the group-specific parameter $r_k > 0$ which is the elasticity of the PI retention rate $(1 - t_k)$ with respect to the SI retention rate $(1 - \tau)$. Chetty and Saez (2009) call this a fiscal externality, which would be absent from informal insurance provided, for example, among members of the same household. Since both components of UI are subject to budget balance, this elasticity indirectly indicates how much of employer-provided coverage is crowded out by SI (however, for simplicity the model does not contain a module representing optimizing behavior of either employers or employees in this respect). The greater r_k, the smaller should be the optimal extent of SI, ceteris paribus. As will be shown in Sect. 4.2, crowding-out effects are indeed substantial, reducing the social benefit to be expected from a (continued) expansion of SI.

The remaining elements of Eq. (3) can be interpreted as follows. The $\overline{\text{Cov}}$ term represents the consumption smoothing benefit of SI and PI combined. For each group k of the population, their marginal utility of income averaged over the loss and the no-loss state $\overline{u}'_k$ is compared with that of that of the population as a whole $\overline{u}'$. This ratio is high (due to the concavity of the risk utility function) when relative income $\left(\overline{W}_k/\overline{W}\right)$ is low, resulting in a negative covariance. The group-specific covariance values have to be averaged to obtain $\overline{\text{Cov}}\,(\cdot)$; together with the negative sign, this amounts to a benefit similar to the social risk-sharing gain in Eq. (2). It has to be qualified in terms of the moral hazard effect $e_{\overline{W},1-\tau}$ discussed above. The summation in the second term is over the population shares s_k and bears close similarity to a covariance between (relative) marginal utility and income levels giving rise to the social redistribution gain in Eq. (2).[9] Here, the moral hazard effect of PI relative to SI given by $e_k/e_{\overline{W},1-\tau}$, the rationing due to adverse selection $\left(t_k^* - t_k\right)$, and the leakage due to the crowding-out effect $(1 - r_k)$ all enter as qualifying factors.

Clearly, the averaged marginal utility of income depends on their values in the loss and the no-loss state, which in turn are determined by the curvature of the risk utility function. The relationship with the coefficient of (relative) risk aversion was established by Chetty (2006), who also showed that in spite of later modifications, an influential finding by Bailey (1978) continues to be valid. It states that the optimal amount of UI benefits is determined by (1) the mitigation of the drop in consumption achieved by UI, (2) the coefficient of relative risk aversion, and (3) the elasticity of unemployment duration with respect to UI benefits. The analysis by Chetty and Saez (2009) shows that these parameters are indeed crucial. In Eq. (3), the mitigation of the drop in consumption is reflected by $\overline{W}_k/\overline{W}$, the role of (differences in) risk aversion indicated by $\overline{u}'_k/\overline{u}'$, while the elasticity of duration has become the elasticity of earnings $e_{\overline{W},1-\tau}$.

Finally, the authors explicitly relate the SI contribution rate τ to a PI contribution rate $\hat{t}$, which is weighted by population shares, relative incomes, and relative marginal utilities,

$$\hat{t} = -\frac{\tau}{1-\tau} + \frac{1-\hat{r}}{e_{\overline{W},1-\tau}} \cdot \frac{-\overline{\text{Cov}}\left(u', W\right)}{\overline{W}}. \tag{4}$$

Clearly, this defines an optimal crowding-out relation in that a higher (optimal) value of τ (which increases with $\tau/(1-\tau)$) implies a lower value $\hat{t}$, ceteris paribus. However, the PI component is also to be credited for its contribution to consumption smoothing (indicated by the average covariance between marginal utility u' and income W, which is usually negative), qualified by the crowding-out

[9] Although the last factor of Eq. (3) is not expressed as deviations from expected values, it still says that the benefit of SI is large if high (relative) marginal utilities of income coincide with low (relative) levels of wealth.

factor, with $\hat{r}$ symbolizing an average value of r_k weighted in the same way as t_k. Equations (3) and (4) together indeed define an optimal division of labor between PI and SI.

Conclusion 4: Given a benevolent government, the optimal degree of SI coverage depends positively on social risk-sharing and a redistribution gain but negatively on the moral hazard caused by it. It needs to be adjusted upward when the rationing effect of adverse selection in PI is substantial but downward when SI crowds out PI to a great extent.

4 The Crowding-Out Phenomenon

4.1 The Public Choice View of Social Insurance

Contrary to the basic assumption adopted in Sect. 3, the government in a democracy cannot act as a benevolent dictator but must act on behalf of citizens who know their risk type and position in the income and wealth distribution. Referring back to Fig. 1, it is evident that the high risks would always gain from a continued expansion of SI to the detriment of PI. At first sight, such a shift appears even more likely if individuals not only differ in terms of risk but also in terms of income. As voters, the poor would have an incentive to see SI expanded because it also redistributes income from the rich to the poor. However, it is the preferences of the median voter who are decisive in determining whether or not a proposal finds a majority in the political process. Indeed, the middle class may favor a mixed system because it can benefit from the better deal offered by PI; on the other, it would in the limit bear an excessive burden due to the income redistribution by SI (Gouveia 1997).

In the following, the model by de Donder and Hindricks (2003) is expounded because it is suited to explain the growing importance SI relative to PI if voters are offered a choice. In this way, it provides a strong theoretical basis for the crowding-out findings to be reported in Sect. 4.2 below. Indeed, the model predicts support for SI against PI even by voters who are both slightly richer and better risks than average, regardless of the distribution of income and risk in the country. This sweeping prediction goes beyond Casamatta et al. (2000), who merely derived conditions that shift the division of labor between PI and SI in favor of SI. It comes at a price, however. Rather than the conventional expected utility (EU) framework, Yaari's (1987) dual formulation is adopted by the authors. In this formulation, risk aversion is expressed not by the concavity of the risk-utility function but by the concavity of a weighting function $\varphi(p)$ defined over probabilities p. This permits to express the utility V^{P} derived from a PI contract directly in terms of predetermined wealth or income W,

$$V^{\mathrm{P}} = \varphi(p)\,(W - P - k \cdot 1) + (1 - \varphi(p)\,(W - P)) = W - P - \varphi(p)\,(1 - k)\,, \tag{5}$$

with $\varphi(p) = (1 + A)p$ (A to be explained below).

Therefore, utility amounts to a weighted average of wealth or income levels in the loss and the no-loss state, respectively. In the loss state, income is net of the premium P and the rate of coinsurance k on a loss that is normalized to 1. This normalization makes the loss a fixed quantity, thus precluding any moral hazard effect, in contradistinction with Sect. 2.3. In the no-loss state, the premium still needs to be paid. The two income levels are weighted by a function $\varphi(\cdot)$ whose parameter $A > 0$ reflects risk aversion; for simplicity, A is assumed independent of W. Note that eq. (5) implies choice of full coverage ($k = 0$) even if the premium exceeds its fair value p (which is commonly observed behavior, however). This differs from the prediction derived from EU theory (see Fig. 1). It is the consequence of the fact that risk aversion has a first-order influence here but merely a second-order influence in EU theory (stemming from the concavity of the risk-utility function).

SI can offer full coverage [$(1-k)=1$] since it suffers neither from moral hazard nor risk selection. On the other hand, contributions $\overline{P}$ are scaled according to the individual's wealth compared to the average $\overline{W}$, reflecting the redistribution characteristic of SI. Therefore, the utility value of SI is given by Eq. (6),

$$V^s = W - \left(W/\overline{W}\right)\overline{P} \tag{6}$$

Before comparing Eqs. (5) and (6), de Donder and Hindricks (2003) proceed to impose the self-selection constraint that renders separating PI contracts viable, in analogy to Fig. 1. They apply the criterion introduced by Mailath (1987), $dV^{\mathrm{P}}/d\hat{p} = 0$ for $\hat{p} \to p$, stating that an individual cannot gain by announcing a risk type $\hat{p}$arbitrarily close to the true one. This permits them to derive an optimal coverage function, relating k^* to p and A. After substitution of k in Eq. (5), the can implicitly define an income level W^o where the two utilities are equal,

$$W^o - (1+A)\,p + Ap\left(p/\overline{\overline{p}}\right)^{1/A} = W^o - \left(W^o/\overline{W}\right)\overline{P} \tag{7}$$

The new symbol $\overline{\overline{p}}$ is the maximum loss probability in the population; it is part of the optimal benefit function. Therefore, the last term on the left-hand side of Eq. (7) expresses the benefit in financial terms that an individual characterized by a loss probability p relative to $\overline{\overline{p}}$ can derive from the separating PI contract. Taking the log of this term and differentiating it w.r.t. A results in a positive value, indicating that this benefit increases with risk aversion A, as one would expect. However, the weighting function $\varphi(\cdot)$ causes the premium paid (the term $-(1+A)p$ on the left-hand side) to increase as well. Solving condition (7) for the income level W^o indicating indifference between PI and SI yields.

$$\frac{W^o}{\overline{W}} = \left[1 + A - A\left(\frac{p}{\overline{\overline{p}}}\right)^{1/A}\right]\frac{p}{\overline{p}} \quad = \quad [1+\varphi(A,p)]\,\frac{p}{\overline{p}}, \text{ with } \varphi(A,p) = A\left[1-\left(p/\overline{\overline{p}}\right)^{1/A}\right]. \tag{8}$$

This condition can be interpreted as follows. Consider an individual with the mean loss probability $\overline{p}$. Since $\varphi(\cdot) > 0$, $W^o/\overline{W} > 1$, implying that such an individual votes in favor of SI although his or her wealth is above-average. Conversely, assume that the pivotal voter has below-average wealth. In that event, he or she continues to support SI rather than PI if his or her loss probability p corresponds to the average $\overline{p}$; indeed, it could even be somewhat below-average. Ultimately, the reason for this dominance of SI lies with the term $\left(p/\overline{\overline{p}}\right)$, which reflects the necessity of sustaining separating contracts. This detracts from the benefit of PI, whereas SI does not need to satisfy this constraint.

The final step is forming the derivative of V^s with respect to $1-k^s$, which the authors show to be positive. In addition, its value does not depend on $1-k^s$, which implies the corner solution with $1-k^{s*}=1$, that is, full coverage by SI. Moreover, individuals with the crucial income W^o turn out to be indifferent to an expansion of SI, while those with less income (the majority in a country with a skewed income distribution) prefer to have full SI coverage. For this reason, a mixed system cannot be sustained politically.

In this way, the analysis provided by de Donder and Hindricks (2003) comes close to explaining the expansion of SI since its inception. It could be interpreted as the transition to a political equilibrium in which a loss of a given type is covered by SI exclusively. The puzzle then becomes why this expansion seems to have slowed down recently (see Table 1). A possible explanation is that several exogenous

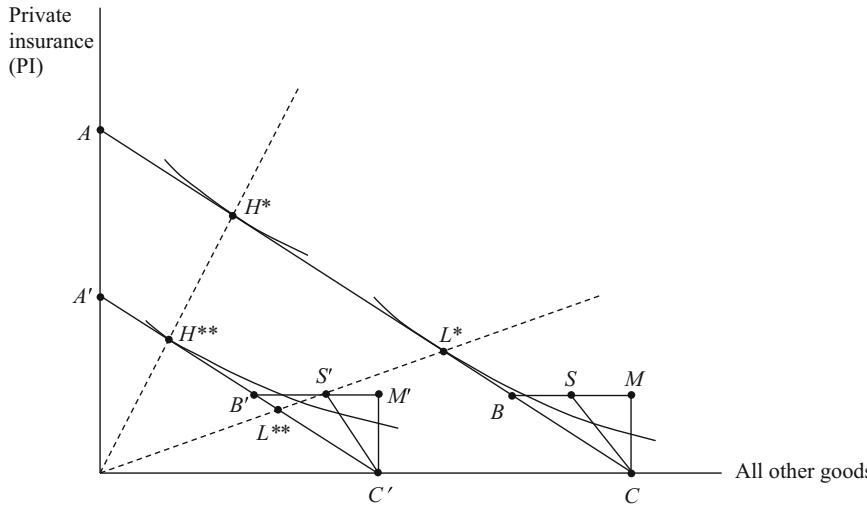


Fig. 2 The effect of public insurance on the demand for private insurance

changes (e.g., increased migration, to be discussed in Sect. 6), increase the cost of enforcement for SI, causing a loading in the contribution to SI $\overline{P}$ of Eq. (6) that may push the threshold value of W^{o} sufficiently below $\overline{W}$ to establish a stalemate between supporters and opponents of SI.

Conclusion 5: The historical expansion of SI in democracies can be explained the fact that even voters with above-average incomes and slightly below-average probability of loss prefer it to PI.

4.2 Evidence on the Crowding-Out Phenomenon

Cutler and Gruber (1996) provide the theoretical underpinnings to the crowding-out of PI by SI in a particularly simple way.[10] While they apply foremost to the United States, most industrial countries have a similar institutional setting. The straight line *ABC* in Fig. 2 depicts the budget constraint of an individual whose income is low enough to qualify for an SI program. Since "All other goods" corresponds closely to income, the slope of the budget constraint (inversely) reflects the amount of income that has to be sacrificed for additional insurance coverage, that is, the premium per unit coverage provided by PI. If highly risk-averse, the individual is characterized by an indifference curve making H^* the optimum; for a less risk-averse type, the optimum is L^*. Now, let the government offer an SI program with a take-it-or-leave-it benefit falling short of L^* at a subsidized contribution rate, causing the budget constraint to become *ABSC*. If the program does not require a contribution at all, the budget constraint becomes *ABMC*. The H-type individual is unaffected; however, the L-type is predicted to typically opt for the SI program at point *S* (or *M*, respectively). These corner points dominate because both are chosen by many types characterized by indifference curves with differing slopes. In either case, the amount of PI demanded decreases.

Now consider a still poorer individual with budget constraint $A'B'C'$, with the two preference types unchanged (homothetic preferences are assumed). In the absence of the SI program, the H-type opts for H^{**}. When the SI program becomes available, the L-type shifts to it for sure (to point S' or M',

[10] For a theoretical analysis predicting a crowding out of private long-term care insurance, see Zweifel and Courbage (2015).

respectively). However, even the H-type is now predicted to opt for the SI alternative, with the result that SI crowds out PI regardless of type. The reason is that for a low-income individual, the SI option has much higher relative importance than for a higher-income individual, a difference which may swamp the influence of a preference structure that in principle favors the PI alternative (see H^{**} again). This effect is reminiscent of Hubbard et al. (1995), who also emphasized the impact of SI on low-income households in the context of the crowding-out of private savings.

Cutler and Gruber (1996) take the expansion of US Medicaid (in particular benefits for mothers with children) as the exogenous impulse to study the impact of SI on the demand for PI. They ignore considerations of adverse selection in PI that might contribute to its crowding-out by the SI alternative, noting that much of US health insurance was contracted through employers at the time, who by imposing open enrollment on health insurers prevented them from cream-skimming. The authors estimate the probability of eligibility for children and women of child-bearing age. For children, eligibility has a positive effect on the take-up of Medicaid but a negative one on the likelihood of having private health insurance amounting to a crowding-out of 31% per percentage point of Medicaid expansion. For women, both the effect on the take-up of Medicaid and (negative) on PI are insignificant, but if taken at face value, they even suggest a crowding-out rate in excess of 100%. For the parents, the authors also check whether the expansion of Medicaid caused employers to curtail their offer of private health insurance coverage. However, their findings suggest that the decision to drop PI was made by the employees—possibly encouraged by their employers. They arrive at an estimated 49% rate of crowding-out, that is, out of two persons additionally covered by Medicaid, one canceled his or her private coverage.

In a follow-up study, Gruber and Simon (2008) review other work on the crowding-out effect of Medicaid. Notably, Dubay and Kenney (1997) found much lower rates of crowding-out by comparing changes in private enrollment over time in populations eligible and not eligible for the Medicaid program. Moreover, in 1998, the State Children's Health Insurance Program (SCHIP) was initiated. Yet beneficiaries had to prove that they had been without private health insurance (usually due to a job loss) for between 6 and 12 months, depending on the state. Indeed, Lo Sasso and Buchmueller (2004) had concluded that these waiting periods served to reduce crowding-out. In their re-estimation using a different data set, Gruber and Simon (2008) are unable to replicate this finding, estimating overall crowding-out rates for SCHIP to be similar to those for Medicaid. As to Medicaid, they obtain rates ranging from 24% to 81%, bracketing their original estimate of 49%. The large variation is caused by differing assumptions regarding the changes in composition of the subgroup that has both private and public coverage.

Another instance of Medicaid displacing PI is coverage for long-term care (LTC).[11] Brown and Finkelstein (2008) emphasize that contrary to the Medicare program (for the elderly), Medicaid (for the poor) does not permit PI to top up benefits; rather, PI (which provides partial coverage) must be used up until wealth falls to the low Medicaid threshold. Using a dynamic optimization algorithm, the authors derive willingness-to-pay (WTP) values for LTC coverage as currently offered in the presence of Medicaid, to find a value of -$20,700 (women) and -$18,200 (men) at the 30th percentile of the US wealth distribution. The WTP values turn positive no sooner than at the 60th percentile. The authors then assume fair premiums for PI combined with full LTC coverage—only to still find negative WTP values up to at least the 50th percentile. The cause is the implicit tax on PI which reaches values close to 100% for wealth levels below the 30th percentile. As a consequence, full loadings on PI approach 100% in this domain, making coverage very expensive indeed. While the authors do not explicitly calculate crowding-out rates, their simulations clearly suggest substantial crowding-out effects of Medicaid in spite of the fact that its coverage is partial only.

[11] For a theoretical analysis predicting a crowding out of private LTC insurance, see Zweifel and Courbage (2015).

Outside the United States, evidence on private LTC insurance being crowded out is rare. In 2012 Xu and Zweifel (2014) conducted a survey among residents of Shanghai (China) in which the 500 respondents aged 30–60 were asked to first adopt the viewpoint of a child and then, that of a parent. In the first scenario, no fewer than 99% indicated that they would scale back the amount of informal care if the parent were to purchase (more) private LTC coverage. Conversely, 28% stated that acting as the parent they would purchase less (no) LTC insurance if the child were to provide more informal care. While the authors concluded that bilateral intergenerational moral hazard was present, confirming Courbage and Zweifel (2011), these results can also be interpreted as suggesting that informal care crowds out LTC insurance. Even more telling is the finding that 61% of the respondents thought that the burden of LTC should be assumed by social insurance in the future. Using expectational data covering 15,172 respondents living in the EU-15[12], Costa-Font and Courbage (2015) also find evidence of a crowding-out of private LTC insurance, this time mainly by support provided by the family. Respondents were asked whether they expect SI, their own wealth, PI, or the family to finance the cost of their LTC. The authors´ regression analysis suggests that support by the family (but less clearly SI) crowds out private LTC insurance.

Conclusion 6: For the case of US Medicaid in particular, there is substantial evidence suggesting a crowding-out effect on PI. The same applies to private LTC insurance according to preliminary evidence from China and European Union member countries.

5 Could the Division of Labor Between Private and Social Insurance Be Improved?

For all their merits, the contributions discussed up to this point have one weakness in that they all consider one risk impinging on the individual at a time. However, PI and SI both consist of a multitude of lines in response to the fact that there are a great many impulses affecting individual assets which comprise wealth, health, and skills (Zweifel et al. 2021, Chap. 1). For instance, an illness episode affects not only health but indirectly wealth (through reduced labor income) and skills (through depreciation of human capital). The same is true of an accident, of disability, and unemployment, resulting in a positive correlation between these three assets. Braveman et al. (2018) present a wide array of evidence suggesting a significant link between health and wealth, while Cutler and Lleras-Muney (2010) document the link between health and education. As Eeckhoudt and Schlesinger (2006) have shown, risk-averse individuals seek to eschew an accumulation of risks so would benefit from benefits from PI and SI coverage jointly neutralizing these correlations.

Consider two out of the three assets, X and Y, with X denoting health transformed into monetary units. Since both changes in the probability of survival and in the quality of life can be valued using willingness-to-pay and time trade-off techniques, the asset "health" can be made commensurable to the other asset, Y (Zweifel et al. (2009), Sect. 2.4). Let this asset Y be skills; their value can be derived from wage differentials on the labor market. In a (μ, σ)-framework, expected returns of PI are close to a (quasi-) risk-free rate of return (mostly on long-term government bonds) because of regulators' solvency concerns. As is well-known (Samuelson 1958), SI also has a rate of return, viz., the growth of the labor income from which benefits for the retired are paid. Other lines of SI are financed through a payroll tax as well so their benefits develop in line with labor income as well (assuming budget balance for SI).

[12] The EU-15 comprises the member states of the European Union before its enlargement in 2004, that is, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxemburg, Netherlands, Portugal, Spain, Sweden, and the United Kingdom.

During the period 1960–2007, the two rates of return have not differed markedly according to internationally comparable OECD estimates (see Schoder et al. (2013) for details). Thus, the expected rates of return on private and social insurance may be set equal as an approximation, permitting the analysis to be limited to the variance component of (μ, σ). Note that PI premiums and contributions to SI are treated as non-stochastic quantities since they are predictable. This leaves the benefits of PI and SI as the quantities of interest.

Specifically, denote by $X^a + x$ the final value of health capital, with X^a the asset value after expected payment by insurance. Under ideal circumstances, X^a would be a fixed (optimal) quantity. However, medical care may fail to restore the health stock to its optimal value; more importantly in the present context, any copayment gives rise to an unexpected variation (denoted by x).

Likewise, the total value of skills can be split into Y^a (after expected insurance benefits) and y (unexpected variation in benefits). Unexpected benefit variation even occurs in SI. For example, a beneficiary may learn that a degree of negligence was found on his or her part in a workplace accident, causing workmen's compensation to be curtailed. Conversely, benefits may be higher than expected because, for example, the presence of a child triggers a supplementary benefit.

Under these circumstances, total asset variance is given by

$$\underset{(+)}{\mathrm{Var}}\left(X^a + x + Y^a + y\right) = \underset{(+)}{\mathrm{Var}}\left(X^a\right) + \underset{(+)}{\mathrm{Var}}(x) + \underset{(+)}{\mathrm{Var}}\left(Y^a\right) + \underset{(+)}{\mathrm{Var}}(y) \tag{9}$$

$$+2\underset{(0)}{\mathrm{Cov}}\left(X^a, x\right) + 2\underset{(+)}{\mathrm{Cov}}\left(X^a, Y^a\right) + 2\underset{(0)}{\mathrm{Cov}}\left(X^a, y\right) + 2\underset{(0)}{\mathrm{Cov}}\left(Y^a, y\right) + 2\underset{(0)}{\mathrm{Cov}}\left(Y^a, x\right) + 2\underset{(?)}{\mathrm{Cov}}\left(x, y\right).$$

Here, $\mathrm{Var}(X^a)$ and $\mathrm{Var}(Y^a)$ remain positive to the extent that insurance stipulates a degree of copayment in response to moral hazard or as a property of separating contracts (see Sects. 2.1 and 2.3). Under the usual assumption that expected and unexpected components of insurance benefits are uncorrelated, $\mathrm{Cov}\,(X^a, x)$ and $\mathrm{Cov}(Y^a, y)$ can be set to zero. Moreover, expected benefits under one title (X, say) presumably are not related to unexpected benefits under another title (Y). Thus, $\mathrm{Cov}(X^a, y) = \mathrm{Cov}(Y^a, x) = 0$. Also, while expected benefits are negatively correlated with variations in the covered asset, some of the post-insurance variation usually remains. Thus, the positive covariance between the two assets will still show after payment of expected benefits, making $\mathrm{Cov}(X^a, Y^a) > 0$. The one term related to insurance that can be used for minimizing total asset variance therefore is $\mathrm{Cov}(x, y)$. The more strongly negative the covariance between unexpected deviations form expected benefits, the smaller total asset variance. Thus, the performance of the insurance system as a whole from the consumer's point of view insurance can be gauged by the direction and amount of correlation in its unexpected benefit components.

By fitting the time series of insurance benefits to quadratic time trends to (roughly) account for both their expansion and inflation in the economy, residuals can be calculated that reflect unexpected deviations of benefits from their expected value. Determine between two lines of insurance. For a given pair of standard errors $\{\sigma_x, \sigma_y\}$, the correlation coefficient $\rho_{x,y}$ indicates the size of $\mathrm{Cov}(x,y)$ (see Schoder et al. (2013) for details)).

In the case of the United States, four lines of personal insurance can be distinguished, reflecting data availability. There are two significantly positive correlations out of six, while one correlation is significantly negative. In an earlier study, using data for 1972–1992, one positive and one negative correlation (out of six) were found (see Zweifel (2000), where the admissibility of a microeconomic interpretation of the aggregate data is also discussed). Still, one might argue that risk diversification by PI could be better, calling for a shift in favor of SI. According to the philosophy of SI, a beneficiary should always be able to count on a minimum level of consumption (which corresponds to a minimum of total assets). Thus, more variance reduction might be expected from SI than from PI. Somewhat

surprisingly, however, the evidence does not point to a superior performance of SI. Unexpected variations in benefits correlate positively in no less than 15 out of 28 cases (especially between workers´ compensation and disability benefits), while not a single negative correlation of statistical significance is found. Again, the earlier study based on 1972–1992 data found six positive correlations (but also three negative ones) out of 21 possible cases. Therefore, SI in the United States seems to expose rather vulnerable individuals to excessive asset variance.

A final consideration is that risk diversification could be also achieved by PI filling unexpected gaps left by SI and vice versa. However, when the four PI lines distinguished are juxtaposed with the seven lines of SI (not shown here), this expectation fails to be confirmed in the case of the United States. Out of 28 correlation coefficients, none is significantly negative. On the whole, then, PI and SI fail to complement each other in a way that would contribute to a maximum reduction of consumers' total asset variance, a finding that however applies to Germany and Japan as well according to the earlier study.

Conclusion 7: Given the similarity of rates of return of PI and SI, their mission becomes to jointly minimize the variance of individual's assets, which calls for negative correlations between unexpected deviations of payments from their expected values. However, preliminary evidence derived from aggregate time series point to a preponderance of positive correlations within PI, SI, and between PI and SI.

6 Concluding Remarks and Outlook

This contribution takes as its starting point the well-known potential failures of private insurance (PI) markets, moral hazard, and adverse selection, which might justify government intervention, possibly in the guise of social insurance (SI). Since moral hazard effects are found to beset SI at least as much as PI, adverse selection constitutes crucial market failure (Conclusion 1). However, at least in the case of private group health insurance, the efficiency loss caused by adverse selection may be rather small, with little evidence suggesting of the occurrence of a "death spiral" (Conclusion 2). While some degree of mandatory SI has the potential to improve the welfare of both high and low risks, more recent work casts doubt on its efficiency-enhancing impact (Conclusion 3). Given a benevolent government, the optimal degree of SI coverage depends positively on a social risk-sharing and a redistribution gain but negatively on the moral hazard caused by it. It needs to be adjusted upward when the rationing effect of adverse selection in PI is substantial but downward when SI crowds out PI to a great extent (Conclusion 4). The pertinent contributions cannot explain the historical expansion of SI in democracies, however. Adopting a public choice view, one finds that even voters with above-average incomes and a slightly below-average probability of loss may prefer it to PI (Conclusion 5). This amounts to a crowding-out of PI by SI, for which there is substantial evidence from expansions of US Medicaid and preliminary evidence from China and European Union member countries concerning private LTC insurance (Conclusion 6). As to the issue of whether the division of labor between PI and SI has scope for improvement, the answer is yes because unexpected deviations of payments from their expected values are correlated positively rather than negatively across the lines of PI, SI, and between the two, exposing individuals to excessive asset variance (Conclusion 7).

Even with these insights in hand, it is difficult to answer the two crucial questions for the future: How will the division of labor between PI and SI evolve, and How should this division be changed, if at all? As to the first question, the expansion of SI seems to have come to a halt in several industrial countries (see Table 1). This may be the consequence of several exogenous developments, which challenge SI more than PI. One such change is the opening of economies not only to the international flows of goods but of labor and capital as well. Private insurers have the freedom to pursue an

investment policy that can benefit from the hedging provided by international capital markets, while institutions of SI are tied to their domestic capital market (provided they dispose of reserve capital at all). In view of their complexity, the theoretical models presented in this Chapter invariably neglect PI benefits that can be financed by investment income. When it comes to international movements of labor, SI schemes of rich countries cannot easily follow highly skilled emigrants whereas PI usually grants full portability of benefits. Conversely, lower-skilled immigrants may even be attracted by the generosity of SI benefits in rich countries while PI benefits reflect risk-based premiums. Emigration and immigration thus threaten to undermine the financial equilibrium of SI but not PI (Zweifel and Eugster 2008).

Another challenge is demographic change. It affects SI with its pay-as-you-go finance much more directly than PI, which frequently is capital-based. In addition, both theory and (rather spotty) empirical evidence suggest that the individual decisions causing demographic change in the aggregate (length of education, marriage, number of children, age at retirement, and even longevity) are in fact influenced by SI, and predominantly in ways exacerbating its financing problems (Zweifel and Eugster 2008). It is conceivable that these changes explain the reversal in the trend of expanding SI coverage noted above.

Turning to the normative issue of what the division of labor between PI and SI should be, the theory discussed in this Chapter certainly provides a measure of guidance. However, the underlying hypothesis is that governments implement and adjust SI reflecting reasonably informed self-interest of the voting public. One undisputed characteristic of the implementation is that it should be long-term since volatility in SI disturbs the life-cycle decisions of individuals. However, in a disquieting piece of research, van Dalen and Swank (1996) find clear empirical evidence suggesting that a "solid" government such as the Dutch boosted SI programs around (re)election times, presumably in an attempt to win pivotal votes. Therefore, even if one were to know whether (and in which lines of insurance) the division of labor between PI and SI ought to be changed, the issue of whether the proposed adjustment will be truly efficiency-enhancing remains. The balance between market failures that may beset PI and political failures that may beset SI is far from evident, and it may well change over time!

Acknowledgments The author wishes to thank David Rowell (University of Queensland, Brisbane) and XY anonymous referees for helpful suggestions and criticisms.

References

Akerlof GA (1970) The market for 'lemons': quality uncertainty and the market mechanism. Q J Econ 84(3):488–500
Arnott R, Stiglitz J (1991) The welfare economics of moral hazard. In: Loubergé H (ed) Risk, information and Insurance. Springer, Dordrecht, pp 91–121
Arrow KJ (1963) Uncertainty and the welfare economics of medical care. Am Ec Rev 53:941–973
Bagchi S, Jung J (2022) Health risk and the welfare effects of social security. Macroec Dyn 27(7):1767–1806
Bailey M (1978) Some aspects of optimum unemployment insurance. J Publ Econ 10:379–402
Beck K, Trottmann M, Zweifel P (2010) Risk adjustment in health insurance and its long-term effectiveness. J Health Econ. 29(4):489–498
Besley T (1989) Publicly provided disaster insurance for health and the control of moral hazard. J Public Econ 39(2):141–156
Blomqvist A, Johansson P-O (1997) Economic efficiency and mixed public/private insurance. J Public Econ 66(3):505–516
Boadway R, Leite-Monteiro M, Marchand M, Pestieau P (2006) Social insurance and redistribution with moral hazard and adverse selection. Scand J Econ 108(2):279–298
Braveman P, Arkin E, Orleans T et al (2018) What is health equity? Behav Sci & Pol 4(1):2–3

Brown JR, Finkelstein A (2008) The interaction of public and private insurance: Medicaid and the long-term care insurance market. Am Econ Rev 98(3):1083–1102
Cabral M, Geruso M, Mahoney N (2018) Do larger health insurance subsidies benefit patients or producers? Evidence from Medicare Advantage. Am Ec Rev 108(8):2048–2087
Casamatta G, Cremer H, Pestieau P (2000) Political sustainability and the design of social insurance. J Public Econ 75(3):341–365
Cawley J, Philipson T (1999) An empirical examination of information barriers to trade in insurance. Am Econ Rev 89(4):827–846
Chandler S (2015) Prices rising, choice declining for 2016. https://acadethspiral.org/tag/adverse-selction/. Accessed 22 Jan 2023
Chetty R (2006) A general formula for the optimal level of social insurance. J Public Econ 90:1879–1901
Chetty R, Saez E (2009) Optimal taxation and social insurance with endogenous private insurance. Am Econ J Econ Pol 1(2):31–52
Chiappori PA, Salanié B (2000) Testing for asymmetric information in insurance markets. J Political Econ 108(1):56–78
Cohen A, Einav L (2007) Estimating risk preferences from deductible choice. Am Econ Rev 97(3):745–788
Costa-Font J, Courbage C (2015) Crowding out of long-term care insurance: evidence from European expectations data. Health Econ 24(S1):1–27
Courbage C, Zweifel P (2011) Two-sided intergenerational moral rd, long-term care insurance, and nursing home use. J Risk & Unc 43:65–80
Cox C, Fehr R, Levitt L (2019) Individual insurance market performance in 2018. KFF Kaiser Family Foundation
Cutler DM, Lleras-Muney A, Vogl T (2011) Socioeconomic status and health: dimensions and mechanisms. In: Glied S, Smith PC (eds) Oxford handbook of health economics. Oxford Academic, Oxford, pp 124–163
Cutler DM, Gruber J (1996) The effect of Medicaid expansions on public insurance, private insurance and redistribution. Am Econ Rev 86(2):378–383
Cutler DM, Lleras-Muney A (2010) Understanding differences in health behavior by education. J Health Econ 29(1):1–28
Cutler DM, Reber SJ (1998) Paying for health insurance: the trade-off between competition and adverse selection. Q J Econ 113(2):433–466
Dahlby B (1981) Adverse selection and Pareto improvements through compulsory insurance. Public Choice 37:547–558
Danzon P (1992) Hidden overhead costs: is Canada really less expensive? Health Aff 11(1):21–43
de Donder P, Hindricks J (2003) The politics of redistributive social insurance. J Public Econ 87(12):2639–2660
Dionne G, St-Michel P (1991) Workers' compensation and moral hazard. Rev Econ St 83(2):236–244
Dubay LG, Kenney G (1997) Did Medicaid expansions for pregnant women crowd-out private insurance? Health Aff 16(1):185–193
Eeckhoudt L, Schlesinger H (2006) Putting risk at its proper place. Am Econ Rev 96(1):280–289
Einav L, Finkelstein A, Cullen MR (2010) Estimating welfare in insurance markets using variation in prices. Q J Econ 125(3):877–921
Einav L, Finkelstein, Ryan S et al (2013) Selection on moral hazard in health insurance. Am Ec Rev 103(1):178–219
Eliason M, Johansson P, Nilsson M (2018) Forward-looking moral hazard in social insurance: evidence from a natural experiment. Lab Econ 60:84–98
Epstein LG, Zin SE (1989) Substitution, risk aversion and the temporal behavior of consumption and asset returns: a theoretical framework. Econometrica 57(4):937–969
Ericson KM, Kircher P, Spinnewijn J et al (2012) Inferring risk perceptions and preferences using choice from insurance menus: theory and evidence. Econ J 131(634):713–744
Feldstein MS (1995) Would privatizing social security raise economic welfare? NBER Working Paper No. 5281, published in Feldstein M (ed) (1998) Privatizing social security. Chicago University Press, Chicago
Frech HE III, Smith MP (2015) Anatomy of a slow-motion health insurance death spiral. N Am Actuar J 19(1):60–72
Frech HE III, Zweifel P (2017) Market socialism and community rating in health insurance. Comp Econ Stud 59(3):405–427
Gouveia M (1997) Majority rule and the public provision of a private good. Public Choice 93:221–244
Gruber J, Simon KI (2008) Crowd-out 10 years later. J Health Econ 27(2):201–217
Hubbard RG, Skinner J, Zeldes SP (1995) Precautionary saving and social insurance. J Political Econ 103(2):360–399
Lo Sasso AT, Buchmueller TC (2004) The effect of state children's health insurance program on health insurance coverage. J Health Econ 23(5):1059–1082
Mailath JG (1987) Incentive compatibility in signaling games with a continuum of types. Econometrica 55(6):1349–1365
McDermott D, Cox C (2020) Insurer participation on the ACA marketplaces, 2014–2021. https://www.kff.org/private-insurance/issue-brief/insurer-participation-on-the-aca-marketplaces-2014-2021/. Accessed 15 Jan 2023

Miller DP, Dafny L, Kowalski A et al (2019) The consequences of a public health insurance option. Am J Health Econ 5(2):191–226
Miyazaki H (1977) The rat race and internal labor markets. Bell J Econ 8:394–418
Newhouse JP (1996) Reimbursing health plans and health providers: efficiency in production versus selection. J Econ Lit 34:1236–1263
Olovsson C (2010) Quantifying the risk-sharing welfare gains of social security. J Monet Econ 57(3):364–375
Pauly MV (1984) Is cream skimming a problem for the competitive medical market? J Health Econ 4:87–95
Petretto A (1999) Optimal social health insurance with supplementary private insurance. J Health Econ 18:727–745
Przemysław J, Krasnokutskaya E, Ceccarini O (2019) Skimming from the bottom: Empirical evidence of adverse selection when poaching customers. Marketing Sci 38(4):543–566
Rothschild M, Stiglitz JE (1976) Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Q J Econ 90:629–649
Saltzman, E, Swanson A, Polsky D (2021) Inertia, market power, and adverse selection in health insurance: evidence from the ACA exchanges. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3908686. Accessed 14 Jan 2023
Samuelson PA (1958) An exact consumption-loan model of interest with or without the contrivance of money. J Political Econ 66:321–338
Schoder J, Eugster P, Zweifel P (2013) Insurers, consumers, and correlated risks. J Insur Iss 36(2):194–233
Silow-Carroll S, Rodin D, Dehner T, Bern J (n.d., ca. 2012). Health insurance exchanges: State roles in selecting health plans and avoiding adverse selection. https://www.commonwealthfund.org/ publications/newsletter-article/health-insurance-exchanges-state-roles- selecting-health-plans-and-avoiding-adverse-selection. Accessed 10 Jan 2023
Tinbergen J (1952) On the theory of economic policy. In: Contributions to economic analysis. North Holland, Amsterdam
Trottmann M, Zweifel P, Beck K (2012) Supply-side and demand-side cost sharing in deregulated social health insurance: which is more effective? J Health Econ 31(1):231–242
van Dalen HP, Swank OA (1996) Government spending cycles: ideological or opportunistic? Public Choice 89:183–200
Wilson CA (1977) A model of insurance markets with incomplete information. J Econ Theory 16:167–207
Xu X, Zweifel P (2014) Bilateral intergenerational moral hazard: empirical evidence from China. Geneva Pap 39:651–667
Yaari ME (1987) The dual theory of choice under risk. Econometrica 55(1):95–115
Zhao K (2017) Social insurance, private health insurance and individual welfare. J Econ Dyn Control 78:102–117
Zweifel P (2000) Criteria for the future division of labor between private and social health insurance. J Health Care Fin 26(3):38–55
Zweifel P, Courbage C (2015) Double crowding-out effects of means-tested public provision for long-term care. Risks 3(1):61–76
Zweifel P, Eugster P (2008) Life-cycle effects of social security in an open economy: a theoretical and empirical survey. German J Risk Ins 97(1):61–77
Zweifel P, Frech HE III (2016) Why 'optimal' payment for healthcare providers can never be optimal under community rating. Appl Health Econ Health Policy 14(1):9–20
Zweifel P, Breyer F, Kifmann M (2009) Health economics, 2nd edn. Springer, Boston
Zweifel P, Eisen R, Eckles DL (2021) Insurance economics, 2nd edn. Springer, Boston

Part V

Empirical Models

Causality in Empirical Analyses with Emphasis on Asymmetric Information and Risk Management

Georges Dionne

1 Introduction

This chapter presents the basis of applied econometric analysis for causal analysis and discusses examples of applications. Our goal is to improve the understanding of the measurement of economic policy analysis in risk management and insurance. According to Heckman and Pinto (2022), policy analysis must be causal analysis. Causal analysis estimates the factors that generate outcomes and isolates the role of interventions or external shocks in resource allocation. Causal analysis must control for possible counterfactual problems from the control group with appropriate methodologies in order to isolate the appropriate causal factors.

The applications in this new chapter concentrate on causality methodologies for testing the presence or absence of a residual information problem in different markets with an emphasis on insurance and for verifying the effect of risk management on firm value and risk. The revised contributions highlight various difficulties that are not always well understood by those who perform the empirical measurement of causal relationships. In fact, conclusions are often made based on correlations instead of on causality.

One application is measuring the presence of asymmetric information in an insurer's portfolio. Is risk classification sufficient to rule out residual asymmetric information, or do we need self-selection mechanisms inside risk classes? We treat the separation issue between moral hazard and adverse selection and explore how dynamic data can be used to develop tests for achieving such separation. We also analyze the effect of a new insurance pricing scheme in the presence of asymmetric information with a DID model and show that moral hazard is not rejected. Other resource allocation applications, such as reinsurance demand and liquidity creation in the insurance industry, CDS central clearing to reduce information problems in OTC markets, and premiums in M&A transactions in the presence of asymmetric information, are presented with causality analysis.

G. Dionne (✉)
Canada Research Chair in Risk Management and Finance Department, HEC Montréal, Montreal, QC, Canada
e-mail: georges.dionne@hec.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_13

We also review the recent literature on the empirical effect of risk management on firm value when essential heterogeneity is potentially present. Does risk management increase firm value, or do high-value firms engage in more risk management? We also look at the relationship between risky firms and risk management.

We do not consider dynamic treatment analysis in this chapter since the covered applications do not contain events and data with dynamic treatments. In fact, dynamic treatment effects have become popular only recently. Basic causality models in applied econometrics do not use dynamic models. For a dynamic treatment effect analysis with instrumental variables or matching modeling, see Heckman et al. (2016), and for an analysis of dynamic heterogeneous treatment effects that generalizes the standard DID model, see Sun and Abraham (2021).

The rest of the chapter is organized as follows. The remainder of the introduction covers the literature review of asymmetric information in insurance markets and of risk management value. Section 2 presents the fundamental contributions to causality analysis made by three Nobel prize winners in 2021. Section 3 discusses the main issues related to causality in applied econometrics, with an emphasis on instrumental variables and essential heterogeneity. Section 4 presents three applications: two with instrumental variables including one with potential essential heterogeneity and one with the use of propensity score matching methodology in a DID application. Sections 5 and 6 are devoted to different tests of asymmetric information in insurance markets using dynamic data and DID analysis. Section 7 presents an application of the generalized method of moments to analyze the causal relationship between reinsurance demand and liquidity creation in a dynamic environment with the GMM, and Sect. 8 concludes the chapter.

1.1 Asymmetric Information in Insurance Markets

The study of information problems in economics began in the early 1960s. The two best-known problems, moral hazard and adverse selection, were introduced in the literature in 1963 by Kenneth Arrow in a classic article published in *American Economic Review*. In 1970, Akerlof proposed the first analysis of market equilibrium in the presence of adverse selection. Optimal contracts were characterized endogenously (security design) for adverse selection in articles by Rothschild and Stiglitz (1976) and Wilson (1977) and for ex ante moral hazard by Holmström (1979) and Shavell (1979). Ex post moral hazard was defined by Pauly (1968) and was later formalized by Townsend (1979) and Gale and Hellwig (1985).

In the early 1980s, several theoretical developments were advanced to account for different facts observed in markets. Specifically, multiperiod contractual relations were introduced; the renegotiation of contracts was formalized; the problem of contractual commitments was analyzed; and simultaneous treatment of several information problems became a consideration. Other noteworthy proposals have been developed to explain hierarchical relations in firms and organizations, often involving multiparty participants and contracts in insurance, education, healthcare, and risk management.

Economic relationships in banking and insurance contracts, labor and sharecropping contracts, and auctions were studied. Several forms of contracts observed in these markets have been cataloged in various theoretical contributions. The best known are partial insurance coverage (coinsurance and deductibles), compensation based on hours worked and performance, executive compensation with stock options, debt with collateral, bonus-malus schemes, temporal deductibles, and venture capital contracts with warrants. In addition, several corporate organizational practices were rationalized, such as the use of foremen, internal and external controls, auditing, decentralization of certain decisions, and centralization of more difficult-to-control decisions.

The empirical study of information problems began much later. The main motivation was to distinguish the stylized (qualitative) facts used to construct certain theoretical models from real or more quantitative facts. For example, in theoretical contributions, different automobile insurance deductibles may well be used to reduce adverse selection, but there is no evidence that insurers establish this partial coverage for that reason. It can also be argued that labor contracts with performance compensation are used to reduce moral hazard in firms, but it has not necessarily been conclusively empirically demonstrated that there is less moral hazard in firms with this form of compensation than in other firms that use fixed compensation, combined with other incentives or control mechanisms to deal with this information problem.

Another strong motivation for empirically verifying the causal effects of information problems is the search for ways to reduce their negative impact on resource allocation. For example, we know that partial insurance is effective in reducing ex ante moral hazard, as it exposes the insured person to risk. However, this mechanism may be ineffective against ex post moral hazard because the accident has already occurred. Partial insurance may even have pernicious effects and encourage the padding of costs (Dionne and Gagné 2001). The fact that the audit of files seems to be the most effective instrument against ex post moral hazard shows the importance of identifying real problems when attempting to correct imperfections and improve resource allocation.

When it comes to empirically measuring information problems and assessing the effectiveness of mechanisms set up to correct them (the relationship between the nature of contracts and their performance), numerous complications soon arise. Several information problems may exist simultaneously in the database studied; the theoretical predictions must then be carefully defined to distinguish the effects of different information problems on the parameters of the models to be estimated. Moreover, firms have a wide range of mechanisms (substitutes or complementary) at their disposal, which may be selected for reasons other than information problems or for information problems other than those investigated in a particular study. In other words, the information problems under consideration are often neither a necessary nor a sufficient condition to justify the existence of certain observed mechanisms.

Treating several information problems simultaneously in empirical studies is difficult: The literature does not yet offer strong theoretical predictions. If we simply examine whether a market contains any residual information asymmetry, regardless of its origin, it is easier to demonstrate its absence because there is no need to distinguish between the different forms of information asymmetry. Otherwise, we have to ascertain which residual form is still present and document its cause to analyze the instruments that could mitigate or eliminate it.

As a rule, the distinction between moral hazard and adverse selection can be reduced to a problem of causality (Chiappori 1994, 2000). With moral hazard, the nonobservable actions of individuals that affect accidents are consequences of the forms of contracts. For example, a nonoptimal contract may increase the risk of driving because it reduces the incentives to act safely. With pure adverse selection, the nature of different risks already exists before contracts are written. The contracts selected are dictated by the risks present. Thus, there is a potential form of reverse causality between the two information problems. When an exogenous change occurs in an insurance market, we can limit our test to the way it affects existing policyholders in the treatment group and isolate a moral hazard effect if insured in the treatment and control groups are almost identical. Alternatively, we could make comparisons to see whether the chance of accidents differs between new and old policyholders and check for any bias caused by adverse selection. Another approach is to use dynamic models and develop causality tests. However, these tests must consider that other information asymmetries may be present, such as the learning of the contract parties over time (Dionne et al. 2013a, b). Learning may or may not be symmetrical. Dynamic data are also useful for separating moral hazard from unobserved heterogeneity (Abbring et al. 2003; Dionne et al. 2011). Even if standard causality models in the

literature (Arkhangelsky et al. 2021; Heckman et al. 2006) use panel data, they do not necessarily incorporate dynamic behavior in their analysis.

Another difficulty in the empirical measurement of information problems is the fact that researchers are not privy to more information than are decision-makers. On the contrary, they often have less information. Two solutions have been adopted to mitigate that difficulty: (1) use of confidential surveys and (2) development of econometric strategies that can isolate the desired effect. The experimental approach is a third avenue that I shall not deal with in detail.

The survey method has the advantage of providing direct access to private information not available to one party to the contract, such as accidents not claimed or risk perception. Such information makes it possible to measure motivations for choosing specific contractual clauses directly, along with agents' behavior. The drawback of this method is that it is very costly. It can also be biased because it is very difficult to explain the complexity of the problem studied to participants and because several alternative explanations, such as risk aversion, might have been overlooked in the questionnaires. Another source of bias is related to the formation of representative samples.

The development of econometric strategies requires knowledge of the theoretical problem under study and of the econometric methods suitable for the project. This is why the most productive research teams are composed of theoreticians and econometricians. The objective is to isolate effects that are not directly observable by both parties to the contract but that are taken into account by certain variables or combinations of variables. As discussed by Chiappori (1994, 2000) and Gouriéroux (1999), econometric research consists in distinguishing between two sources of information. The first type is composed of variables observable by the two parties to the contract. These variables can be used to make estimates conditional on the characteristics observed. The second type is linked to the information that is not observable by econometricians (and by at least one contractual party) but that may explain choices of contracts or behaviors. In the case of adverse selection, choices of contracts can be interpreted by econometricians as being an endogenous selection. One way to take this into account is to estimate agents' decisions simultaneously by introducing hidden connections (or informational asymmetries) between the decisions. One known test is the nonnull correlation between the random terms of the different equations (contract choice and accident distributions; Chiappori and Salanié 2000). Another test entails estimating the parameters of contract choice on the contract result (Dionne et al. 2001). However, these early studies did not identify the causality effect of information problems on resource allocation (Dionne et al. 2009). They are limited to the presence or absence of residual asymmetric information and to correlations of contract forms and accidents. They cannot conclude that accidents cause contract forms (adverse selection) or that contract forms cause accidents (moral hazard).

The quality of data is a determining factor in the measurement of desired effects. The data must correspond directly to the contractual relations studied and to the duration of the contractual periods. There must also be access to data broken down by contract. The effort involved in formulating raw data for research purposes should not be underestimated. Raw data are used in the day-to-day operations of firms that are not concerned with research problems and do not always contain direct information on the variables needed for the problem studied.

Econometric specifications must correspond to the theoretical models under consideration if erroneous conclusions are to be avoided. Often, researchers choose (or are forced) to use only part of the information available to decision-makers and thus bias the effects of certain treatments such that they capture the effects of other forgotten or inaccessible variables (unobserved counterfactuals) and obtain false conclusions.

Finally, the participants in different contracts are often risk averse to varying degrees. This characteristic is also difficult to observe and can be a source of asymmetric information. Some authors have recently proposed models that take into account the varying degrees of risk aversion (Cohen and

Einab 2007), but very few predictions can isolate the effects of information problems, as they relate to varying degrees of risk aversion among agents.

1.2 Risk Management Value

When there are no market imperfections, market prices contain all the information, making it impossible to generate a profit based on informational advantages. Although this result is widespread, many managers continue to believe that they possess comparative advantages in certain markets. Consequently, firms use their resources to develop investment strategies that are risky because a high return is generally accompanied by high risk. However, these practices are not followed by firms that realize that they do not actually possess comparative advantages within their sector or those that have bad experiences resulting from the inappropriate use of hedging instruments. In fact, firms do not necessarily need to hedge against all the financial risks they may face, particularly when they are already well diversified internally.

The main goal of risk management is to increase firm value by reducing the cost of risk when there are market imperfections. The four main sources of market imperfections are default costs, agency costs, investment financing, and taxes. Dividend payments, managers' risk attitudes, and corporate governance problems may also explain the risk management of nonregulated firms (Dionne and Ouederni 2011; Dionne 2019).

Market imperfections generate default costs. Default costs refer to the costs associated with default, not bankruptcy (Stulz 1996). Default costs can be divided into two categories: direct costs, such as lawyer fees, consultant fees, and court-related expenses, and indirect costs incurred when a firm is under bankruptcy protection laws, such as reorganizational costs. These two categories of costs are directly reflected in a firm's valuation. The goal of an efficient risk management strategy is to maintain these costs at an optimal level while taking into consideration the cost of hedging activities.

Risk management can allow a firm to reduce the expected tax payments when the taxation function is convex with respect to profits or firm value (Graham and Smith 1999; Graham and Rogers 2002). A good risk management strategy may increase a firm's debt capacity and its capital structure. In other words, risk management can be interpreted as a substitute for equity by reducing the default probability and hence the default risk premium imposed by banks or investors. By reducing the risk premium, hedging can create new investment opportunities financed by debt (Smith and Stulz 1985; Campello et al. 2011; Dionne and Triki 2013). Conversely, capital structure can also impact how a firm approaches risk management (Stulz 1996); as a result, causality analysis is essential for establishing real effects.

Under asymmetric information, the external financial costs of investment are much higher than the internal financial costs (Froot et al. 1993). This situation increases the incentives to protect internal financing with risk management. Firms whose managers are also shareholders (meaning that they also benefit from the firm's profits) are usually poorly diversified. Tufano (1996) tested this premise for firms in the gold mining industry. He found that managers who have a large portion of their human capital and compensation invested within their firm seek to protect themselves more by using firm risk management. Attributing firm equity to managers is beneficial when it comes to risk management, yet this incentive can often be more costly than stock options. Stulz (1996) explains why firms that compensate managers with stock options may be more lax with respect to risk management since risk management may reduce the probability of the option being in the money at the end of the period and the possibility of making money with options. Good corporate governance reduces the agency costs of risk management between the board and managers (Dionne et al. 2019).

There are many other motivations for firm risk management. They include a lack of liquidity, mergers and acquisitions, higher productivity in producing goods and services, and other strategic behaviors (Dionne 2019). The main question is the extent to which risk management increases firm value and reduces risk. We will see below that the current empirical evidence is ambiguous. We argue that this can be due to methodological problems such as the absence of causality analysis.

2 Three Nobel Lectures in Economics on Causality

2.1 Basic Model

The three 2021 Nobel lectures in economics were on causality analysis. Establishing causal effects is very important for decision-makers. Three distinct literature studies proposed developments of methodologies for estimating causal effects: in statistics, econometrics, and computer science.

Research on causality in statistics started in the first half of the twentieth century (Imbens 2022, Nobel lecture). It was not until the 1990s that the term causality began to become popular in applied econometrics and statistics. Before that period, correlation was often associated with implicit causality without appropriate methodology.

The research in the statistical literature started with randomized experiments where the treatment and the control groups were formed randomly by applying, for example, the randomized controlled trial (RCT) methodology. Randomization simplifies the comparison of treatment and control groups before treatment. Randomization balances the covariates (observable and nonobservable) between the two groups. However, many real-life applications do not correspond to randomization. Empirical studies with different groups are more common.

Let us start with an example presented by Imbens (2022), where causality is not limited to testing and estimating causal effects in randomized experiments. Rubin (1974) defined the causal inference problem where one compares two potential outcomes defined for the same agent i. $Y_i(C)$ is the agent's outcome in the control group, and $Y_i(E)$ is the outcome for the same agent in the experimental group, at the same time, with exposure to the treatment. The causal effect is the difference $Y_i(E) - Y_i(C)$, or the comparison of outcomes from a treatment and a control situation. The average causal effect over the population is equal to

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i(E) - Y_i(C)\right), \tag{1}$$

where N is the number of agents in the studied population. This causal example has an identification problem: the two potential outcomes cannot be observed for the same agent at the same time. Holland (1986) referred to this as the fundamental problem of causal inference, implying that causal analysis must be performed on different groups or populations, which complicates the exercise because the two groups must be comparable in order to isolate a causal effect. The composition of the groups should not be affected by the treatment, and the nonobservable characteristics of the subjects should be balanced between the groups. In a study with randomized group formation, the two groups obtained from a fair lottery on a given population are similar before the treatment if the lottery is well managed.

Rosenbaum and Rubin (1983a, b) focused on the case where assignment to treatment is not completely random but is conditional on some observed confounders (covariates) that may affect both the treatment and the outcome. If the observed confounders for agent i are denoted by X_i, the absence of confoundedness is measured by the conditional independence on X_i of the treatment W_i and potential outcomes Y_i. Let $W_i \in \{C, E\}$, we can write

$$W_i \perp (Y_i(C), Y_i(E)) \,|X_i. \tag{2}$$

where $\perp$ is for independence. The distinction between confounders or covariates X_i and causal variables W_i is important in this framework. The absence of confoundedness is an assumption (to be tested) in the process that determines the effect of W_i on outcomes Y_i. Without this absence, one cannot draw any conclusions about the causality of the treatment.

Formally, unconfoundedness requires that the probability of treatment assignment is free of dependence on the potential outcomes. "Outcome under the control treatment, $Y_i(C)$, given active treatment and given covariates, is identical to its distribution conditional on control treatment and conditional on covariates, and second, that, analogously, the conditional distribution of the outcome under the active treatment $Y_i(E)$, given receipt of the control treatment and conditional on covariates, is identical to its distribution given the active treatment and conditional on covariates" (Imbens and Rubin 2015). Informally, unconfoundedness requires a sufficiently rich set of pretreatment variables that remove systematic biases from comparisons between treatment and control units. The unconfoundedness assumption is not directly testable. The issue is that the data are not directly informative about the distribution of the control income $Y_i(C)$ for those who received the active treatment (for those with $W_i = E$). Furthermore, the data is not directly informative about the distribution of the active treatment outcome given a receipt of the control treatment (for those with $W_i = C$, for which we never observe $Y_i(E)$). Thus, the data cannot directly provide evidence to support the validity of the unconfoundedness assumption. A nonobserved covariate or confounder could explain the treatment result. Causal estimations can be used to isolate the effect of the treatment. The instrumental variable approach can be useful under assumptions more restrictive than under unconfoundedness. Replication is another way to validate the treatment result when the first study uses a small sample. This method reduces bias in the original study.

This statistical approach differs from early studies in econometrics, where treatments stem from economic agents' welfare maximization behavior. Leamer (1983) and Lalonde (1986) significantly influenced the development of causality in applied econometric studies, such as the research of the Princeton labor group (Ashenfelter, Card, and Krueger) and the credibility revolution of causality analysis in the econometric literature, to which Imbens and Angrist have contributed.[1]

2.2 Local Average Treatment Effects (LATEs)[2]

By testing which instrumental variable (IV) estimates are valid, the local average treatment effects (LATEs) framework is appropriate for measuring causality under certain circumstances. In his Nobel lecture, Angrist (2022) presented empirical examples to illustrate the value of the LATE framework for causal inference. LATE identifies independence conditions satisfied by random or nonrandom assignment from other exclusion restrictions.

An empirical strategy for policy evaluation is a research strategy that includes data collection, identification of causal factors, and estimation. Angrist (2022) considers the randomized clinical trials (RCTs) model a powerful research approach. As mentioned above, random assignment ensures that the treatment and control groups are comparable in the absence of treatment. As a result, post-treatment differences in average outcomes reflect the treatment effect.

[1] Well-known examples of early studies include those of Angrist (1990), Angrist and Krueger (1991), Card (1990), Ashenfelter and Krueger (1994), Card and Krueger (1994), and Imbens and Angrist (1994).

[2] LATE is a discrete approximation of the marginal treatment effect (MTE) discussed below (Heckman et al. 2006).

Angrist (2022) identified instrumental variable (IV) methods and regression discontinuity (RD) designs for RCTs when experimental random assignment was not possible. In applications of IV and RD, causal variables of interest are referred to as treatment variables.

Angrist (2022) presented a contribution by Bloom (1984) that inspired the LATE model. Consider a clinical trial that offers a treatment randomly. Proportion π receives the treatment, while the rest of the group does not. Those who are offered treatment are identified with the dummy variable Z_i, and those who receive the treatment are identified with the dummy variable D_i. Potential outcomes for subject i in the treated and untreated groups are denoted by Y_{1i} and Y_{0i}, respectively. The observed outcome is

$$Y_i = Y_{0i} + D_i \left[Y_{1i} - Y_{0i} \right]. \tag{3}$$

$Y_{1i} - Y_{0i}$ measures the causal effect of treatment on individual i. Bloom (1984) shows how to compute the average effect on the treated in this scenario. Let δ be the causal effect of the treatment assigned on Y_i. Then,

$$E\left[Y_{1j} - Y_{01} \mid D_i = 1 \right] = \frac{\delta}{\pi}. \tag{4}$$

This is an IV experiment that uses the treatment assigned as an instrument for the treatment received (D_i). The LATE model generalizes this result.

Imbens and Angrist (1994) were among the first to propose conditions for considering essential heterogeneity and developed the LATE measure of policy intervention. The local average treatment effect can be measured as

$$\text{LATE} = E\left(Y_i(E) - Y_i(C) \mid W > 0 \right) \tag{5}$$

where W is a policy variable. The LATE theorem generalizes Bloom's result by adding the monotonicity condition.

One concern is that LATE estimates the average effect for a subpopulation that may not be identified. Moreover, LATE is not necessarily invariant to the assignment mechanism. Some inferences can be made, however, about this subpopulation (Abadie 2003).

Angrist and Imbens (1995) extended the basic model to multiple endogenous variables. Motivated by Angrist and Krueger (1991), Angrist and Imbens (1995) also extended the LATE model to ordered treatments such as years of schooling, while Angrist et al. (2000) added continuous treatments and simultaneous equation models.[3]

2.3 Instrumental Variable May Not Be Sufficient

To better identify the average treatment for a subpopulation, Angrist (1990) proposed the following instrumental variable application. The author was interested in the causal effect of serving in the military ($W_i = 1$) on earnings (Y_i). The simple comparison of earnings between veterans $Y_i(1)$ and nonveterans $Y_i(0)$ may not be sufficient to establish a causality effect of serving in the military. Omitted covariates may cause bias in a comparison of veterans' and nonveterans' earnings, even

[3] These extensions are discussed in Imbens (2022).

when controlling for various observed characteristics of the individuals under the assumption of the absence of confounders.

Angrist (1990) used draft eligibility as an econometric instrument (Z). It is clear that the instrument can affect the probability of serving in the military. If the instrument is correlated with the endogenous variable of serving in the military and not with estimated earnings, it satisfies the relevancy condition (Staiger and Stock 1997). It could also have satisfied the exclusion restriction, namely, that the only effect of being draft-eligible was serving in the military. Even if, from standard econometric modeling, this approach seems adequate, it is not clear that it is sufficient to isolate the causal effect of military service on earnings.

Indeed, Heckman (1990) and Manski (1990) showed that it is not possible to identify the average causal effect of military service without additional assumptions. Heckman (1990) argued that identification of this average effect requires z values for the instrument, such that the probability of being a veteran, conditional on the instrument, $P(W_i = 1 | Z_i = z)$, is arbitrarily close to zero or one. Manski (1990) defined sample bonds with the same properties as the condition proposed by Heckman (1990). These contributions motivated the development of the monotonicity condition by Imbens and Angrist (1994).

The development of methods to isolate causal effects has inspired research in statistics on instrumental variables and research in econometrics on matching methods and experimental design. These developments are well described in the works of Imbens (2022) and Angrist (2022).

2.4 Propensity Score

The rise of applied econometrics has generated a wave of theoretical econometric innovations. One methodology application was built on Rosenbaum and Rubin's (1983b) propensity score theorem. The propensity score is the probability that an individual with different covariates is assigned to the treatment group. This theorem contributed to the credibility revolution of causality analysis in the econometric profession by shifting econometricians' attention to the foundations of treatment analysis rather than simply estimating models for outcomes. Dehejia and Wahba (1999) were the first to demonstrate the value of the propensity score for applied work, while Hirano et al. (2003) and Angrist and Hahn (2004) addressed important theoretical questions about the score model. We present an application of the propensity score approach in Sect. 4.

2.5 Difference-in-Differences (DID)

Transparency around critical assumptions is crucial when studying causality in economics. Understanding the limits of what data can credibly tell us and making the research accessible to a broader social science community were important goals for David Card. In his Nobel lecture, Card (2022) presented the design-based perspective approach, with an emphasis on understanding the assignment mechanism. He contended that design-based studies put causality analysis front and center in research projects. Design-based studies are particularly useful for testing basic predictions of a theory or for testing between competing theories.

For Card (2022), design-based research is a solution to the problems of credible inference identified by Ashenfelter and Heckman (1974), Leamer (1978), and Hendry (1980), among others. Ashenfelter and Card proposed the term difference-in-differences in 1985 in a paper in which they identified problems with the newest generation of models that were supposed to separate causal effects from selection biases in longitudinal models. Does the control group evolve in the same way as the

treatment group before the treatment date? The parallel trends test can verify this. Card contributed to the definition and importance of the control groups. An application of a DID analysis with a parallel trends test is presented in Sect. 6.3.

In most design-based studies, such as a classical randomized controlled trial (RCT) study, the counterfactual state for treatment is the state experienced by the comparison group. In RCT studies, the groups are often randomly chosen. One concern for Card (2022) is that the particular counterfactual identified by the design may be too restrictive.

Card was also very active in discussions with practitioners (Goux and Maurin 2023). He contributed to many debates on subjects such as the minimum wage, unemployment insurance effects, and inequalities in wages in developed countries. Education and immigration were important research topics for him.

2.6 Synthetic Control (SC)

Synthetic control (SC) is intended to improve the quality of the control or comparison group when the data do not permit robust parallel trends analysis with only one control group (Abadie and Gardeazabal 2003; Abadie 2021; Abadie and L'Hour 2021; Abadie et al. 2010, 2015). The goal is to obtain the best control group that reproduces the counterfactual outcome that the treated agents would have experienced in the absence of the treatment. The idea behind the SC method is to use information from many potential control groups instead of only one to ensure a better comparison with the treated group. The created SC group is a weighted average of chosen control groups.

2.7 Synthetic Difference-in-Differences (SDID)

A new estimator for causal effects with panel data that builds on difference-in-differences (DID) and synthetic control (SC) methods was recently published (Arkhangelsky et al. 2021). The authors found, theoretically and empirically, that the synthetic difference-in-differences (SDID) estimator has robust properties and that it performs well where other estimators are commonly used. They studied the asymptotic behavior of the estimator when the systematic part of the outcome model included latent unit factors that interact with latent time factors.

As Currie et al. (2020) demonstrated, DID methods have been widely used in applied economics over the last three decades. More recently, SC methods have emerged as important alternative methods for comparative case studies. DID methods are often applied in cases with a large number of agents or units that are exposed to the policy in a panel data study. The parallel trends assumption must be tested in this environment. SC methods, often used in applications with small number of agents exposed, compensate for the lack of parallel trends by weighting control groups. Similar to SC, the SDID method weights and matches preexposure trends of different control groups. Like DID, the SDID method allows for valid large-panel inference. The model also extends the difference-in-differences approach by permitting the effect of unobserved factors on outcomes to vary with time. Dionne et al. (2023) presents an application of the model.

2.8 Regression Discontinuity (RD)

Regression discontinuity was introduced by Thistlethwaite and Campbell (1960) for estimating treatment effects in a nonexperimental setting and was used intensively by Angrist (2022) in different

studies of attendance at school. The treatment is obtained when an observed assignment variable D exceeds a known cutoff point c. Consider the example in which Thistlethwaite and Campbell (1960) analyzed the impact of merit awards on future academic outcomes (Lee and Lemieux 2010). The treatment was given to agents with a score value D greater than or equal to c. Individuals with a score value just below the cutoff are quite comparable to those just above the cutoff, so they can be considered to form the control group.

It is clear that agents should not be able to manipulate D, the causal variable. With no manipulation, the variation in the treatment near c can be considered random, as in a randomized experiment. Other covariates do not matter in that case, contrary to the IV and DID regressions, where tests must be done on potential counterfactuals to validate the causal conclusions. In that sense, RD is a less restrictive research design, which may explain why it has become very popular in recent years.

2.9 Causality and Computer Science

In terms of causality within the computer science literature, Pearl (2009) developed Directed acyclic graphs (DAGs) and structural causal models (SCMs) to study causality. According to Imbens (2022), the DAG approach has two distinct benefits. The first is pedagogical, which involves formulating critical assumptions that capture causal relationships. A DAG can be a powerful way to illustrate the key assumptions underlying causal models. A second potential benefit lies in the mathematical tools developed in the recent DAG literature. For example, Do-calculus, developed by Pearl (2012), can be used to answer causal questions differently than in econometrics.

3 Causality in Applied Econometrics

3.1 Introduction

An econometric model is based on interpretable behavior that comes from economic theory, where economic agents maximize their behavior. The model allows one to test the sources of potential outcomes. One of its activities consists of constructing credible policy counterfactuals, including the forecasting of policy impacts in new environments. The econometric approach considers the definition of causal parameters and their identification as two different tasks.

Two well-known approaches (the Neyman-Rubin-Holland approach in statistics and the Do-calculus approach in computer science) address some of the same problems as those in the econometric approach. All approaches start from the basic intuitive definition of a causal effect as a ceteris paribus consequence of a policy change. However, the rules of constructing and identifying counterfactuals in statistics and computer science are very different from those in econometric analysis.

Heckman and Pinto (2022) presented the econometric causal model for policy analysis developed from the seminal contributions of Frisch (1930, 1938) and Haavelmo (1943, 1944). They compared the econometric causal approach with the two other causal frameworks discussed above: the Neyman-Rubin (1923) and Holland (1986) (NH) causal model and the Do-calculus (DoC) causal model (Pearl 2012).

Heckman and Pinto (2022) argued that economists who use these two causal frameworks often discount the benefits of the econometric causal approach. These two models may have limitations when applied to problems analyzed by economists. The NH approach does not consider unobservable variables and some restrictions on empirical relationships derived from economic theory. In contrast,

Table 1 Four potential causal models

Causal Model 1	Causal Model 2	Causal Model 3	Causal Model 4
$T = f_T(\varepsilon_T)$	$T = f_T(\varepsilon_T, \varepsilon_V)$	$T = f_T(\varepsilon_T, U)$	$T = f_T(\varepsilon_T)$
$U = f_U(\varepsilon_U)$	$U = f_U(\varepsilon_U, \varepsilon_V)$	$U = f_U(\varepsilon_U)$	$U = f_U(\varepsilon_U, T)$
$Y = T\beta + U$	$Y = T\beta + U$	$Y = T\beta + U$	$Y = T\beta + U$

Source: Heckman and Pinto (2022)

the DoC approach does not treat the functional restrictions and covariance information used in econometrics. It cannot incorporate monotonicity and separability restrictions, which are important components of instrumental variable analysis. For example, the generalized Roy model cannot be identified with this approach.

We now present the causality framework proposed by Heckman and Pinto (2022).

3.2 Causality and Mapping

Heckman and Pinto (2022) start with a formal definition of causality that relies on mapping inputs X (usually a vector) to an output Y. A map is stable if changing its arguments over the domain of X preserves the mapping relationship. A simple linear model obtained from a map between X and Y is

$$Y = \alpha + \beta X. \tag{6}$$

Stability means that α and β are not affected when X is changed. This is the invariance of relationships concept proposed by Frisch (1938). Directionality is central to causality analysis. Inverting a map may produce a stable relationship that is not necessarily causal. In their framework, other potential outcomes associated with different values of X are defined as counterfactuals associated with X. The causal effect is the comparison of Y values for different values of X, ceteris paribus.

3.3 Regression Analysis

Equation (7) presents a simple regression of Y on T, where Y and T are observed variables and U is an unobserved variable:

$$Y = T\beta + U. \tag{7}$$

In terms of (6), $X = (T, U)$. The random link between T and Y operates through two channels: β and $E(U|T = t)$. The presence of U complicates causal analysis because U is not observable.

Heckman and Pinto (2022) considered four potential causal models, as illustrated in Table 1, where the different ε variables are independent external random variables not caused by T, U, and Y.

In Model 1, T does not cause U, nor does U cause T. β measures the causal effect of varying T on Y for a fixed value of U. Since variables T and U are independent, the parameter β can be estimated by OLS. In the second causal model, once again, T does not cause U, and U does not cause T. Parameter β is still the causal effect of T on Y. T and U are no longer statistically independent because they are related to a common confounding variable ε_V. The OLS estimator of β could be biased. In Model 3, U causes T and the causal effect of T on Y remains β. In the fourth model, T causes U and the OLS

Table 2 Three distinct tasks in causal policy analysis

Task	Description	Requirements	Types of analysis
1: Model creation	Defining the class of hypotheticals or counterfactuals by thought experiments (models)	A scientific theory: a purely mental activity	Outside statistics; hypothetical worlds
2: Identification	Identifying causal parameters from a hypothetical population	Mathematical analysis of point or set identification: a purely mental activity	Probability theory
3: Estimation	Estimating parameters from real data	Estimation and testing theory	Statistical analysis

Source: Heckman and Pinto (2022)

estimator of the parameter β does not identify this causal effect because T also affects Y. The OLS estimator of β captures both the direct and indirect effects of T on Y.

As observed, the OLS model is not necessarily related to a causal interpretation. The OLS estimates the value for Y conditioning on $T = t$. It evaluates the conditional expectation $E(Y|\, T = t)$ instead of the causal expectation $E(Y(t)|T = t)$, where $Y(t)$ is the value of Y when T is externally fixed to a value t in a causality analysis.

The econometric approach to causality develops explicit hypothetical models in which inputs cause outcomes. According to Heckman and Pinto (2022), the mechanisms governing the choice of inputs are central to studying the causal effect of a treatment on outcomes. Identification of empirical counterparts to the hypothetical counterfactuals requires careful analysis of unobserved variables (U) that cause both input choice and outcomes. Structural econometric models can do that.

Heckman and Pinto (2022) present, in Table 2, three tasks of econometric causal policy analysis. This framework, inspired by the works of Haavelmo (1943, 1944), distinguishes the types of policy analysis where model creation does not require statistical analysis, identification is based on probability theory, and estimation implies statistical analysis.

3.4 Generalized Roy Model

The Model

Let us consider the Generalized Roy model as an example of a structural model. The Generalized Roy model of counterfactuals analyzes earnings in two sectors of the economy. All agents have two potential incomes: $Y(0)$ in Sector 0 and $Y(1)$ in Sector 1. Agents choose a sector from their potential benefit $I = Y(1) - Y(0)$. Potential incomes depend on observed variables X, while I may depend on X and on an external variable Z, which may be a policy variable that influences participation. The agent's choice of Sector 1 ($T = 1$) is identified by $T = 1[I(X, Z) > 0]$.

The Generalized Roy model can be represented by four variables $\{Z, V, T, Y\}$. Z is an external policy variable that causes the treatment T, which causes the outcome Y. Z is an instrumental variable. It causes Y only through its effects on T. V contains confounding variables that jointly cause T and Y. Only variables Z, T, and Y are observed. V is a source of bias in treatment choice, which makes the evaluation of the causal effect of T on Y more difficult. The observed relationship between T and Y may be due to the common effect of V on both T and Y, instead of the causal effect of T on Y. The average treatment effect (ATE) is the mean treatment effect over all agents, knowing that each agent cannot be in both sectors at the same time: ATE $= E(Y(1) - Y(0))$.

Equations (8) to (11) represent the Generalized Roy model. The four error terms (ε) are independent, so $Z \perp V$ and $Y \perp Z|(T, V)$. Eq. (8) states that the identification of causal effects in the Generalized Roy model entails controlling for the unobserved confounding variables V. A popular approach for doing so uses instrumental variables (IVs) that are independent of V. They control for V by shifting T without affecting the distribution of V, as in (10):

$$V = f_V(\varepsilon_V), \tag{8}$$

$$Z = f_Z(\varepsilon_Z), \tag{9}$$

$$T = f_T(Z, V, \varepsilon_T), \tag{10}$$

$$Y = f_Y(T, V, \varepsilon_Y). \tag{11}$$

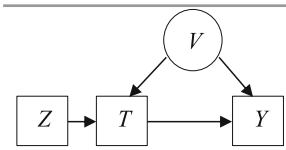
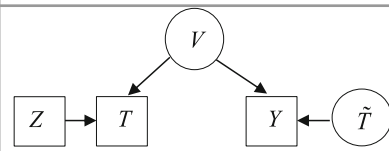
Table 3, from Heckman and Pinto (2022), represents the empirical Generalized Roy model. Directed arrows identify causal relationships, circles denote unobserved variables, and squares denote observed variables. Hypotheses for each variable of the model are presented in the LMC column.

Table 3 also represents the hypothetical causal model developed by Frisch (1938). This model has the same structure as the empirical Generalized Roy model. It adds an external variable $\tilde{T}$ that replaces T. Both variables are independent. $\tilde{T}$ is directly related to Y in the hypothetical model. It replaces the link between T and Y. The hypothetical model is related to experiments, while the empirical model is generated from data.

Counterfactuals are obtained by external manipulations of treatments. These are produced in the hypothetical model by conditioning on the variable $\tilde{T}$. For instance, the distribution of the counterfactual outcome Y, when the treatment is externally set to a value t, is $P_h\left(Y \middle| \tilde{T} = t\right)$, and the counterfactual outcome mean is given by $E_h\left(Y \middle| \tilde{T} = t\right)$. These values are different from those in the empirical counterparts $P_e(Y|T = t)$ and $E_e(Y|T = t)$.

Counterfactuals are said to be identified when they can be expressed in terms of the probability distributions of the observed data in the empirical model. This identification requires the econometrician

Table 3 Generalized Roy model: empirical and hypothetical causal models

Empirical model		Hypothetical model
 LMC		 LMC
V:	$V \perp Z$	$V \perp \left(Z, \tilde{T}\right)$
Z:	$Z \perp V$	$Z \perp \left(V, Y, \tilde{T}\right)$
T:	$T \perp \varnothing \mid (Z, V)$	$T \perp \left(\tilde{T}, Y\right) \mid (Z, V)$
Y:	$Y \perp Z \mid (T, V)$	$Y \perp (Z, T) \mid \left(\tilde{T}, V\right)$
$\tilde{T}$:	Not defined for the model	$\tilde{T} \perp (T, V, Z)$

Source: Heckman and Pinto (2022)
Where $\perp$ means independent of and LMC means local Markov condition. The directions and illustrations are from DAGs (directed acyclic graphs).

to connect the probability distribution of the hypothetical model with the probability distributions of the empirical model.

Counterfactuals in the Generalized Roy Model

The Generalized Roy model can be used to identify counterfactuals. Heckman and Pinto (2022) described several of these approaches. We limit ourselves to the IV model and the Matching model and present three applications. The standard IV model described by Eqs. (10') and (11') with Z as an instrument cannot identify all counterfactuals without additional assumptions:

$$T = \alpha_0 + \alpha_1 Z + \varepsilon_T \tag{10'}$$

$$Y = \beta_0 + \beta_1 \overline{T} + \varepsilon_Y \tag{11'}$$

where β_1 measures the causal effect of $\overline{T}$ on Y and $\overline{T}$ is the predicted value of T conditional on Z.

The Generalized Roy model is not necessarily captured by this system of two equations often used in applied econometrics. The causal effect, $\beta_1 = Y(1) - Y(0)$, is, in general, a random variable and not a constant, so treating the causal parameter in Eq. (11') as a constant does not capture the essential heterogeneity of treatment effects across agents or units. In many applications, the causal effect is stochastically dependent on $\overline{T}$.

LIV Model

Heckman and Vytlacil (1999, 2001, 2005) and Heckman et al. (2006) presented the IV problem by introducing a separable choice equation. Their approach enables analysts to control for V and to identify counterfactual outcomes.

The local instrumental variable (LIV) model considers a binary treatment $T \in \{0, 1\}$. The separability assumption states that treatment is given by a latent equation that includes the instrumental variable Z and the confounder V. Separability enables us to rewrite the choice equation as

$$T = 1\,[P(Z) \geq U]\,; \quad P(Z) = P_e\,(T = 1\,|Z)\,, \tag{12}$$

where $P(Z)$ is the propensity score obtained from the probit model.

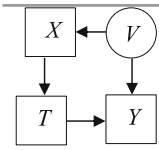
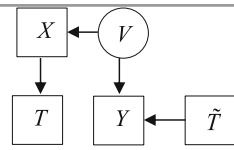
The LIV model (Heckman and Vytlacil 1999) can be used to identify the distributions of counterfactual outcomes conditioned on U by taking the derivative of the observed outcome with respect to the propensity score. More generally, the counterfactual expectation is identified if there is sufficient variation in the propensity score $P(Z)$ around the value $u \in (0, 1)$.

If $P(Z)$ has full support, the average treatment effect (ATE) can be evaluated by

$$\text{ATE} \equiv E_h\left(Y \,\middle|\, \tilde{T} = 1\right) - E_h\left(Y \,\middle|\, \tilde{T} = 0\right) = \int_0^1 \left(E_h\left(Y \mid T = 1, U = u\right) - E_h\left(Y \mid T = 0, U = u\right)\right) du. \tag{13}$$

The solution of the LIV model implies a separability assumption that generates a score U for V. Equation (13) assumes that the sample propensity score has enough variation around the value $u \in (0, 1)$. The equation is not directly applicable to discrete instruments. Discrete instruments are discussed by Heckman and Pinto (2022).

Table 4 Matching model: empirical and hypothetical causal models

Empirical model	Hypothetical model
	

Source: Heckman and Pinto (2022)

Matching Model

Another method for identifying treatment effects assumes that a set of observed pretreatment variables suffices to control for the confounding variable V. Matching assumes that the observed variable X is a balancing score for the confounding variable V.

The average causal effect of a binary treatment $T \in \{0, 1\}$ is evaluated by the weighted average of the mean difference between the treated and nontreated participants that match X, namely,

$$\text{ATE} = \int \left(E_e\left(Y \mid T = 1, X = x\right) - E_e\left(Y \mid T = 0, X = x\right)\right) \; dF_{e,X}(x) \tag{14}$$

where e is used to identify the matching relationships.

The matching assumption replaces the unobserved variable U of the Generalized Roy model in Table 4 with the observed variable X. In practice, it assumes that potential bias can be eliminated by controlling for observed pretreatment variables between the agents in the treated and nontreated groups. Under matching, the identification of treatment effects does not require an instrumental variable or separability. This assumption enables us to solve the problem of selection bias induced by unobserved variable V by conditioning on the observed variable X, which makes the two populations of agents comparable, as in a randomized study.

4 Three Applications of Econometric Models

Two applications apply the IV model while the third uses the Matching model.

4.1 Effect of Risk Management on Firm Value[4]

Reverse causality between firm hedging behavior and other firm financial variables is a crucial concern in risk management studies; it has been identified as the major source of inconsistency in the literature. To control for this endogeneity, Dionne and Mnasri (2018) studied the real effects of hedging using an instrumental variable approach applied to the essential heterogeneity model (Heckman et al. 2006). They controlled for biases related to selection of unobservables and self-selection in the estimation of the marginal treatment effects (MTEs) of hedging choice on firm market value, risk, and accounting performance. They also estimated the average treatment effects (ATEs),

[4] We do not study enterprise risk management (ERM) in this chapter. Hedging is a risk management strategy while ERM is a framework for risk management in an enterprise. On ERM, see the chapter by Hoyt and Liebenberg (2025) in this book.

which can be interpreted as the mean of the MTEs. They estimated the hedging equation with the probit model to reduce the potential bias associated with unobserved heterogeneity.

In the last three decades, the risk management literature has improved in terms of data availability and theoretical research on corporate risk management. Mayers and Smith (1982) and Stulz (1984) were the first to build a hedging theory that incorporated the introduction of frictions into financial markets and to show that market frictions (e.g., default costs, tax shields, agency costs, and asymmetric information) enable firms to create value by hedging actively. The subsequent empirical literature extended the knowledge of hedging determinants (e.g., Tufano 1996; Haushalter 2000; Dionne and Garand 2003; Adam and Fernando 2006; Dionne and Triki 2013). Other contributions in the literature focus on hedging value and risk implications for firms (e.g., Guay 1999; Allayannis and Weston 2001; Jin and Jorion 2006). Empirical findings on the value implications of risk management are still fairly mixed and inconclusive. Methodological problems related to the endogeneity of derivatives use and other firm decisions, sample selection, sample size, and the existence of other potential hedging mechanisms (e.g., operational hedging) are often blamed for this mixed empirical evidence.

To overcome the major source of inconsistency in the findings in the empirical literature (i.e., endogeneity), Dionne and Mnasri (2018) applied the instrumental variable approach with a model that considers essential heterogeneity inspired by the work of Heckman et al. (2006), which controls for the individual-specific unobserved heterogeneity in the estimation of marginal treatment effects of using high hedging ratios (i.e., upper quartile) versus low hedging ratios (i.e., lower quartile). Heckman et al. (2006) showed that the basic method of instrumental variables appears to be inappropriate when firms exhibit heterogeneous responses to treatment. In their application of the essential heterogeneity model, Dionne and Mnasri (2018) identified a credible instrument arising from the economic literature pertaining to macroeconomic responses to crude oil price shocks, namely, the Kilian (2009) index, which gives a measure of the demand for industrial commodities driven by the economic perspective.

Their evidence suggests that marginal firm financial value (marginal treatment effect, MTE), as measured by Tobin's q, increases with oil producers' propensity to hedge their oil production to a greater extent (i.e., upper quartile). This finding corroborates the stream of literature that argues for the existence of a hedging premium for nonfinancial firms (Allayannis and Weston 2001; Carter et al. 2006; Adam and Fernando 2006; Pérez-Gonzalez and Yun 2013). Consistent with the literature (e.g., Guay 1999; Bartram et al. 2011), Dionne and Mnasri (2018) find that marginal firm riskiness, as measured by its systematic and idiosyncratic risk, decreases with oil producers' propensity to be high-intensity rather than low-intensity hedgers. The oil beta, representing firms' stock returns' sensitivity to fluctuations in oil prices, decreases with the propensity to hedge to larger extents, albeit with no statistical significance. Altogether, these findings suggest that any potential positive effects associated with oil hedging should translate into value enhancement for shareholders because of the decrease in the required cost of equity due to the lower riskiness of oil producers, in particular lower systematic risk, as suggested by Gay et al. (2011). Dionne and Mnasri (2018) also found that a firm's marginal accounting performance, as measured by the return on equity, is lower for oil producers that are low-intensity hedgers. Finally, the researchers obtained a significant average treatment effect (ATE) on firm value and risk for Tobin's q (positive), idiosyncratic risk (negative), and systematic risk (negative). ATE is not significant for accounting performance even if some firm types obtain significant MTE values.

Instrumental Variables

For the choice of a candidate instrument, Dionne and Mnasri (2018) built on two previous studies that demonstrated a significant impact of oil market conditions (oil spot price and volatility) on oil

hedging design in terms of maturity and type of derivatives (Mnasri et al. 2017; Dionne et al. 2018). They looked for an instrument that can explain fluctuations in real oil prices and that cannot directly affect the value, riskiness, and accounting performance of an oil producer. A large body of economic literature affirms that one of the most important fundamental factors that determines industrial commodity prices is demand pressures or shocks induced by real economic activity. Consequently, they chose the Kilian (2009) index as an instrument. This instrument measures the component of true global economic activity that drives demand for industrial commodities. This index is based on dry cargo (grain, crude oil, coal, iron ore, etc.) and single-voyage ocean freight rates and captures demand shifts in global industrial commodity markets. The Kilian index, constructed monthly, accounts for fixed effects for different routes, commodities, and ship sizes. It is also deflated with the US consumer price index and linearly detrended to remove the decrease in real terms over time of the dry cargo shipping cost. Kilian (2009) showed that aggregate shocks to industrial commodities cause long swings in real oil prices. This differs from the variations in the price of oil induced by oil market-specific supply shocks, which are more transitory. Oil market-specific supply shocks also differ from shocks related to shifts in the precautionary demand for oil, which arise from uncertainty about expected supply shortfalls relative to expected demand. Dionne and Mnasri (2018) calculated the changes in the Kilian (2009) index for each fiscal quarter in the sample. These changes in the index are calculated by taking the index's level at the end of the current fiscal quarter (i.e., at the end of the fiscal quarter's last month) minus its level at the end of the previous fiscal quarter. They showed that an increase in demand for industrial commodities is correlated with an increase in the prices of derivatives such as futures contracts. Consequently, oil hedging intensity should have a negative relationship with the Kilian index.

Essential Heterogeneity Model

The essential heterogeneity model usually begins with a Mincer–like equation (Mincer 1974), as follows:

$$Y_{i,t} = \alpha + \beta D_{i,t} + \Delta X_{i,t} + u_{i,t}, \tag{15}$$

where $Y_{i,t}$ is the risk or value of oil producer i at the end of quarter t and where $D_{i,t}$ is the observed value of a dummy variable $D = (0, 1)$ representing whether oil producer i uses low (0) or high (1) intensity hedging during quarter t. The control variables include a set of observable covariates ($X_{i,t}$). The term $u_{i,t}$ is an individual-specific error term, and β represents the average return from using high-intensity hedging (average treatment effect, ATE). Δ is a vector of parameters.

Two sources of bias could affect the estimates of β. The first is related to the standard problem of selection bias when $D_{i,t}$ is correlated with $u_{i,t}$. This bias could be resolved by using an instrumental variable (IV) method, among others. However, this method may not be sufficient. The second source of bias occurs if the returns from using high-intensity hedging vary across oil producers (i.e., β is random because of firm nonobserved factors that can influence both the firm target and the hedging decision, such as governance of the firm or manager risk aversion), even after conditioning on observable characteristics leading to heterogeneous treatment effects. Moreover, oil producers make their hedging level choice (low versus high intensity) with at least partial knowledge of the expected idiosyncratic gains from this decision.

Heckman et al. (2006) developed an econometric methodology based on IVs to solve the problem of essential heterogeneity (i.e., β is correlated with $D_{i,t}$) in the estimation of marginal treatment effects (MTE). Their model is an extension of the LATE model developed by Imbens and Angrist (1994). MTE is a limit measure of LATE. Heckman et al.'s (2006) methodology is built on the Generalized Roy model, which is an example of a treatment effect model for economic policy

evaluation. The Generalized Roy model involves a joint estimation of an observed continuous outcome and its binary treatment.

We now present the linear and separable version of the essential heterogeneity model. Let (Y_0, Y_1) be the potential outcomes observed under the counterfactual states of treatment (Y_1) and no treatment (Y_0). These outcomes are supposed to depend linearly upon observed characteristics (X_{it}) and unobservable characteristics (U_0, U_1) as follows:

$$Y_1 = \alpha_1 + \beta + \Delta_1 X_{it} + U_1 \tag{16}$$

$$Y_0 = \alpha_0 + \Delta_0 X_{it} + U_0 \tag{17}$$

where β is the benefit related to the treatment $D_{i,\,t} = 1$.

The selection process is represented by $I_D = \gamma Z - V$, which depends on the observed values of the Z variables and an unobservable disturbance term (V) given X. The selection process, related to whether low- or high-intensity hedging is used, is linked to the observed outcome through the latent variable I_D, which gives the dummy variable D, which represents the treatment status:

$$D = \begin{cases} 1 \text{ if } I_{D_{i,t}} > 0 \\ 0 \text{ if } I_{D_{i,t}} \leq 0 \end{cases} \tag{18}$$

where the vector of observed Z variables includes instrumental variables Z_{IV} and all the components of $X_{i,\,t}$ in the outcome equation. The variable Z_{IV} satisfies the following constraints: $\text{Cov}(Z_{\text{IV}}, U_0) = 0$, $\text{Cov}(Z_{\text{IV}}, U_1) = 0$, and $\gamma_{IV} \neq 0$. The unobservable set of (U_0, U_1, V) is assumed to be statistically independent of Z, given X. In this application, one must first estimate the probability of participation in high-intensity hedging or the propensity score and then analyze how this participation affects firm values and risks.

We can assume the joint normality of the outcome's unobservable components and decision equations $(U_0,\ \ U_1,\ \ V) \sim N(0, \Sigma)$, where Σ is the variance–covariance matrix of the three unobservable variables and $\sigma_{1V} = \text{Cov}(U_1, V)$, $\sigma_{0V} = \text{Cov}(U_0, V)$, and $\sigma_{\text{VV}} = 1$ following standard hypotheses. Under this parametric approach, the discrete choice model is a conventional probit with $V \sim N(0, 1)$ where the propensity score $(P(z))$ is given by:

$$P(z) = \Pr(D = 1 | Z = z) = \Pr(\gamma z > V) = \Phi(\gamma z), \tag{19}$$

where $\Phi(\cdot)$ is the cumulative distribution of *V,* a standard normal variable. The term $P(z)$ denotes the selection probability of using high-intensity hedging conditional on $Z = z$. We can therefore write

$$\Phi(\gamma Z) > \Phi(V) \iff P(Z) > U_D, \tag{20}$$

where

$$U_D = \Phi(V) \text{ and } P(Z) = \Phi(\gamma Z) = \Pr(D = 1 | \text{Z}).$$

The term U_D is a uniformly distributed random variable between zero and one representing different quantiles of the unobserved component V in the selection process. These two quantities, $P(Z)$ and U_D, play crucial roles in the essential heterogeneity model. The quantity $P(Z)$ could be interpreted as the probability of going into treatment, and U_D could be interpreted as a measure of individual-specific resistance to undertaking treatment (or, alternatively, the propensity to not be treated as a

high-intensity hedger). In the application, the higher the $P(Z)$ is, the more the oil producer is induced to hedge its oil production to a larger extent due to Z. Conversely, the higher the U_D is, the more resistant the oil producer is to using higher hedging extents due to a larger unobserved component. $P(Z) = U_D$ is thus the margin of indifference for oil producers that are indifferent between low- and high-intensity hedging.

The marginal treatment effects (MTEs) can be measured as follows:

$$\text{MTE}\left(X = x, U_X = u_X\right) = \left(\alpha_1 + \beta - \alpha_0\right) + \left(\Delta_1 - \Delta_0\right) x + \left(\lambda_{1V} - \lambda_{0V}\right) \Phi^{-1}\left(u_X\right) \tag{21}$$

Under the assumption of joint normality, λ_{1V} and λ_{0V} are the inverse Mills ratio coefficients used to reduce selection bias. They are estimated along with the other parameters in the two following equations:

$$E\left(\text{Y}|\text{X} = x, D = 1, P(Z) = p\right) = \alpha_1 + \beta + \Delta_1 x + \lambda_{1V}\left(-\frac{\phi\left(\Phi^{-1}(p)\right)}{p}\right) \tag{22}$$

$$E\left(\text{Y}|\text{X} = x, D = 0, P(Z) = p\right) = \alpha_0 + \Delta_0 x + \lambda_{0V}\left(\frac{\phi\left(\Phi^{-1}(p)\right)}{1 - p}\right) \tag{23}$$

to obtain the $\widehat{\text{MTE}}$ values. Using the estimated propensity score,

$$\widehat{\text{MTE}}\left(X = x, U_X = u_X\right) = \widehat{\alpha_1 + \beta} - \widehat{\alpha_0} + \left(\widehat{\Delta_1} - \widehat{\Delta_0}\right) x + \left(\widehat{\lambda}_{1V} - \widehat{\lambda}_{0V}\right) \Phi^{-1}\left(u_X\right). \tag{24}$$

Intuitively, how the MTE evolves over the range of U_D illustrates the heterogeneity in treatment effects among oil producers, that is, how the coefficient β is correlated with the treatment indicator D in (15). Equivalently, the estimated MTE shows how the increment in the marginal firm value, risk, and accounting performance from choice 0 to choice 1 varies with different quantiles of the unobserved component V in the choice equation. Whether MTE increases or decreases with U_D tells us whether the coefficient β in (15) is negatively or positively correlated with the latent tendency to use high-intensity hedging for oil production. Detailed estimation results, including firms' individual parameters (MTE), are presented by Dionne and Mnasri (2018). They clearly show that MTE varies between firms, meaning that the effect of hedging on firm value and risk is not homogenous between firms.

4.2 Asymmetric Information in Mergers and Acquisitions

Dionne et al. (2015) studied the effect of asymmetric information in auctions. Their application concerns mergers and acquisitions of nonfinancial firms under asymmetric information. Their econometric analysis served to identify the causal effect on the premium of transaction value associated with the presence of blockholders. Does the presence of a blockholder on the board of a target firm affect the premium paid by the winner of the auction? This bidder should have more information on the target than the other potential bidders.

Econometric Model

The authors postulate a linear relationship between the premium from a merger and acquisition and the candidate explanatory variables. They assume that more informed bidders (blockholders) pay a

lower premium when they acquire a firm because other bidders know that the blockholders have more information about the target. Consequently, they bid less actively because they do not want to pay an overly high price (overbidding). The presence of a blockholder on a board is public information in the market since blockholders, in this study, hold more than 5% of the capital stock value of the target. Let X_i denote a row vector that contains all available regressors, including the constant but excluding Blockholders_i. The model becomes

$$Y_i = X_i \underline{\beta} + \beta_{15}\text{Blockholders}_i + \varepsilon_i \tag{25}$$

where $Y_i = \text{Premium}_i$ and $\underline{\beta} = (\beta_0, \ldots \beta_{14})'$. The test for the null hypothesis (H_0) of no information asymmetry consists in verifying $\beta_{15} = 0$ against the alternative (H_1) $\beta_{15} < 0$. β_{15} measures the effect of blockholders' presence. In this application, β_{15} is assumed constant across firms in the absence of essential heterogeneity analysis.

It is straightforward to estimate this model using ordinary least squares. However, there are reasons to suspect that the blockholders dummy variable is correlated with unobservable factors in (25), in which case the ordinary least square estimate of β_{15} would be biased. For instance, blockholders may submit bids during the acquisition process because they want to be more active in a particular industry, and the current offer by a competitor may reduce this opportunity. This behavior is not observed by market participants. The endogeneity of Blockholders_i would imply that $E(\varepsilon_i | X_i, \text{Blockholders}_i) \neq 0$. Below, we examine two approaches to deal with this endogeneity issue.

Treatment Effect Approach

The endogeneity of the Blockholders_i variable can be addressed in line with the literature on the treatment effect (Heckman and Vytlacil 2001). Indeed, the presence of blockholders in an auction may be viewed as a treatment that affects the distribution of the error term in (25) such that:

$$Y_i = X_i \underline{\beta} + \beta_{15} + \varepsilon_{i1} = Y_{i1}, \text{ if } \text{Blockholders}_i = 1, \tag{26}$$

and

$$Y_i = X_i \underline{\beta} + \varepsilon_{i2} = Y_{i2}, \text{ if } \text{Blockholders}_i = 0, \tag{27}$$

where ε_{i1} and ε_{i2} are errors with potentially distinct distributions. One may further ask: Is there any hidden self-selection effect in the process that drives the presence of blockholders in auctions? Indeed, blockholders may be keener to attend auctions where their informational advantage is higher. As a result, their ability to lower the premium would be higher for the auctions in which they participate to win than for other auctions.

According to the treatment effect formulation, the error term of (25) is given by

$$\varepsilon_i = \varepsilon_{i1}\text{Blockholders}_i + (1 - \text{Blockholders}_i)\, \varepsilon_{i2}. \tag{28}$$

To estimate the coefficients of (25) by OLS, the initial set of predictors is augmented with a set of instrumental variables denoted Z_i. Consequently, the premium can be represented as:

$$Y_i = X_i \underline{\beta} + \beta_{15}\text{Blockholders}_i + E\left(\varepsilon_i | X_i, Z_i, \text{Blockholders}_i\right) + \tilde{\varepsilon}_i \tag{29}$$

where the new error term satisfies $\widetilde{\varepsilon}_i = \varepsilon_i - E(\varepsilon_i|X_i, Z_i, \text{Blockholders}_i)$ so that $E\left(\widetilde{\varepsilon}_i|X_i, Z_i, \text{Blockholders}_i\right) = 0$. When estimating (29), we can assume the existence of a latent variable k_i such that $\text{Blockholders}_i = 0$ if $k_i \leq 0$ and $\text{Blockholders}_i = 1$ if $k_i > 0$. This latent variable further satisfies

$$k_i = X_i\underline{\delta} + Z_i\underline{\gamma} + e_i \tag{30}$$

where $e_i \sim N(0, 1)$ and $\underline{\delta}$ and $\underline{\gamma}$ are vectors of parameters.

Equation (30) was estimated by a probit model using three variables as instruments: (i) intrastate, (ii) regulated industry, and (iii) an interaction variable between Intrastate and Performance of the target. The first variable comes from the study by Kang and Kim (2008). They found that blockholders prefer targets in the same state because proximity reduces transaction costs, yielding higher returns. They also showed that the monitoring of intrastate firms is more valuable for targets that have poor past performance. Blockholders are consequently more likely to buy shares in intrastate underperforming firms to better exploit their informational advantage. Blockholders may be more present in poorly performing targets of their state because they have a higher probability of obtaining long-run value by exploiting asymmetric information with other competitors. The other variable controls for the fact that the industry of the target is regulated. Blockholders should better exploit their informational advantage for these firms because they are more knowledgeable about the different laws regulating the target. Considering only auctions where blockholders are present, the effect of their presence compared with the outcome if they were absent is given by

$$E(Y_{i1} - Y_{i2}|X_i, Z_i, \text{Blockholders}_i = 1) = \beta_{15} + \lambda_1 \frac{\varphi\left(\hat{\eta}_i\right)}{\Phi\left(\hat{\eta}_i\right)}, \tag{31}$$

where λ_1 is the coefficient of the inverse Mill ratio (IMR). Similarly, considering only auctions where blockholders are not present, their absence would have induced the following effect:

$$E(Y_{i1} - Y_{i2}|X_i, Z_i, \text{Blockholders}_i = 0) = \beta_{15} - \lambda_2 \frac{\varphi\left(\hat{\eta}_i\right)}{1 - \Phi\left(\hat{\eta}_i\right)}. \tag{32}$$

Accordingly, the average treatment effect (ATE) of the presence of blockholders on the premium is

$$E(Y_{i1} - Y_{i2}|X_i, Z_i) = \beta_{15}.$$

Finally, the equation that allows us to estimate all the parameters without bias is given by

$$\text{Premium}_i = X_i\underline{\beta} + \beta_{15}\text{Blockholders}_i + \text{IMR}_i\underline{\lambda} + \widetilde{\varepsilon}_i \tag{33}$$

where $\underline{\lambda} = (\lambda_1, \lambda_2)$, $\text{IMR}_i = [\text{IMR}_{1i}, \text{IMR}_{0i}]$, $\text{IMR}_{1i} = \frac{\varphi(\hat{\eta}_i)}{\Phi(\hat{\eta}_i)}\text{Blockholders}_i$, and $\text{IMR}_{0i} = \frac{-\varphi(\hat{\eta}_i)}{1-\Phi(\hat{\eta}_i)}(1 - \text{Blockholders}_i)$.

Two Estimation Approaches

Given that Blockholders_i is a binary variable, two approaches can be considered to proxy its expectation conditional on X_i and Z_i. The first approach relies on the linear probability model (LPM):

$$\text{Blockholders}_i = X_i \underline{\delta} + Z_i \underline{\gamma} + e_i, \tag{34}$$

from which the fitted values are deduced as $\widehat{\text{Blockholders}}_i = X_i \underline{\hat{\delta}} + Z_i \underline{\hat{\gamma}}$.

The second approach is based on the probit model presented in the previous subsection, from which fitted values are deduced as $\widehat{\text{Blockholders}}_i = \Phi\left(\hat{\eta}_i\right)$. Either the LPM or the probit model can be a good proxy for $E(\text{Blockholders}_i = 1 | X_i, Z_i)$ (see, however, Heckman and Pinto 2022). Hence, the first-stage functional form does not affect the consistency of the second-stage estimates. The second-stage estimation equation is

$$Y_i = X_i \underline{\beta} + \beta_{15} \widehat{\text{Blockholders}}_i + \hat{\varepsilon}_i. \tag{35}$$

Two formal tests are done to assess the validity of the results: Sargan's overidentifying restrictions test for the exogeneity of the instruments (H0: the instruments are exogenous) and the Durbin-Wu-Hausman test for the relevance of the instrumental variable method (H0: the blockholders variable is exogenous). These tests are performed only within the 2SLS approach, which relies on the LPM (linear) at the first stage of the estimation. The tests cannot be used with the probit (nonlinear) model.

Dionne et al. (2001) (see also Blundell and Smith 1989) propose the following extension of Eq. (35) to make it robust to nonlinearity and misspecification within the framework that employs a probit at the first stage:

$$Y_i = X_i \underline{\beta} + \tilde{\beta}_{15} \Phi\left(\hat{\eta}_i\right) + \beta_{15} \text{Blockholders}_i + \mu_i. \tag{36}$$

By adding and subtracting $\text{Blockholders}_i - \Phi\left(\hat{\eta}_i\right)$ to $\Phi\left(\hat{\eta}_i\right)$, Eq. (36) becomes:

$$Y_i = X_i \underline{\beta} + \left(\beta_{15} + \tilde{\beta}_{15}\right) \text{Blockholders}_i - \tilde{\beta}_{15} \left(\text{Blockholders}_i - \Phi\left(\hat{\eta}_i\right)\right) + \mu_i. \tag{37}$$

Blockholders_i is endogenous if $\underline{\delta}$ in (34) is significantly different from zero. By comparing (35) and (36), we see that Blockholders_i is endogenous if $\tilde{\beta}_{15}$ is significantly different from zero in the model of Dionne et al. (2001). In that event, the (unbiased) coefficient of Blockholders_i is equal to $\beta_{15} + \tilde{\beta}_{15}$, as indicated by (37). The estimation results clearly show that more informed bidders pay a lower premium in mergers and acquisitions (Dionne et al. 2015). The three instrumental variables are valid instruments, meaning that the presence of blockholders is a causal variable that explains the lower premium paid by the blockholders in the presence of asymmetric information.

The estimation results, which consist of the results for the standard OLS, the treatment effect approach, the 2SLS approaches (2SLS-LPM and 2SLS-probit), and Dionne et al. (2001), are presented in Table 5 of the study.

The results from all the models clearly support the hypothesis that information asymmetry between potential buyers significantly influences the premium paid during an acquisition. This result is consistent with the theoretical and empirical study (Dionne et al. 2009) showing that information asymmetry between participants and even adverse selection influences the equilibrium price of an auction.

4.3 CDS Central Clearing

Akari et al. (2021) revisited the impact of the voluntary central clearing scheme on the CDS market. Central clearing is intended to improve transparency in the derivatives market by reducing asymmetric information between market participants in over-the-counter transactions (OTC). This article is part of the ongoing research on the impact of introducing a central counterparty (CCP) that stands between buyers and sellers of default protection in the CDS market.

Akari et al. (2021) applied the dynamic propensity score matching methodology combined with difference-in-differences (DID). Their empirical findings show that central clearing results in a small, but significant, increase (estimated at 19 bps) in CDS spreads, while there is no evidence of an associated improvement in CDS market liquidity and trading activity or of a deterioration in the default risk of the underlying bond. These results suggest that the increase in CDS spreads can be mainly attributed to a reduction in CDS counterparty risk in the central market because CDS buyers are protected against issuer default. The spread, paid by risk-averse investors, is used to finance part of the CCP activities.

Their approach aims at eliminating selection bias in causal analysis. First, they used a dynamic matching technique, relying on dynamic firm data just prior to the move to central clearing, instead of using a fixed estimation period to match all the firms. Second, they estimated a DID model including time and firm fixed effects.

In order to study the impact of central clearing, they compared the spreads of single-name CDS contracts in two groups of firms, namely, cleared reference entities that are members of the clearinghouse and noncleared reference entities that are in the OTC market; this comparison is undertaken before and after adhesion to the CCP in a DID framework.

The classical two-by-two design uses data from a treatment group and from a control group, measured at two different dates: before treatment and after treatment. This methodology is flexible and can be generalized to the case of multiple groups and multiple time periods (Bertrand et al. 2004; Imbens and Wooldridge 2009; Gormley and Matsa 2011). In this application, since the authors were dealing with multiple treatment (clearing) dates, they opted for a DID framework with firm and time-fixed effects.

To apply DID with matching, one needs to form a treatment and a control group containing firms that have similar characteristics just before the treatment event. The first step consists of constructing a sample of candidate treatment and control entities and computing their propensity scores on the basis of preclearing characteristics. Specifically, the authors considered the 29 clearing dates as the various possible times for adhering to a CCP. These treatment dates can be interpreted as hypothetical events for the control group. Each noncleared firm thus generates up to 29 firm-event entities. The sample also contains the cleared firms, paired with the event corresponding to their clearing date.

The probit model was then estimated by using the sample of cleared and noncleared firm-event entities and the corresponding observable variables that are relevant to clearinghouses:

$$\Pr(Y = 1 \,|X) = \Phi(X\beta), \tag{38}$$

where Y is a binary random variable that equals 1 if the firm is centrally cleared and 0 otherwise. Φ is the standard normal cumulative distribution function, X is the vector of regressors that influence the outcome Y, and β is a vector of parameters. The vector β is obtained by maximum likelihood and is used to estimate the probability, for each firm-event entity, of being accepted for central clearing. This probability is the propensity score associated with a combination of firm characteristics and a possible clearing date.

The second step consists of matching cleared and noncleared entities on the basis of the propensity scores. One can match with replacement each cleared firm with its closest neighbor from the group of noncleared firm-event entities. The final sample is then composed of matched firm-event entities.

A DID regression to the matched sample can be applied to test for the presence of statistically significant impact factors such as liquidity. More specifically, to isolate the effect of central clearing on a given factor, one can estimate the following DID equation:

$$\text{Factor}_{i,t} = \beta_0 + \beta_1 \text{Cleared}_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t}, \tag{39}$$

where subscript i denotes a firm-event entity and subscript t denotes a date in the event window. The dependent variable $\text{Factor}_{i,t}$ takes various definitions in order to investigate the impact of central clearing on CDS spreads, liquidity, and trading activity, as well as on bond default spreads. The main explanatory variable $\text{Cleared}_{i,t}$ is a binary variable that indicates whether reference entity i is centrally cleared or not on date t. This variable is the equivalent of the interaction term in the classic two-by-two DID design because matching eliminated the differences between units before the events. The treatment effect is given by the corresponding coefficient β_1. The fixed effects of the generalized DID setting help control for unobserved heterogeneity across time and reference entities, thereby alleviating concerns about any omitted variables that might affect both groups in the same way. The firm fixed effect, α_i, captures differences across firms that are constant over time, while the time fixed effect, γ_t, captures differences over time that are common to all firms. The authors deliberately did not control for specific time-varying variables to avoid confounding estimates of β_1 since these variables might also be affected by the move to central clearing.

5 Measurement of Residual Asymmetric Information in Insurance Data

The objective of this section is to present various tests for the presence of residual asymmetric information in insurance markets. From the theoretical literature (Dionne et al. 2025; Picard 2025; Parra and Winter 2025; Crocker et al. 2025), we know that the potential presence of asymmetric information between insured and insurers regarding individual risk motivates partial insurance, risk classification, long-term contracts, and the auditing of claims. It is also well known from the insurance literature that risk classification is due, in part, to asymmetric information between the insurer and the insured (Hoy 1982; Crocker and Snow 1985, 1986; Dionne and Rothschild 2014). Full efficiency in risk classification should separate individual risks and generate different actuarial insurance premiums that reflect these risks. This means that there should not be any residual asymmetric information between the insurer and the insured inside the risk classes when risk classification is costless and not regulated. In the presence of proportional transaction costs, partial insurance should be the optimal contract, but there should be no correlation between insurance coverage and individual risk. However, in the real world of insurance contracting, there may be numerous constraints that limit efficiency in risk classification. Incentive contracting to obtain more information on individual risk in different risk classes thus becomes important, and the empirical question is how efficiently this mechanism reduces asymmetric information in insurers' portfolios.

Cohen and Siegelman (2010) present a survey of empirical studies of adverse selection in insurance markets. They argued that the coverage-risk correlation is particular to each market. Accordingly, the presence of a significant coverage-risk correlation has different meanings in different markets and even in different risk pools in a given market, depending on the type of insured service, the participants' characteristics, institutional factors, and regulation. This means that when testing for the presence of residual asymmetric information, one must also control for these factors. What

characteristics and factors explain the absence of coverage-risk correlation in automobile insurance markets? Some studies using the conditional correlation approach on cross-sectional data found evidence of asymmetric information (Dahlby 1983, 1992; Devlin 2002; Puelz and Snow 1994; Richaudeau 1999; Cohen 2005), while others did not (Chiappori and Salanié 2000; Dionne et al. 2001; Saito 2006).[5] One major criticism of the conditional correlation approach with cross-sectional data is that it does not allow one to conclude on causality.

The first empirical question in insurance markets can be summarized as follows: Is there any residual causal effect between chosen insurance coverage and risk within risk classes? The second question is how to identify which information problem remains when the first test rejects the null hypothesis that there is no residual information problem. This step is important for the insurer because it must be implemented to improve resource allocation. A deductible efficiently reduces *ex ante* moral hazard but not necessarily *ex post* moral hazard because, often, the accident has already occurred when the action is taken. A high deductible value may even have an adverse effect and encourage accident cost building. As is well known in the empirical literature, a positive correlation between insurance coverage and risk is a necessary condition for the presence of asymmetric residual information, but it does not shed light on the nature of the information problem. The third question is how improving contracts can reduce the negative impact of asymmetric information on resource allocation. These resource allocation objectives must take into account other issues, such as risk aversion, fairness, and the accessibility of insurance. This last issue is particularly important in many insurance markets. A decrease in insurance coverage may reduce *ex ante* moral hazard because it exposes the insured person to risk, but it also significantly reduces accessibility to insurance protection for risky and poor people, who are not always responsible for their risk type and financial conditions.

Econometricians analyze two types of information when studying insurers' data (Boyer and Dionne 1989; Dionne and Vanasse 1989, 1992; Chiappori 1994; Puelz and Snow 1994; Gouriéroux 1999; Richaudeau 1999; Dionne and Ghali 2005; Saito 2006; Dionne et al. 2006; Lee 2013; Spindler 2014; Dionne and Harrington 2014; Zavadil 2015; Rowell et al., 2017a, b). The first type contains variables that are observable by both parties to the insurance contract. The risk classification variables are one example. Econometricians/insurers combine these variables to create risk classes when estimating accident distributions. The observed variables can be used to make estimates conditional on the risk classes or within the risk classes. The second type of information is related to what is not observed by the insurer or the econometrician during contract duration and at contract renegotiations but can explain the insured's choice of contracts or safety in driving. If we limit our interpretation to asymmetric information (moral hazard, adverse selection, or both), we can test the conditional residual presence of asymmetric information in an insurer's portfolio or look for a correlation between the contract coverage and the realization of the risk variable during a contract period. Two parametric tests have been proposed in the literature (Chiappori and Salanié 2000; Dionne et al. 2001; see Chiappori and Salanié 2013, for a detailed analysis). One parametric test (Dionne et al. 2001) estimates the following relationship:

$$y_i = g\left(\alpha + \beta X_i + \gamma d_i + \delta E\left(d_i | X_i\right)\right) + \varepsilon_i, \tag{40}$$

where y_i is the contract choice by individual i (level of deductible, for example), X_i is a vector of control variables such as the observable characteristics used in risk classification and control variables for risk aversion, β is a vector of parameters to be estimated, d_i is the realization of the random variable observed at the end of the contract period (accident or not, for example), $E(d_i | X_i)$ is

[5] Other recent studies include those by Kim et al. (2009), Olivella and Vera-Hernández (2013), Dardanoni et al. (2018), Geyer et al. (2020), and Rowell and Zweifel (2024).

the conditional expected value of the random variable obtained from the estimation of the accident distribution, and ε_i is the residual of the regression. A positive sign is usually anticipated for the coefficient d_i when residual asymmetric information remains (higher coverage is related to more accidents or higher risk). The seminal theories of Rothschild and Stiglitz (1976) and Wilson (1977) strongly predict that such a correlation should be observed in the data in the presence of adverse selection, while Holmström (1979) and Shavell (1979) strongly predict that the correlation is due to moral hazard. Note that the dependent variable in the above regression can be the risk variable d_i while coverage y_i is an independent variable. The presence of the variable d_i is not necessarily exogenous in Eq. (40). It is better to instrument this variable in order to obtain a causality result, as shown in Dionne et al. (2009) and Rowell et al. (2017a, b).

The presence of $E(d_i|\, X_i)$ is necessary to control for specification errors (missing variables) or for potential nonlinearity not modeled in the equation. Without this control, the coefficient (γ) of d_i can be significant for reasons other than the presence of residual asymmetric information in the risk classes.

If the coefficient of d_i is not significant, one can reject the presence of residual asymmetric information in the risk classes when all other factors are well controlled. This does not mean that there was no asymmetric information in this market; rather, the insurers' risk classification system efficiently eliminated asymmetric information and that there was no *residual* asymmetric information within the risk classes. In other words, when risk classification is done properly, it is not necessary to choose the insurance contract form within the risk classes to reduce asymmetric information. Therefore, the observed presence of a deductible must be explained by other factors in the insurance market, such as transaction costs.

An equivalent parametric model was proposed by Chiappori and Salanié (2000). Here, two equations are estimated simultaneously, one for the contract choice and the other for the accident distribution. An example is the bivariate probit model:

$$y_i = f\,(x_i, \beta) + \varepsilon_i \tag{41}$$

$$d_i = g\,(x_i, \beta) + \eta_i. \tag{42}$$

The test consists in verifying whether there is a dependence between the residuals of the two equations. An absence of conditional correlation is interpreted as an absence of residual asymmetric information in the data. The authors performed an additional nonparametric test that is independent of the functional forms of the above model. It is based on a chi-square test of independence. However, this test seems to be limited to discrete variables, contrary to the two parametric tests presented above (see Su and Spindler 2013, for a longer discussion).

Many extensions of these models have been presented in the literature. Chiappori and Salanié (2013) specified conditions to obtain robustness of the test when insured may have different degrees of risk aversion. They showed that if insurers maximize profits in competitive markets, the results of the above test are robust to heterogeneity in preferences. Such robustness is less evident in noncompetitive insurance markets.[6]

Fang et al. (2008) did not reject asymmetric information in the medical insurance market, nor did they find evidence of adverse selection. Their results are consistent with multidimensional private information along with advantageous selection (De Meza and Webb 2001). They obtained a *negative* correlation between risk and insurance coverage. Risk aversion is not a source of advantageous

[6] Maliar and Salanié (2022) showed that adding deep learning to the test model does not affect the correlation results obtained by Chiappori and Salanié (2000).

selection in their data. The significant sources are income, education, longevity expectations, financial planning horizons, and, most importantly, cognitive ability (see also Finkelstein and McGarry 2006, on this issue).

To separate moral hazard from adverse selection, econometricians need a supplementary step. An additional market relationship can be estimated to look for adverse selection (conditional on the fact that the null hypothesis of no asymmetric information was rejected), as Dionne et al. (2009) did for auctions where moral hazard is not a significant problem.

In insurance markets, dynamic data are often available. Time adds an additional degree of freedom to test for asymmetric information (Dionne and Lasserre 1985, 1987; Dionne and Vanasse 1989, 1992; D'Arcy and Doherty 1990; Dionne and Doherty 1994; Chiappori et al. 1994; Hendel and Lizzeri 2003). This information can be used in many insurance markets where past experience information is available and when usage is possible. For ethical reasons, this information is not often utilized on an individual basis in health insurance and for bodily injury insurance in many countries. The experience rating works at two levels in insurance. Past accidents implicitly reflect unobservable characteristics of the insured (adverse selection) and introduce additional incentives for prevention (moral hazard). Experience ratings can therefore directly mitigate problems of adverse selection and moral hazard, which often hinder risk allocation in the insurance market.

The experience rating not only provides additional information on risk but also may play an important role in the dynamic relationship between policyholders' insurance claims and contract choice. The theoretical literature on repeated insurance contracting over time clearly indicates that these features may help overcome problems of moral hazard when risks known to the policyholder (endogenous) are unobservable by the insurer (moral hazard, Parra and Winter 2025) or when exogenous characteristics are unobservable (adverse selection, Dionne et al. 2025). Contract choice is influenced by the evolution of the premium, which is closely linked to the insured's risk or past experience. Because increased insurance coverage tends to lower the expected cost of accidents for the insured, incentives for safe behavior are weakened for all risks. Under the experience rating, the subsequent rise in accidents increases the marginal costs of future accidents when the experience rating is taken into account. An experience rating may therefore offset the disincentive effect created by single-period insurance coverage.

Abbring et al. (2003) applied a multiperiod incentive mechanism by focusing on the dynamics of claims but not on the dynamics of contract choice (because of data limitations). Proposing specific assumptions about the wealth effects of accidents to policyholders who differ only in their claim records (thus their experience rating), their model predicts that subjects with the worst claims records should try harder to increase safety and thereby, ceteris paribus, file fewer claims in the future. However, their data do not support the presence of moral hazard. Dionne et al. (2011) extended their model and did not reject the presence of moral hazard by using a different dataset. The potential presence of adverse selection in their data was not a real problem because all drivers must be insured for bodily injuries.

Dionne et al. (2013a) showed that failure to detect residual asymmetric information, and more specifically, moral hazard and adverse selection in insurance data, is due to the failure of previous econometric approaches to model the dynamic relationship between contract choice and claims adequately and simultaneously when the experience rating is considered. Intuitively, because there are at least two potential information problems in the data, an additional relationship with the correlation between risk and insurance coverage is necessary to test for causality between risk and insurance coverage. Using a unique longitudinal survey of policyholders from France, they propose a methodology to disentangle the historical pathways in claims and premiums. They explained how

causality tests can be used to differentiate moral hazard from asymmetric learning (and eventually adverse selection). They did not reject moral hazard for a given group of policyholders and did not reject asymmetric learning for younger drivers. The empirical methodologies of Dionne et al. (2011) and Dionne et al. (2013a) are reviewed in detail below, as is the recent study of Dionne and Liu (2021), which applied a difference-in-differences (DID) approach to separate moral hazard from adverse selection and evaluated the introduction of a new regulation on insurance pricing in China.

6 Causal Tests for Moral Hazard in the Automobile Insurance Market

6.1 Moral Hazard as a Function of Accumulated Demerit Points

We now analyze moral hazard as a function of demerit points. Because no-fault environments are common in North America, traffic violations are events likely to be used in insurance experience rating schemes. In Quebec, the public insurer in charge of the compensation of bodily injuries uses an experience rating scheme based on demerit points.[7] The same public enterprise is also in charge of the point-record license system. Dionne et al. (2011) show that the new insurance pricing scheme based on demerit points introduced in 1992 reduced the number of traffic violations by 15%. They also determined that there is residual *ex ante* moral hazard in road safety management. The discussion below focuses on the methodology they developed for obtaining these results.

The methodology extends the empirical model of Abbring et al. (2003). Over time, a driver's observed demerit points inform two effects: an unobserved heterogeneity effect and an incentive effect. Drivers with more demerit points accumulated during a period are riskier with respect to hidden features in risk distributions. Hence, unobserved heterogeneity is a form of risk reassessment in the sense that those who accumulate demerit points represent higher risks over time. This time effect is in the opposite direction of the incentive effect. For the incentive effects, accumulating demerit points should increase the incentive for safe driving to reduce the probability of receiving a higher penalty, such as losing the driving license for a certain period.

The model proposed by Dionne et al. (2011) tests for an increasing link between traffic violations and the number of accumulated demerit points over time. Rejecting the positive link is necessary to verify the presence of moral hazard. They estimated the following hazard function (Cox 1972):

$$\lambda_{it} = \exp\left(x_{it}\beta\right) + g\left(\text{adp}_{it}\right) \times h\left(c_{it}\right) \tag{43}$$

where λ_{it} is the hazard function for driver i at date t, x_{it} is a vector of control variables, β represents the corresponding coefficients, adp_{it} is the number of demerit points accumulated over the 2 previous years at time t, and c_{it} is the contract time at date t.

In the absence of moral hazard, g should increase because of unobserved heterogeneity. They found that g decreased when drivers accumulated more than seven demerit points. This means that beyond seven demerit points, drivers become safer because they do not want to lose their driver's license when they have accumulated 15 demerit points. This is evidence of the presence of moral hazard in the data: these drivers were negligent when the accumulated record was below seven demerit points (see Pinquet 2025, and Dionne et al. 2013b, for more details).

[7] On point-record driver's licenses, see the study by Bourgeon and Picard (2007).

6.2 Separating Moral Hazard from Learning and Adverse Selection with Dynamic Data

To separate learning leading to adverse selection (asymmetric learning) from moral hazard, Dionne et al. (2013a) considered the case where information on contracts and accidents is available over time in the form of panel data. They exploited dynamics in accidents and insurance coverage while controlling for dynamic selection due to unobserved heterogeneity. They constructed two additional tests based on changes in insurance coverage. Coupled with the negative occurrence tests of Abbring et al. (2003) and Dionne et al. (2011), their two additional tests allowed them to separate moral hazard from asymmetric learning (which should become adverse selection in the long run).

They analyzed the identification of asymmetric learning and moral hazard within the context of a tractable structural dynamic insurance model. From the solution of their theoretical model, they simulated a panel of drivers behaving under different information regimes or data-generating processes (with or without both moral hazard and asymmetric learning). They validated their empirical tests on simulated data generated from these different information regimes. They then applied these tests to longitudinal data on accidents, contract choice, and experience rating for the 1995–1997 period in France (Dionne 2001). They found no evidence of information problems among experienced drivers (with more than 15 years of experience). For drivers with less than 15 years of experience, they found strong evidence of moral hazard but little evidence of asymmetric learning. They obtained evidence of asymmetric learning, despite the small sample size, when they focused on drivers with less than 5 years of experience. To obtain these results, they estimated the following model.

They considered a joint parametric model for the probabilities of accidents and contract choice. Each equation corresponds to a dynamic binary choice model with predetermined regressors and an error component structure. The error component structure is important given the likelihood of serial correlation in contract and accident outcomes. They used the solution proposed by Wooldridge (2010) to take the potential left censoring effect into account. More specifically, the process for accidents is specified as

$$n_{it} = I\left(x_{it}\beta_n + \phi_{nd}d_{it-1} + \phi_{nn}n_{it-1} + \phi_{nb}b_{it} + \varepsilon_{n,it} > 0\right) \quad i = 1, \ldots, N, t = 1, \ldots, T \tag{44}$$

where $\varepsilon_{n,\,it}$ has an error component structure $\varepsilon_{n,\,it} = \alpha_{ni} + v_{n,\,it}$, n_{it} is a binary variable for the number of accidents of individual i at time t, $d_{it\,-\,1}$ is the individual's contract choice in period $t - 1$, $n_{it\,-\,1}$ is the number of accidents in period $t - 1$, and b_{it} is the bonus-malus score at period t. The presence of moral hazard is confirmed by a positive sign for ϕ_{nd} (more insurance coverage-more accidents) and a negative sign for ϕ_{nb} (a higher malus creates more incentives for safe driving, similar to the test presented in the previous section with accumulated demerit points.) Here, a high malus means an accumulation of accidents over the previous periods. Dionne et al. (2013a) specify a similar equation for contract choice:

$$d_{it} = I\left(x_{it}\beta_d + \phi_{dd}d_{it-1} + \phi_{dn}n_{it-1} + \phi_{db}b_{it} + \varepsilon_{d,it} > 0\right) \quad i = 1, \ldots, N, t = 1, \ldots, T \tag{45}$$

where again $\varepsilon_{d,\,it} = \alpha_{di} + v_{d,\,it}$. The asymmetric learning test is a test of whether an accident (not necessarily a claim) in the last period, conditional on the bonus-malus, leads to an increase in coverage during this period. In this model, the insured observes all accidents while the insurer observes only the

claims. Drivers thus learn that they are riskier than anticipated and increase their insurance coverage accordingly. It is a test of whether $\phi_{dn} > 0$ or not.

6.3 Separating Moral Hazard from Adverse Selection with a DID Model

We now consider the DID approach to separate moral hazard from adverse selection by reviewing the study of Dionne and Liu (2021) (see also Abay 2018). In this study, the authors analyzed the effect of public reform of the Chinese bonus-malus system on road safety and showed how the observed variation in accident rates could be attributed to a reduction in moral hazard. They had access to panel data from two cities, one affected by the reform (treatment group) and one not affected (control group).

DID Matching Estimators

One potential concern is that the control vehicles may differ from the treated vehicles along most of the observable dimensions. The authors use the propensity score matching (PSM) methodology to balance the two groups in terms of observables. The goal of the reform was to improve road safety (fewer accidents) by introducing more incentives for safe driving in the pricing of auto insurance. They first estimated the propensity score or the likelihood of being treated by using a standard logit model. Their results affirmed that they obtained covariate balances between the matched treated and control groups for most of the dimensions they used. Potential adverse selection was thus eliminated. They then estimated DID matching estimators and determined that the number of accidents in the treated city dropped after the two-stage reform.

DID with Parallel Trends

The DID approach allows analysts to test the dynamics of the treatment effect induced by each of the two stages of the reform on road safety. A relative decline in claim frequency was expected in the treated group compared with the control group after both stages of the reform. Equation (46) presents the basic regression equation:

$$\text{Claims}_{it} = \alpha \quad + \sum_{s=1}^{2} \beta_s \text{treated}_i \times \text{post-event}_{st} + \kappa_i + \eta_t + \varepsilon_{it} \tag{46}$$

where $s = 1$ and $s = 2$ denote the first-stage and second-stage reforms, respectively. β_1 and β_2 are the main parameters of interest: they evaluate the differential effects of the two-stage reform at each stage across the treated and control groups. Claims_{it} measures the claim frequency for vehicle i in year t. α is a constant term. $\text{treated}_i = 1$ when the policy is issued in the treated city, and 0 otherwise. $\text{treated}_i = 1$ and post - $\text{event}_{2t} = 1$ when the due date of the policy is after the first-stage and second-stage reform, respectively, and 0 otherwise. κ_i is the vehicle fixed effects, the proxy for driver fixed effects, which control for vehicle/driver-level time-invariant heterogeneity. η_t is the year fixed effects, which account for common aggregate shocks.

In addition, to verify the parallel trends in claim frequency across these two groups before the reform (no difference in claims variations), Dionne and Liu (2021) estimated the time-varying effects of the reform by year using the following distributed lag model in Eq. (47):

$$\text{Claims}_{it} = \alpha + \sum_{y=-1}^{2} \beta_y \text{treated}_i \times \text{year}_{yt} + \kappa_i + \eta_t + \varepsilon_{it} \tag{47}$$

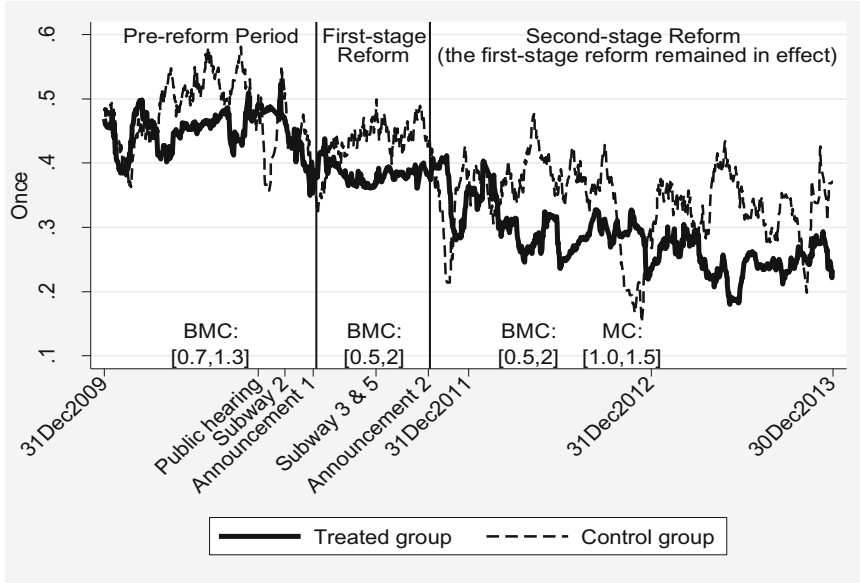


Fig. 1 Evolution of two-stage reform and claim frequency by group and period. (Source: Dionne and Liu (2021))

where year_{yt} represents a set of 4 yearly dummies from 1 year before the reform until 2 years after the first-stage reform year. β_y measures the differential trend in claim frequency across the treated and control groups during the pre-reform period. The analysis was repeated with semester and quarterly observations. In all the cases, the authors did not reject parallel trends between claims in the two cities before the reform, a necessary step to consolidate the DID methodology. Their estimation of Eq. (46) confirmed the presence of moral hazard in the data. Adverse selection was ruled out by showing that there were no significant differences between drivers' and cars' characteristics between the two cities before and after the reform.

Figure 1 shows the evolution of the accident rates during the period of analysis. We observe a clear difference between the two frequencies after the two reforms. The new pricing scheme of automobile insurance reduced the accident rates in the treated city, which is attributed to a reduction in ex ante moral hazard. Figure 2 does not reject the presence of parallel trends between accident distributions before the first reform in the two pre-reform periods $(-2, -1)$ and confirms the differences in accident rates after the reform (starting in period 3).

7 Causality Analysis Based on the Generalized Method of Moments

7.1 Econometric Model

Desjardins et al. (2022) studied the causal reciprocal relationship between liquidity creation and reinsurance demand. One objective of their research was to examine the reciprocal relationship between insurers' liquidity creation and their demand for reinsurance. One of the most efficient models that can be used to evaluate this reciprocal relation is the structural equation model (Low and Meghir 2017). To evaluate the reciprocal relationship between reinsurance demand and liquidity creation with a structural equation model, Desjardins et al. (2022) specified a dynamic data panel incorporating unobserved heterogeneity. In this model, the lagged value of liquidity creation was

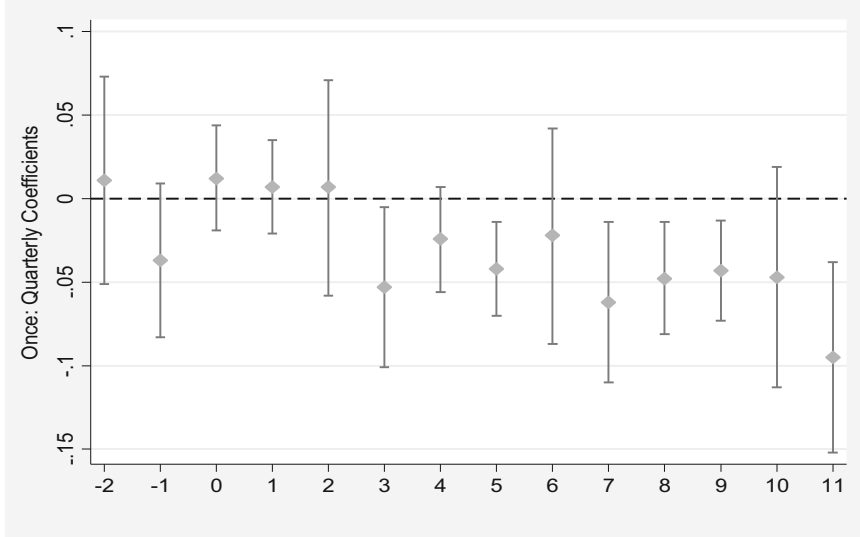


Fig. 2 Quarterly estimates of the effects of insurance incentives on road safety with two pre-reform periods. (Source: Dionne and Liu (2021))

added to the equation of reinsurance demand as one of the key explanatory variables, and the lagged reinsurance demand variable was added to the equation of liquidity creation as an explanatory variable. In this specification of the model the parameters associated with lagged variables reflect a causal link that takes some time to become effective. In fact, most of the insurers' strategic decisions, such as investments (liquidity creation) and reinsurance demand (risk management), are generally made by the board of directors once a year and may take several months to materialize. The decisions are very unlikely to produce causal effects during the same year. Therefore, the authors focused on yearly lagged values of the key variables to analyze their causal relationships. Moreover, this specification of the model fits well with Granger causality.

Desjardins et al. (2022) analyzed the causality between reinsurance demand and liquidity creation by applying a robust GMM procedure to estimate their parameters. More precisely, they used the regularized GMM procedure proposed by Carrasco and Nayihouba (2024) for dynamic panel data. They estimated Eq. (48), where $y_{i,t}$ is for reinsurance demand and $x_{i,t}$ is for the liquidity creation ratio:

$$\begin{cases} y_{i,t} = \beta_1 x_{i,t-1} + \beta_2 y_{i,t-1} + \Delta_1 Z_{i,t} + \kappa_i + \varepsilon_{i,t} \\ x_{i,t} = \beta_3 x_{i,t-1} + \beta_4 y_{i,t-1} + \Delta_2 \Xi_{it} + \theta_i + \nu_{i,t} \end{cases} \tag{48}$$

In Eq. (48), the liquidity creation ratio at time t is regressed on the control variables at time t, and the reinsurance demand at time t is regressed on the control variables at time t. Each equation of the model is in fact a dynamic panel data relationship with a lagged dependent variable, a lagged endogenous variable, individual fixed effects (κ_i, θ_i) and vectors of covariates $(Z_{it}, \Xi_{i,t})$. The terms $\varepsilon_{i,t}$ and $\nu_{i,t}$ are error terms with zero mean and positive variance for $i = 1 \ldots N$ and $t = 1 \ldots T$,, where N is the number of firms and T is the number of periods.

The estimation of Eq. (48) is done equation by equation. Using the above specification of the model, we may face some severe endogeneity problems that need to be solved in the estimation process. The first endogeneity problem is due to the presence of individual fixed effects in the model, which creates a correlation between the error term and one of the explanatory variables, namely, the

lagged value of the dependent variable. This implies that the lagged value of the dependent variable should be treated as an endogenous variable in the estimation process. Moreover, this problem is amplified by the fact that the lagged value of liquidity creation in the equation of reinsurance demand and the lagged value of reinsurance demand in the equation of liquidity creation are also endogenous variables, as was shown empirically by the authors using Hausman's (1978) test procedure. Therefore, the standard OLS method with fixed effects may yield bias estimates. To tackle these two endogeneity problems and avoid the problem involved in finding valid instruments for the two-stage least square (2SLS) regression method, the GMM model is employed to estimate parameters in (48), with lagged levels of the set of explanatory variables as instruments. This model has an important feature: if a variable at a certain period can be used as an instrument, then all the past realizations of that variable can also be used as instruments. Therefore, the number of moment conditions can be very large, even if the time duration of panel T is finite.

The presence of this large set of moment conditions may create a variance bias that is also referred to as the many instruments problem. Moreover, the lagged levels of the dependent variable, which appear in the explanatory variables, can become weak instruments when the autoregressive parameter is close to unity (Blundell and Bond 1998). As a solution to these problems (many instruments and weak instruments), one can add some regularization methods to the standard GMM method, as Carrasco and Nayihouba (2024) did, to evaluate the relationship between liquidity creation and reinsurance demand.

7.2 Regularization Procedures for Estimation

Several methods have been proposed in the context of cross-sectional data models to deal with these problems of instruments. To manage this problem in a dynamic setting, Okui (2009) recommends choosing the optimal number of moment conditions to minimize the mean square error of the estimation in order to improve the finite sample properties. However, the finite sample problem is not completely solved since there may be a large bias in estimated cross-lagged parameters when the autoregressive coefficient in the dynamic panel is close to unity. Carrasco and Nayihouba (2024) proposed a more general method based on different ways of inverting the covariance matrix of instruments. They showed that this method improves the properties of the GMM estimation even if the autoregressive coefficient is close to unity. To analyze the causality relationships in (48), Desjardins et al. (2022) focused on two of the regularization procedures proposed by Carrasco and Nayihouba (2024) in the context of their dynamic panel data.

When the number of moment conditions exceeds the number of unknown parameters to be estimated by the GMM, the model validity must be confirmed by testing the overidentifying restriction before making any inference regarding the resulting estimation. A common test for this purpose is the J-test proposed by Sargan (1958) and Hansen (1982). To test if their models were well specified, Desjardins et al. (2022) applied the modified version of the J-test to the context of dynamic panel data models (Arellano and Bond 1991). They found a causal relationship in the two equations, meaning that insurers who invest more in the economy buy more reinsurance for protection, and those who buy more reinsurance invest more in the economy because they are well protected.

8 Conclusion

We explored the empirical measurement of asymmetric information on resource allocation. We studied causality effects with the applied DID methodology (with parallel trends or propensity score modeling) and with instrumental variables (with essential heterogeneity or not). Other tests were applied with dynamic data. We did not cover the literature with dynamically treated effects because our applications did not contain such effects.

Three information problems drew our attention: moral hazard, asymmetric learning, and adverse selection. One conclusion that seems to be accepted by many authors is that information problems may create distortions in the economy, in contrast with a situation of full and perfect information. Effective mechanisms have been established to reduce these distortions and eliminate residual information effects at the margin. In this new version of our survey, we have emphasized the role of causality in identifying different information problems. We have showed that causality, in dynamic models, can be used to separate unobserved heterogeneity from moral hazard and to apply tests to separate moral hazard from adverse selection and asymmetric learning. We have also showed how instrumental variables and DID methodologies permit to identify moral hazard in auction and insurance applications and to identify the effect of a new regulation in the CDS market. Finally, we have used the GMM methodology to study the dynamic links between reinsurance demand and liquidity creation.

We have studied the causal effect of risk management on firm value and risk. We have showed that when appropriate methodology is applied (instrumental variable with essential heterogeneity), risk management is a causal source of improvement in the value and risk of nonfinancial firms.

Acknowledgements The author would like to thank SSHRC-Canada for financial support. He would also like to acknowledge the researchers who helped him develop several ideas on the subject over the years: P.A. Chiappori, D. Desjardins, A.T. Fenou, R. Gagné, N. Koné, Y. Liu, P.C. Michaud, M. Mnasri, J. Pinquet, B. Salanié, and C. Vanasse. He thanks Claire Boisvert and Karen Sherman for their collaboration in the preparation of the chapter.

References

Abadie A (2003) Semiparametric instrumental variable estimation of treatment response model. J Econ 113:231–263

Abadie A (2021) Using synthetic controls: feasibility, data requirements, and methodological aspects. J Econ Lit 59(2):391–425

Abadie A, Gardeazabal J (2003) The economic costs of conflict: a case study of the Basque country. Am Econ Rev 93(1):113–132

Abadie A, L'Hour J (2021) A penalized synthetic control estimator for disaggregated data. J Am Stat Assoc 116(536):1817–1834

Abadie A, Diamond A, Hainmueller J (2010) Synchetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc 105(490):493–505

Abadie A, Diamond A, Hainmueller J (2015) Comparative politics and the synthetic control method. Am J Polit Sci 59(2):495–510

Abay K (2018) How effective are non-monetary instruments for safe driving? Panel data evidence on the effect of the demerit point system in Denmark. Scand J Econ 120(3):894–924

Abbring JH, Chiappori PA, Pinquet J (2003) Moral hazard and dynamic insurance data. J Eur Econ Assoc 1:767–820

Adam T, Fernando CS (2006) Hedging, speculation, and shareholder value. J Financ Econ 81:283–309

Akari MA, Ben-Abdallah R, Breton M, Dionne G (2021) The impact of central clearing on the market for single-name credit default swaps. N Am J Econ Financ 56:101346

Akerlof GA (1970) The market for 'lemons': quality uncertainty and the market mechanism. Q J Econ 84:488–500

Allayannis G, Weston JP (2001) The use of foreign currency derivatives and firm market value. Rev Financ Stud 14:243–276

Angrist JD (1990) Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. Am Econ Rev 80(3):313–336

Angrist JD (2022) Empirical strategies in economics: illuminating the path from cause to effect. Econometrica 90(6):2509–2530
Angrist JD, Hahn J (2004) When to control for covariates? Panel asymptotics for estimates of treatment effects. Rev Econ Stat 86(1):58–72
Angrist JD, Imbens GW (1995) Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J Am Stat Assoc 90(430):431–442
Angrist JD, Krueger A (1991) Does compulsory schooling affect schooling and earnings? Q J Econ CVI:979–1014
Angrist JD, Graddy K, Imbens GW (2000) The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Rev Econ Stud 67(3):499–527
Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev Econ Stud 58(2):277–297
Arkhangelsky D, Athey S, Hirshberg DA, Imbens GW, Wager S (2021) Synthetic difference-in-differences. Am Econ Rev 111(12):4088–4118
Arrow K (1963) Uncertainty and the welfare economics of medical care. Am Econ Rev 53:941–973
Ashenfelter O, Card D (1985) Using the longitudinal structure of earnings to estimate the effect of training programs. Rev Econ Stat 67(4):648–660
Ashenfelter O, Heckman J (1974) The estimation of income and substitution effects in a model of family labor supply. Econometrica 42(1):73–85
Ashenfelter O, Krueger A (1994) Estimates of the economic return to schooling from a new sample of twins. Am Econ Rev 84:1157–1173
Bartram SM, Brown GW, Conrad J (2011) The effects of derivatives on firm risk and value. J Financ Quant Anal 46:967–999
Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Q J Econ 119(1):249–275
Bloom HS (1984) Accounting for no-shows in experimental evaluation design. Eval Rev 8:225–246
Blundell R, Bond S (1998) Initial conditions and moment restrictions in dynamic panel data models. J Econ 87:115–143
Blundell RW, Smith RJ (1989) Estimation in a class of simultaneous equation limited dependent variable models. Rev Econ Stud 56(1):37–47
Bourgeon JM, Picard P (2007) Point-record driver's license and road safety: an economic approach. J Public Econ 91:235–258
Boyer M, Dionne G (1989) An empirical analysis of moral hazard and experience rating. Rev Econ Stat LXXXI(1):128–134
Campello M, Lin C, Ma Y, Zou H (2011) The real and financial implications of corporate hedging. J Financ 66:1615–1647
Card D (1990) The impact of the Mariel boatlift on the Miami labor market. Ind Labor Relat 43:245–257
Card D (2022) Design-based research in empirical economics. Am Econ Rev 112:1773–1781
Card D, Krueger A (1994) Minimum wages and employment: case study of the fast-food industry in New Jersey and Pennsylvania. Am Econ Rev 84:772–793
Carrasco M, Nayihouba A (2024) Regularized estimation of dynamic panel models. Econ Theory 40(2):360–418. First view article
Carter DA, Rogers D, Simkins BJ (2006) Does hedging affect firm value? Evidence from the US airline industry. Financ Manag 35:53–87
Chiappori PA (1994) Théorie des contrats et économétrie de l'assurance: quelques pistes de recherche, Mimeo, Chaire d'économie et d'économétrie de l'assurance, EHESS – ENSAE, DELTA
Chiappori PA (2000) Econometric models of insurance under asymmetric information. In: Dionne G (ed) Handbook of insurance. Springer, Dordrecht, pp 365–393
Chiappori PA, Salanié B (2000) Testing for asymmetric information in insurance markets. J Polit Econ 108:56–78
Chiappori PA, Salanié B (2013) Asymmetric information in insurance markets: predictions and tests. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, New York, pp 397–422
Chiappori PA, Macho I, Rey P, Salanié B (1994) Repeated moral hazard: the role of memory, commitment, and the access to credit markets. Eur Econ Rev 38:1527–1553
Cohen A (2005) Asymmetric information and learning in the automobile insurance market. Rev Econ Stat 87:197–207
Cohen A, Einab L (2007) Estimating risk preferences from deductible choices. Am Econ Rev 97(3):745–788
Cohen A, Siegelman P (2010) Testing for adverse selection in insurance markets. J Risk Insur 77(1):39–84
Cox DR (1972) Regression models and life tables. J R Stat Soc Ser B 34:187–220
Crocker K, Snow A (1985) The efficiency of competitive equilibria in insurance markets with asymmetric information. J Public Econ 26:207–219
Crocker K, Snow A (1986) The efficiency effect of categorical discrimination in the insurance industry. J Polit Econ 94:321–344

Crocker K, Snow A, Rothschild (2025) The theory of risk classification. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer, New York
Currie J, Kleven H, Zwiers E (2020) Technology and big data are changing economics: mining text to track methods. AEA Pap Proc 110:42–48
D'Arcy SP, Doherty NA (1990) Adverse selection, private information, and lowballing in insurance markets. J Bus 63(2):145–164
Dahlby BG (1983) Adverse selection and statistical discrimination. An analysis of Canadian automobile insurance. J Public Econ 20:121–130. Reprinted in: Dionne G, Harrington S (eds) (1992) Foundations of insurance economics – readings in economics and finance. Kluwer Academic Publishers
Dahlby BG (1992) Testing for assymmetric information in Canadian automobile insurance. In: Dionne G (ed) Contributions to insurance economics. Kluwer Academic Publishers
Dardanoni V, Forcina A, Li Donni P (2018) Testing for asymmetric information in insurance markets: a multivariate ordered regression approach. J Risk Insur 85(1):107–125
De Meza D, Webb DC (2001) Advantageous selection in insurance markets. RAND J Econ 32(2):249–262
Dehejia RH, Wahba S (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc 94:1053–1062
Desjardins D, Dionne G, Koné N (2022) Reinsurance demand and liquidity creation: a search for bicausality. J Empir Financ 66:137–154
Devlin RA (2002) Determinants of no-fault insurance measures. J Risk Insur 69(4):555–576
Dionne G (2001) Insurance regulation in other industrial countries. In: Cummins JD (ed) Deregulating property–liability insurance. AEI–Brookings, Washington, pp 362–390
Dionne G (2019) Corporate risk management: theories and applications. Wiley
Dionne G, Doherty N (1994) Adverse selection, commitment and renegotiation: extension to and evidence from insurance markets. J Pol Econ 102(2):209–235. Reproduced in: Niehaus GR (ed) Insurance and risk management, vol. I, Economics of insurance markets. Elgar Reference Collection, USA
Dionne G, Gagné R (2001) Deductible contracts against fraudulent claims: evidence from automobile insurance. Rev Econ Stat 83(2):290–301
Dionne G, Garand M (2003) Risk management determinants affecting firms' values in the gold mining industry: new empirical results. Econ Lett 79:43–52
Dionne G, Ghali O (2005) The (1992) bonus-malus system in Tunisia: an empirical evaluation. J Risk Insur 72:609–633
Dionne G, Harrington SE (2014) Insurance and insurance markets. In: Viscusi WK, Machina M (eds) Handbook of the economics of risk and uncertainty, 1st edn. North Holland, Amsterdam, pp 203–261
Dionne G, Lasserre P (1985) Adverse selection, repeated insurance contracts and announcement strategy. Rev Econ Stud 70(4):719–724
Dionne G, Lasserre P (1987) Adverse selection and finite-horizon insurance contracts. Eur Econ Rev 31(4):843–862
Dionne G, Liu Y (2021) Effects of insurance incentives on road safety: evidence from a natural experiment in China. Scand J Econ 123(2):453–477
Dionne G, Mnasri M (2018) Real implications of corporate risk management: review of main results and new evidence from a different methodology. Actual Econ 94(4):407–452
Dionne G, Ouederni K (2011) Corporate risk management and dividend signaling theory. Financ Res Lett 8:188–195
Dionne G, Rothschild C (2014) Economic effects of risk classification bans. Geneva Risk Insur Rev 39:184–221
Dionne G, Triki T (2013) On risk management determinants: what really matters? Eur J Financ 19:145–164
Dionne G, Vanasse C (1989) A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. Astin Bull 19(2):199–212
Dionne G, Vanasse C (1992) Automobile insurance ratemaking in the presence of asymmetrical information. J Appl Econ 7:149–165
Dionne G, Gouriéroux C, Vanasse C (2001) Testing for evidence of adverse selection in the automobile insurance market: a comment. J Polit Econ 109:444–453
Dionne G, Gouriéroux C, Vanasse C (2006) The informational content of household decisions with applications to insurance under asymmetric information. In: Chiappori PA, Gollier C (eds) Competitive failures in insurance markets. MIT Press Book, pp 159–184
Dionne G, St-Amour P, Vencatachellum D (2009) Asymmetric information and adverse selection in Mauritian slave auctions. Rev Econ Stud 76:1269–1295
Dionne G, Pinquet J, Maurice M, Vanasse C (2011) Incentive mechanisms for safe driving: a comparative analysis with dynamic data. Rev Econ Stat 93:218–227
Dionne G, Michaud PC, Dahchour M (2013a) Separating moral hazard from adverse selection and learning in automobile insurance: longitudinal evidence from France. J Eur Econ Assoc 11:897–917
Dionne G, Michaud PC, Pinquet J (2013b) A review of recent theoretical and empirical analyses of asymmetric information in road safety and automobile insurance. Res Transp Econ 43:85–97

Dionne G, La Haye M, Bergerès AS (2015) Does asymmetric information affect the premium in mergers and acquisitions? Can J Econ 48(3):819–852
Dionne G, Gueyie JP, Mnasri M (2018) Dynamic corporate risk management: motivations and real implications. J Bank Financ 95:97–111
Dionne G, Maalaoui Chun O, Triki T (2019) The governance of risk management: the importance of directors' independence and financial knowledge. Risk Manag Insur Rev 22:247–277
Dionne G, Fenou A, Mnasri M (2023) Insurers' M&A in the United States during the 1990–2022 period. Canada Research Chair in Risk Management, HEC Montréal, Mimeo
Dionne G, Fombaron N, Mimra W (2025) Adverse selection in insurance. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer, New York
Fang H, Keane MP, Silverman D (2008) Sources of advantageous selection: evidence from the Medigap insurance market. J Polit Econ 116(2):303–350
Finkelstein A, McGarry K (2006) Multiple dimensions of private information: evidence from the long-term care insurance market. Am Econ Rev 96:938–958
Frisch R (1930) A dynamic approach to economic theory: lectures by Ragnar Frisch at Yale University. Lectures at Yale University beginning September 1930. Mimeographed, Frisch Archives, Department of Economics, University of Oslo
Frisch R (1938) Autonomy of economic relations: statistical versus theoretical relations in economic macrodynamics. Paper given at League of Nations. Reprinted in: Hendry DF, Morgan MS (1995) The foundations of econometric analysis. Cambridge University Press
Froot KA, Scharfstein D, Stein J (1993) Risk management: coordinating corporate investment and financing policies. J Financ 48:1629–1658
Gale D, Hellwig M (1985) Incentive-compatible debt contracts: the one-period problem. Rev Econ Stud 5(4):647–663
Gay GD, Lin CM, Smith SD (2011) Corporate derivatives use and the cost of equity. J Bank Financ 35:1491–1506
Geyer A, Kremslehner D, Mürmann A (2020) Asymmetric information in automobile insurance: evidence from driving behavior. J Risk Insur 87(4):969–995
Gormley TA, Matsa DA (2011) Growing out of trouble? Corporate responses to liability risk. Rev Financ Stud 24(8):2781–2821
Gouriéroux C (1999) The econometrics of risk classification in insurance. Geneva Pap Risk Insur Theory 24:119–138
Goux D, Maurin E (2023) David Card, Nobel Prize 2021: the design-based revolution. Rev Econ Polit 133(1):25–46
Graham JR, Rogers DA (2002) Do firms hedge in response to tax incentives? J Financ 57(2):815–839
Graham JR, Smith CW (1999) Tax incentives to hedge. J Financ 54:2241–2262
Guay WR (1999) The impact of derivatives on firm risk: an empirical examination of new derivatives users. J Account Econ 26:319–351
Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11(1):1–12
Haavelmo T (1944) The probability approach in econometrics. Econometrica 12(supplement):iii–vi, 1–115
Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50(4):1029–1054
Haushalter D (2000) Financing policy, basis risk, and corporate hedging: evidence from oil and gas producers. J Financ 55:107–152
Hausman JA (1978) Specification tests in econometrics. Econometrica 46:1251–1272
Heckman JJ (1990) Varieties of selection bias. Am Econ Rev 80:313–318
Heckman JJ, Pinto R (2022) Causality in econometrics. NBER working paper 29787
Heckman JJ, Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci 96(8):4730–4734
Heckman JJ, Vytlacil EJ (2001) Policy-relevant treatment effects. Papers and proceedings of the 113th annual meeting of the American Economic Association 91(2):107–111
Heckman JJ, Vytlacil EJ (2005) Structural equations, treatment effects and econometric policy evaluation. Econometrica 73(3):669–738
Heckman JJ, Urzua S, Vytlacil EJ (2006) Understanding instrumental variables in models with essential heterogeneity. Rev Econ Stat 88:389–432
Heckman JJ, Humphries JE, Veramendi G (2016) Dynamic treatment effects. J Econ 191:276–292
Hendel I, Lizzeri A (2003) The role of commitment in dynamic contracts: evidence from life insurance. Q J Econ 118(1):299–327
Hendry DF (1980) Econometrics – alchemy or science? Economica 47(188):387–406
Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71(4):1161–1189
Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81:945–960
Holmström B (1979) Moral hazard and observability. Bell J Econ 10(1):74–91
Hoy M (1982) Categorizing risk in the insurance industry. Q J Econ 97:321–336

Hoyt RE, Liebenberg AP (2025) Enterprise risk management. In: Dionne G (ed), Handbook of insurance, 3rd edn. Springer, New York
Imbens GW (2022) Causality in econometrics: choice vs chances. Econometrica 90(6):2541–2566
Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 61:467–476
Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, New York
Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47(1):5–86
Jin Y, Jorion P (2006) Firm value and hedging: evidence from U.S. oil and gas producers. J Financ 61:893–919
Kang JK, Kim JM (2008) The geography of block acquisitions. J Financ 63(6):2817–2858
Kilian L (2009) Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market. Am Econ Rev 99:1053–1069
Kim H, Kim D, Im S, Hardin JW (2009) Evidence of asymmetric information in the automobile insurance market: dichotomous versus multinomial measurement of insurance coverage. J Risk Insur 76(2):343–366
Lalonde RJ (1986) Evaluating the econometric evaluation of training programs with experimental data. Am Econ Rev 76(4):604–620
Leamer EE (1978) Specification searches: Ad hoc inference with nonexperimental data. Wiley Series in Probability and Mathematical Statistics
Leamer EE (1983) Let's take the con out of econometrics. Am Econ Rev 73(1):31–43
Lee YW (2013) Testing for the presence of moral hazard using the regulatory reform in the car insurance market: case of Korea. Jpn Econ Rev 64:414–429
Lee DS, Lemieux T (2010) Regression discontinuity designs in economics. J Econ Lit 48:281–355
Low H, Meghir C (2017) The use of structural models in econometrics. J Econ Perspect 31(2):33–58
Maliar S, Salanié B (2022) Testing for asymmetric information in insurance with deep learning. Working paper, Columbia University, 31 p
Manski CF (1990) Nonparametric bounds on treatment effects. Am Econ Rev 80:319–323
Mayers D, Smith CW (1982) On the corporate demand for insurance. J Bus 55:281–296
Mincer J (1974) Schooling, experience and earnings. National Bureau of Economic Research, New York
Mnasri M, Dionne G, Gueyie JP (2017) The use of nonlinear hedging strategies by US oil producers: motivations and implications. Energy Econ 63:348–364
Neyman J (1923) Statistical problems in agricultural experiments. J R Stat Soc 2(Supplement):107–180
Okui R (2009) The optimal choice of moments in dynamic panel data models. J Econ 151:1–16
Olivella P, Vera-Hernández M (2013) Testing for asymmetric information in private health insurance. Econ J 123(567):96–130
Parra A, Winter R (2025) Optimal insurance contracts under moral hazard. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer, New York
Pauly MV (1968) The economics of moral hazard: comment. Am Econ Rev 58(3):531–537
Pearl J (2009) Causal inference in statistics: an overview. Stat Surv 3:96–146
Pearl J (2012) The do-calculus revisited. CORR abs/1210.4852
Pérez-Gonzalez F, Yun H (2013) Risk management and firm value: evidence from weather derivatives. J Financ 68:2143–2176
Picard P (2025) Economic analysis of insurance fraud. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer, New York
Pinquet J (2025) Nonnegative second-order semiparametric analysis and experience rating in non-life insurance. In: Dionne G (ed) Handbook of insurance, 3rd edn. Springer, New York
Puelz R, Snow A (1994) Evidence of adverse selection: equilibrium signaling and cross-subsidization in the insurance market. J Polit Econ 102:236–257
Richaudeau D (1999) Automobile insurance contracts and risk of accident: an empirical test using french individual data. Geneva Pap Risk Insur Theory 24(1):97–114
Rosenbaum PR, Rubin DB (1983a) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J Royal Stat Soc Ser B (Methodol) 45:212–218
Rosenbaum PR, Rubin DB (1983b) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rothschild M, Stiglitz J (1976) Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. Q J Econ 90:629–650. Reprinted in: Dionne G, Harrington S (eds) (1992) Foundations of insurance economics – readings in economics and finance. Kluwer Academic Publishers
Rowell D, Zweifel P (2024) Separating equilibria with search and selection effort: evidence from the auto insurance market. J Risk Financ Manag 17:154. https://doi.org/10.3390/jrfm17040154

Rowell D, Nghiem SH, Connelly LB (2017a) Two tests for ex ante moral hazard in a market for automobile insurance. J Risk Insur 84:1103–1126
Rowell D, Nghiem SH, Connelly LB (2017b) Testing for asymmetric information in insurance markets: a test for ex ante moral hazard revisited. Econ Lett 150:4–5
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688–701
Saito K (2006) Testing for asymmetric information in the automobile insurance market under rate regulation. J Risk Insur 73(2):335–356
Sargan JD (1958) The estimation of economic relationships using instrumental variables. Econometrica 26:393–415
Shavell S (1979) Risk sharing and incentives in the principal and agent relationship. Bell J Econ 10(1):55–73
Smith CW, Stulz RM (1985) The determinants of firms' hedging policies. J Financ Quant Anal 20(4):391–405
Spindler M (2014) Econometric methods for testing for asymmetric information: a comparison of parametric and nonparametric methods with an application to hospital daily benefits. Geneva Risk Insur Rev 39:254–266
Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments. Econometrica 65:557–586
Stulz RM (1984) Optimal hedging policies. J Financ Quant Anal 19:127–140
Stulz RM (1996) Rethinking risk management. J Appl Corp Financ 9:8–24
Su L, Spindler M (2013) Nonparametric testing for asymmetric information. J Bus Econ Stat 31(2):208–225
Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. J Econ 225:175–199
Thistlethwaite DL, Campbell DT (1960) Regression discontinuity analysis: an alternative to the ex post facto experiment. J Educ Psychol 51(6):309–317
Townsend RM (1979) Optimal contracts and competitive markets with costly state verification. J Econ Theory 21(2):265–293
Tufano P (1996) Who manages risk? An empirical examination of risk management practices in the gold mining industry. J Financ 51:1097–1137
Wilson C (1977) A model of insurance markets with incomplete information. J Econ Theory 16:167–207
Wooldridge JM (2010) Econometric analysis of cross section and panel data, 2nd edn. MIT Press, Boston
Zavadil T (2015) Do the better insured cause more damage? Testing for asymmetric information in car insurance. J Risk Insur 82:865–889

Econometrics of Insurance Based on Telematics Information and Machine Learning

Montserrat Guillén

1 Introduction: What Is Telematics?

Telematics is the use of advanced communication technologies, such as GPS, cellular, satellite communication, and the Internet, to gather and transmit data on the location, movements, status, and behavior of an object, typically a vehicle. In insurance, telematics is used to gather information about a policyholder's behavior, habits, and risk factors. This information provides a more accurate assessment of risk than traditional non-telematics information and may be included as rating factors to determine insurance policy premiums.

In non-life insurance, telematics is increasingly being adopted for tracking or monitoring. In motor insurance, products whose premiums are based on driving features observed through telematics are known as usage-based insurance (UBI). Other areas in which telematics may be considered in the rate-making process include health insurance, which may incorporate the use of self-tracking wearables, and household insurance, incorporating the installation in buildings of Internet-connected sensors, including fire and leak detectors and intrusion sensors. Schanz and Sommerrock (2016) believe that motor, health, and homeowners are the three personal lines of insurance that may experience the most significant impact from the potential of telematics. The rest of this chapter is, however, dedicated exclusively to the application of telematics in motor insurance, reflecting the significant amount of research in this field and the number of innovations made to date.

Classical rating factors in motor insurance can be broken down into the characteristics of the insured vehicle, policyholder attributes, and the traits describing the contract that is underwritten. The use of telematics information combined with classical rating factors, such as age, zip code, or vehicle model, has been disruptive for rate making in this particular line of business.

The basic telematics motor insurance rate-making scheme, pay-as-you-drive (PAYD), charges the policyholder a regular baseline premium plus a cost dependent on mileage. PAYD products are sold all over the world, gaining in momentum after the pandemic lockdowns when insureds realized that they were not using their vehicles as often as before and considered that a tariff linked to mileage would prove cheaper than a conventional premium.

M. Guillén (✉)
Department of Econometrics, Statistics and Applied Economics, Universitat de Barcelona, Barcelona, Spain
e-mail: mguillen@ub.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_14

1.1 Distance Driven

The innovative feature of the pioneer telematics devices of the early 2000s was that they could track the distance driven by a vehicle in real time. Distance-driven alone constitutes a rich source of data for the insurer. Litman (2011) shows that a driver who drives a long distance in a certain period of time is generally at greater risk of being involved in an accident than a driver who only drives a short distance in the same period, although the proportion may not be linear.

The use of annual mileage as a rating variable is explored by Lemaire et al. (2016). Before telematics data collection systems became available, policyholders could declare estimates of the distance they drove each year to their insurer; however, this information was usually inaccurate or deliberately lower than what it actually was because policyholders believe that declaring a vehicle to be barely used might reduce the premium they have to pay. With distance driven provided by odometers, Lemaire et al. (2016) find annual mileage to be an extremely powerful predictor of the number of claims at-fault. Regardless of other factors that might be considered to determine insurance premiums, this conclusion remains the same and it has been consistently corroborated (see Huang and Meng 2019; Duval et al. 2022). However, according to Lemaire et al. (2016) while annual mileage is an essential variable, it should not replace bonus-malus systems, which penalize policyholders when claims are reported and reward them with a lower premium if no claims are reported. Therefore, they conclude, a reliable rating system should include both annual mileage and bonus-malus as the two main factors, supplemented by other variables such as age, territory, and engine cubic capacity.

As more and more insurance companies offer telematics motor insurance, the role of distance driven as a measure of exposure to risk in PAYD is gradually replacing that of policy duration and bonus-malus systems will need to adapt to rate-making schemes that operate on a more frequent basis than that of traditional yearly updates. Experience rating or bonus-malus may gradually lose prominence, opening the door to PAYD schemes that operate on a monthly basis rather than an annual premium update rule.

1.2 Other Quantitative Telematics Data

Eling and Kraft (2020) and Gao et al. (2022a) undertake exhaustive reviews of the literature to illustrate how motor insurers consider many other telematics characteristics than that of just distance driven, including, but not limited to, driving at night, exceeding the posted speed limit, accelerating, braking, cornering, average speed, and maximum speed peak depending on the type of road traversed (e.g., urban, rural, or motorway). Additionally, So et al. (2021b) provide a list of quantitative telematics data that are commonly available by the country in which the data were collected.

Most driving patterns and driving scores are derived from the analysis of telematics characteristics that capture driving style and are generally considered to represent equally well all groups of drivers and all types of circumstances, including both weather and personal conditions. Indeed, Masello et al. (2023) and Reig Torra et al. (2023) show that telematics data could be combined with contextual information about the weather, traffic congestion, and road quality status.

1.3 Class Imbalance and Risky Events

Data on traffic accidents in insurance portfolios are naturally subject to what is known as the class imbalance problem due to the infrequency of insurance claims. The number of claims-free drivers greatly surpasses that of drivers who have reported an accident and it is challenging to identify

the main factors that can predict claims, even in datasets containing information on thousands of policyholders. To address this challenge, Guillen et al. (2020, 2021a, 2021b) expand the category of drivers who have experienced accidents to include those who experience near-miss incidents, a term borrowed from aviation that refers to an accident that was avoided, and which points to the occurrence of a risky event. In this way, riskiness is evaluated by the frequency of near-misses detected by telematics, rather than by the frequency of reported claims. One of the primary obstacles to conducting risk analyses with near-misses is defining near-miss events that accurately capture relevant information, for example, extreme acceleration and deceleration or excessive smartphone use while driving. This definition of risky events, and the fact that their frequency is correlated to the frequency of claims (Ma et al. 2018), has become increasingly popular, but it has also ignited a debate as to what should or should not be considered a near-miss. For example, in this context, harsh acceleration has been defined by many authors as acceleration that exceeds 3 m/s^2 (Tselentis et al. 2017; Henckaerts and Antonio 2022; Richman and Wüthrich 2022).

1.4 Preprocessing Telematics Information

Telematics information is recorded uninterruptedly while the vehicle's engine is running. As such, one of the greatest challenges is how to store these raw data and how to transform them into pertinent summaries without being a massive drain on time.

One example of raw telematics data records for one specific vehicle comprises a file in which each row contains information on the position of that vehicle (but note that some jurisdictions do not allow GPS tracking data to be kept for reasons of privacy and only permit the type of road to be recorded), its speed, or alternatively the distance driven (in meters) since the last measurement, and engine records (rotational speed, fuel instant consumption, pedal and gear positions, etc.). Typically, a raw data file contains at least one measurement per second and a clock timestamp to track the exact time of the recording.

The first step in preprocessing raw telematics data involves scanning for errors in this raw file. Unfortunately, errors occur more often than desired and need to be filtered or removed. After data cleaning, the information for each vehicle can be summarized at different levels: per trip (Baecke and Bocca 2017; Verbelen et al. 2018; Ma et al. 2018; Winlaw et al. 2019; Gao et al. 2022a), per day, per week, or per year (Ayuso et al. 2019). Summarizing raw telematics data can be interesting in its own right as an exercise in basic statistical description or can be addressed with advanced data science tools.

Assuming a daily collection of 100 KB of telematics data per driver (as estimated by Gao et al. 2022a), approximately 40 MB of data will be accumulated per driver each year. This means that even a relatively small insurer of, say, 100,000 drivers has to handle a volume of information that equates to a few terabytes of data annually. Storing such quantities of data most likely requires the use of cloud data storage solutions and, furthermore, analyzing a 1-year insurance portfolio of high-frequency time series data necessitates intensive computing and is extremely time-consuming. A further hurdle that has to be overcome is the current absence of homogeneity standards among telematics data providers around the world (see Tselentis et al. 2017) and, this problem has become even more relevant as different manufacturers incorporate their exclusively designed telematics on-board systems in new models to provide feedback to drivers on their driving experience.

Finding appropriate risk scores, indices, and aggregations from raw data can be a tedious process and requires machine learning algorithms to identify suitable, insightful data transformations.

Once telematics data have been cleaned and summarized, they can then be combined with claims information, a necessary step for insurance rate making. Telematics data per se do not contain

information on the occurrence of accidents, and collaboration between the telematics provider and the insurance company is needed to connect telematics data with the information stored in the insurance company's records. This is the final step before analysts can proceed with their rate-making procedures based on claim frequency and severity modeling.

Accountability and transparency are required in the connecting operations conducted between the telematics provider and the insurance company. If only black-box driving scores are provided to the insurer from a third party, and these are not fully reproducible or include potentially discriminating factors, then the company may not be able to use them in their pricing calculations.

Post-processing of telematics data is usually done internally by insurers as part of the risk analysis and rate making procedures. Some sophisticated approaches for feature engineering on telematics data are possible. Discrete Fourier transforms can be implemented to extract covariates from time series data, as proposed by Weidner et al. (2017). Scores, such as average speed, 90% quantile of acceleration rates, and proportions of driving distances on different types of roads, can be calculated (see Paefgen et al. 2014; Verbelen et al. 2018; Huang and Meng 2019; Sun et al. 2020). Moreover, advanced patterns can be found, as proposed by Gao et al. (2019, 2022a), who recommend extracting covariates from velocity-acceleration heatmaps using principal component analysis and neural network architectures.

2 Modeling Frequency Claims with Telematics Information

2.1 Overview

In presenting classical econometric models of telematics insurance claim frequency, we will assume some notation (see Dionne and Vanasse 1992; Frees et al. 2014; Wüthrich and Merz 2023). Suppose a fixed period of time, usually 1 year, but momentarily, we drop the time subscript t from our notation, and a portfolio of n policies. Let Y_i denote the number of claims for policy i in year t, with $i = 1, \ldots, n$. Note that frequency models usually focus on claims at fault for rate making, but other types of claims may also be of interest. In the rest of this chapter, we refer to claims in general without making this distinction. Let υ_i denote the magnitude of risk exposure or the offset for policy i in year t. Risk exposure refers to that fraction of the time period considered when the particular policy is operational, otherwise known as the duration fraction in the [0, 1] interval. For example, in the case of annual data, a policy that started on July 1 would receive a risk exposure equal to 0.5.

Once telematics data become available, risk exposure may be considered a combination of duration and distance driven. In that case, the role that exposure plays in telematics rate making follows from rather strong assumptions, it being usually accepted that proportionality exists between the expected claims and either the duration of the policy or the distance driven, or both. A monotonic transformation of distance can also be adopted.

Let x_i and z_i be the vectors of rate making factors. We denote traditional characteristics as x_i and telematics information as z_i. A common assumption is that the number of claims, Y_i, is independent across all policyholders and that they are modeled by a count-data model. The simplest model in this class is the Poisson regression model. So, under the Poisson model assumption, the expected number of claims conditional on the observed covariates and the exposure is specified as:

$$E\left(Y_i|x_i, z_i, \upsilon_i\right) = \upsilon_i \exp\left(x_i'\alpha + z_i{}'\beta\right) = \upsilon_i \exp\left(x_i'\alpha\right) \exp\left(z_i{}'\beta\right) = \mu\left(x_i, z_i, \upsilon_i\right). \tag{1}$$

In general, function $\mu(\cdot)$ links the covariates and the exposure to the conditional number of claims and has no other requirement than being non-negative. The canonical form of the Poisson model in generalized linear models (GLM) corresponds to the first expression in (1), where proportionality with respect to the exposure factor, υ_i, is assumed. This is the reason why exposure multiplies the exponential of the linear predictor, $x_i'\alpha + z_i{}'\beta$, where α and β are the parameters to be estimated. Model estimation with a standard method, such as maximum likelihood, leads to our obtaining $\hat{\mu}(\cdot)$ or the parameter estimates. A learning sample $\mathcal{L}$ and a test sample $\mathcal{T}$, with corresponding sizes $|\mathcal{L}|$ and $|\mathcal{T}|$, are used for estimation and evaluation, respectively. Goodness of fit can be given by the AIC criterion, and the Poisson deviance loss is used for evaluation in the test sample as follows:

$$\frac{2}{|\mathcal{T}|}\sum_{\substack{i\in\mathcal{T}\\ Y_i\neq 0}} Y_i\left(\frac{\hat{\mu}(x_i,z_i,\upsilon_i)}{Y_i}-1-\log\left(\frac{\hat{\mu}(x_i,z_i,\upsilon_i)}{Y_i}\right)\right)+\frac{2}{|\mathcal{T}|}\sum_{\substack{i\in\mathcal{T}\\ Y_i=0}} 2\cdot\hat{\mu}(x_i,z_i,\upsilon_i). \tag{2}$$

If telematics variables z_i are included in the frequency model as just an additional set of covariates, as in the first two equalities of (1), then there is not much more to say about the claims frequency model. From that point on, telematics information in the form of a set of covariates is simply an additional group of rate-making factors and all the traditional tools can be implemented accordingly. In doing so, if claims are observed yearly, we also implicitly assume that telematics data, z_i, are also updated only once a year.

2.2 Boosting Formulas

Model boosting formulas seek to separate the specification of the conditional claim frequency in three parts: namely, exposure, non-telematics covariates, and telematics covariates. Therefore, the model specification can be presented as:

$$E(Y_i|x_i,z_i,\upsilon_i)=\upsilon_i\exp\left(x_i'\alpha\right)\rho(z_i)=\mu(x_i,z_i,\upsilon_i), \tag{3}$$

where $\rho(\cdot)$ is a telematics correction factor, which allows model (3) to be estimated in two steps. First, the part corresponding to the coefficients of x_i and, second, the (possibly nonlinear) part that contains z_i. This is the method advocated by Ayuso et al. (2019) and Gao et al. (2022b). It is a two-step procedure for updating a conditional expectation of the baseline level $\upsilon_i\exp\left(x_i'\alpha\right)$ with additional information, $\rho(z_i)$, each time the telematics summary is updated. The boosted specification formula in (3) includes, in this particular case, the combination of a GLM and a neural network approach, or any other regression function.

Classical models with telematics covariates have been extended to generalized additive models (Verbelen et al. 2018) and to other extensions of the Poisson model, the negative binomial model and their corresponding zero-inflated versions (Guillen et al. 2019).

2.3 Desynchronized Information

One of the most disruptive practical issues when combining telematics data with claims data and classical rate-making factors is that telematics information may not have been collected during the exact same period as the observed claims. For example, claims associated with each vehicle may refer

to one entire year, while telematics data may only be available for 1 month. In this case, the role of exposure should compensate for the lack of synchronization.

Assume that Y_i denotes the number of claims made on policy i in 1 year, but assume that telematics information captures features that accumulate over time such as distance driven, and weekday and night driving, for example, which have only been obtained for 1 month. We denote month m vector information by z_i^m, where the time reference is a fraction of 1 year, which we denote by f^m. The question is how to unify these two desynchronized sources of data. The ideal solution would be to look at dates and find how many claims were reported in the fraction of the year corresponding to month m, Y_i^m, which would then be the number of claims that were observed in the same month whenever data in z_i^m were also observed. However, merging yearly records with monthly telematics data may not be possible for an incipient telematics system because claims are so infrequent that no driver may have reported any claims at all.

Artificial synchronization is a practical approximation which allows us to consider that $\mathrm{Y_i}$ (claims in 1 year) has a total exposure equal to 12 times the duration of the observation period for the telematics data z_i^m (information collected in 1 month). As Guillen et al. (2021a) propose, when combining Y_i with the month's observations of telematics covariates z_i^m,the original exposure for policy i, υ_i, should be replaced by υ_i^*, so that $\upsilon_i^* = \upsilon_i \times 12$. Note that $(1/f^m) = 12$. Similarly, if telematics information, z_i^m, is available and has been collected during 16 weeks and we assume that claims information correspond to 1 year (i.e., 52 weeks), then exposure should be multiplied by 52/16. Note that the duration of the observation interval is relevant, but it does not have to coincide for all policies in the portfolio, which means that the observation period for telematics can be no more than a few weeks for some policies and longer for others, so f^m can incorporate a subscript i.

The above method is especially useful when a company has to combine claims information with telematics information that has been obtained in a system which has only been operating for a short period of time. In the suggested approximation, assumptions are strong. In fact, by solving the synchronization problem in the way proposed by Guillen et al. (2021a, 2021b), it is assumed that drivers' behavior over time is stable, something that has been found to be questionable, especially after being involved in an accident (Pérez-Marín et al. 2019).

Duval et al. (2022), in conducting an analysis of logistic regression with lasso penalty for the occurrence of claims, find that telematics data become redundant after 3 months or 4000 km. This is why it should not be necessary to monitor drivers constantly if the only objective is to calculate a risk score, whereas continuous monitoring would be required if telematics were to be linked to dynamic pricing.

2.4 Modeling Frequency Panels of Claims with Telematics Information

When analyzing telematics data and summarizing information at a granularity level of less than 1 year, claims are so infrequent that it is highly unlikely for a policyholder to report more than one claim in that time interval. This is especially true of weekly datasets for exploring at-fault claims, because in practice no policyholder reports more than one claim at-fault in a week or less. As a result, claims frequency modeling becomes the analysis of binary outcomes in a panel data structure of policies observed over several time periods (days, weeks, and months), usually smaller than 1 year, or trips as the basic observation unit.

Let us consider Y_{it} a binary response that equals 1 if claim at-fault is reported by policy i, in time t, $i = 1, \ldots, n$, $t = 1, \ldots, W_i$, where W_i is the number of periods observed for policy i. Let υ_{it} be the corresponding risk exposure offset and x_i, z_{it} the vectors of rate-making factors where traditional and telematics covariates, are denoted by x_i and z_{it}, respectively.

When a panel data structure is assumed, Y_{it} is independent across all policyholders, and if there is conditional independence over time, then a binary response panel data model can be expressed as:

$$E\left(Y_{it}|x_i, z_{it}, \upsilon_{it}\right) = \mu\left(x_i, z_{it}, \upsilon_{it}\right) = \text{Prob}\left(Y_{it} = 1|x_i, z_{it}, \upsilon_{it}\right) = p_{it} \tag{4}$$

Again, the imbalance property represents a practical problem, since the occurrence of claims is far less frequent than their non-occurrence, and all subsequent problems arise from the fact that many policyholders do not report a single claim during the whole period W_i. When the dependent variable is all zeros for many individuals, Semykina and Wooldridge (2018) and Wooldridge (2019) mention that some parameters in the panel model may not be identified, because some individual effects cannot be identified. Chamberlain (2010) considers the econometric binary response model for panel data with the logistic link function and just two time points, and shows that parameter identification is possible. In addition, (4) should be improved by including past claims for performing experience rating (see Pinquet 2020; Ahn et al. 2023).

2.5 Impact of Telematics on Contract Choice and Behavior

Econometric models of insurance telematics enhance our understanding of the impact of this technology on pricing, contract choice, and behavior.

Barry and Charpentier (2020) stress that the current shift in data analysis toward predictive analytics conceptually contradicts the collective basis of insurance, since excessive personalization limits the mutualization of risk. Geyer et al. (2020) find that individuals' decisions about driving, insurance, and the type of contract underwritten (with or without telematics) may not be systematically linked to their decisions about driving carefully to avoid accidents. Eling and Kraft (2020) look at the advantages of letting policyholders agree to provide telematics data under their policy contract, and Holzapfel et al. (2023) argue that imperfect informative monitoring signals and the small impact on the premium explain why the demand for UBI products is still low.

Stevenson et al. (2021) conducted a randomized control trial to investigate whether driving behavior could be influenced by feedback from telematics, financial incentives, or neither. The study finds no significant differences between the groups, but the intervention groups showed a reduction in hard braking and hard acceleration compared to the control group.

3 Machine Learning Analysis of Telematics Information

3.1 Overview

Machine learning methods, a general class of techniques that include statistical models and algorithms, enable computers to learn patterns and relationships in data. Traditional models, such as GLM for frequency and severity models, which are used in non-life insurance rate making, can be considered machine learning tools. In general, machine learning methods can be classified as either unsupervised methods, whenever no response output is defined before fitting, and supervised methods, whenever a target outcome depends on a number of inputs.

The following models might be labeled as machine learning tools (Wüthrich and Merz 2023):

- *Decision trees and random forests*: These models are oriented toward the recursive partition of data into smaller subsets, based on the features considered most informative for predicting the target variable.
- *Support vector machines*: These models use a hyperplane to separate the data into different classes and find the optimal hyperplane that maximizes the margin between the classes.
- *Neural networks*: These models simulate the behavior of the human brain, using layers of interconnected nodes that learn to recognize patterns in the data. Bayesian networks represent the relationships between variables using a directed acyclic graph, and use Bayes' theorem to update the probability of each variable based on the observed data.
- *Ensemble methods*: These combine multiple models, such as bagging, boosting, and stacking, to improve overall performance.

Deep learning is a subfield of machine learning that involves training deep neural networks with many layers, often using specialized hardware such as graphics processing units (GPUs) to accelerate computations.

3.2 Some Applications in Telematics

Telematics data can be analyzed using techniques from machine learning (Wüthrich 2017) either by employing unsupervised techniques, where drivers are classified according to their driving patterns based exclusively on their similarities, or by employing supervised techniques, in which a response is approximated from the observed inputs or covariates. The latter can also be useful for generating scores.

Gao and Wüthrich (2019) summarize the driving information for each driver by transforming the information on their trips into a graphical representation between velocity and acceleration, or v-a heat map, which shows how they accelerate and brake at different speeds. In this way, they compress a large volume of data into a small matrix. So et al. (2021a) present an algorithm that uses cost-sensitive multi-class adaptive boosting to handle class imbalances in telematics data, and show, using empirical data collected in Canada, that it provides a better assessment of driving behavior than traditional risk variables and outperforms other learning models designed to handle class imbalances.

Duval et al. (2022) find that logistic regression with a lasso penalty is the most suitable classification algorithm. Richman and Wüthrich (2023) explore different interpretations of deep learning models, which select features internally and introduce a novel network architecture that incorporates similar characteristics to generalized linear models yet delivers better predictive capabilities thanks to representation learning. This is a promising area for model interpretability.

Machine learning methods also allow claim severity to be explored (see Fung et al. 2022) by studying a mixture composite regression model for the simultaneous modeling of attritional and large claims, while dealing with the challenges posed by multimodality and heavy-tailed data.

The main disadvantage of implementing some of the machine learning algorithms mentioned above in insurance rate making is that some methods are black boxes. That is, they do not provide an explicit analytic connection between inputs and outputs. This is not a limitation for the risk analysis and prediction undertaken within insurance companies, but it does create a considerable disadvantage if machine learning forms part of premium calculation. Insurance rates need to be fully transparent to consumers, and regulators must guarantee that premiums comply with ethical principles and standards. According to Richman (2022), the actuarial literature has yet to explore the effects of using sophisticated machine learning algorithms. For example, hyperparameter selection, which is needed to implement some of the more complex models, may have a huge impact on predictions. Model

interpretability is a challenge, especially in deep neural network models. Unless these knowledge gaps in facilitating the safe and effective adoption of deep learning models are overcome, the extensive application of machine learning techniques will remain limited in insurance rate making.

Richman (2021) exemplifies recent advances in machine learning to extract information about driving behavior patterns from telematics data. Maillart (2021) presents new efforts to explain black-box models by using a surrogate model. Meng et al. (2022a) describe claim frequency modeling with driving behavior features and improved boosted trees. Zhu and Wüthrich (2021) present clustering driving styles via image processing. Pesantez-Narvaez et al. (2019) introduce XGBoost methods that are based on improving loss optimization in claims frequency models with telematics data, but they recognize that, compared to GLMs, internal tuning is necessary and interpretability is much harder to achieve.

Wüthrich and Merz (2019) advocate the combination of neural networks and classical modeling, which provides a unique way of blending the two approaches and which they refer to as a combined actuarial neural net (CANN). The implementation of CANN starts with a classical model like the one described in (1) and then, in a second step, the exact same problem of modeling claims frequency is tackled within a neural network. By so doing, all covariates are transformed in hidden layers and final scores are produced from the network. These newly transformed values can be understood as complex combinations of the initial covariates that, in particular, contain high-dimensional covariate interactions that are relevant and which are difficult to obtain manually. The outputs of the neural network are introduced linearly in (1) as additional covariates.

3.3 The Perils of Overfitting

Methods, such as CANN, work well with large datasets. However, a potential disadvantage when applying machine learning algorithms is that of overfitting, as explicitly discussed by McDonnell et al. (2023). Overfitting refers to a common problem encountered in the training of neural networks, whereby the model is excessively tailored to fit the training dataset, but where this results in poor generalization performance on new, unseen data. This can lead to inaccurate predictions and unreliable risk assessments, and may even result in the loss of the autocalibration property by which the sum of the observed claims equals the sum of predicted expected claims, a property that is essential for actuarial fairness.

When building a CANN from a training dataset, the model optimizes the internal loss functions to capture the relevant patterns in the data, but without careful parameter tuning the network may select overly complex input-output links so that the approximation may not be valid for another set of data. To address the overfitting problem, regularization methods, dropout and early stopping are implemented. These techniques can help prevent the model from overfitting by imposing constraints on the model's parameters or limiting the number of iterations during training. Additionally, techniques such as cross-validation and data augmentation can be employed to reduce overfitting by increasing the effective size of the training dataset and evaluating the model's performance on multiple subsets of the data.

3.4 Autocalibration

In the realm of machine learning, numerous tools are available to select appropriate models. However, employing these tools without due diligence can result in an incorrect model selection. One of the most detrimental missteps is the selection of a non-autocalibrated model. In the domain of insurance

rate making, it is imperative that the total predictions (premium rates) closely align with the aggregate responses (losses). Achieving this alignment is crucial, as highlighted in both Wüthrich (2023), Denuit and Trufin (2023), and Frees et al. (2021). This alignment serves as a fundamental property in insurance pricing, ensuring that each group of policyholders paying an identical premium is, on average, self-financing. The cumulative premiums from such a group precisely cover the expected loss of that particular group.

Lindholm et al. (2023) have introduced a distribution-independent predictor for the risk premium, designed to be locally unbiased and derived from any initially proposed predictor. This predictor exhibits a piecewise constant nature, aligning with a partition of the covariate space, and is inherently autocalibrated due to its construction. Notably, this partition is thoughtfully designed, considering duration or observed exposure based on telematics as a significant factor.

3.5 Summarizing Trips with Machine Learning

In telematics systems, driver behavior is processed and saved at the end of each trip. The number of trips in the observed period is informative and obviously does not need to be the same for all drivers.

The procedure proposed by Wüthrich (2017) is an allocation technique that belongs to the unsupervised learning methods to analyze driving styles with v-a heatmaps that compare average speed (velocity) versus acceleration.

Wüthrich and Buser (2016, 2023), in an analysis of trip data, demonstrate how to compare three different drivers. Gao et al. (2022b) extend the comparison of good and bad drivers and integrate trip information in the model specified in (3). They start by selecting two small groups of drivers that are manually labeled as good and bad. From this sample of trips taken by drivers in both groups, a network is trained and then used to classify all trips by all drivers based on telematics data $z_{i,j}$ from individual i and trip j. The classifier is denoted as $\hat{\psi}\left(z_{i,j}\right)$, $i = 1, \ldots, n; j = 1, \ldots, J_i$, where J_i is the number of available trips for i. Then a score is defined for each driver as:

$$\overline{\psi}_i = \frac{1}{J_i}\sum\nolimits_{j=1}^{J_i} \hat{\psi}\left(z_{i,j}\right).$$

Finally, the individual score is inserted in model (1) or, alternatively, a smoothed credibility version can be used:

$$E\left(Y_i|x_i, z_i, \nu_i\right) = \nu_i \exp\left(x_i'\beta\right) \exp\left(\alpha_0 + \alpha_1 \overline{\psi}_i\right).$$

4 Pricing Methods That Incorporate Telematics Data

All the methods presented in the previous sections are integrated in pricing schemes. In this way they provide the added-value of updated motor insurance prices derived from the fact that the telematics driving data send continuous signal information, or at least much more frequently than once a year.

Ayuso et al. (2016, 2019) and Henckaerts and Antonio (2022) propose, first, estimating a baseline pricing model for the complete portfolio with claim history and self-reported risk characteristics as usual. The baseline price can then be updated using the drivers' telematics information. All schemes of this sort correspond to UBI pricing.

In a general setting, Guillen et al. (2024) propose a framework for a simple linear pricing scheme that can be expressed as follows:

$$f\left(\sum_{k=1}^{K}\alpha_k x_{ki}\right)+\gamma g\left(D_{it}\right)+(1-\gamma)\,h\left(\sum_{j=1}^{J}\beta_{1j}z_{jit}+\sum_{l=1}^{L}\beta_{2j}z_{li(t-1)}\right), \tag{5}$$

where K baseline characteristics, x_{ki}, are weighted with coefficients α_k for policy i in the first term and $f(\cdot)$ is a transformation link. This first term involves all items used in classical rate making, that is, engine cubic capacity, age and driving experience, zip code, vehicle model, and all a company's rate-making information. This term should already take into consideration the expected severity of claims and general expenses plus solvency loadings in the final tariff. The second term corresponds to the price that is paid as a function, $g(\cdot)$, of distance driven by i in period t, which is usually the identity, but might be monotonically increasing and concave, and weighted by parameter γ, with $\gamma \in [0, 1]$ which balances the role of distance versus other telematics features considered in the third term. Note that, without loss of generality, distance can be fully integrated in the third term with $\gamma = 0$, and $g(D_{it})$ equal to one of the regular telematics rating factor, z_{jit}. Distance can also be considered exclusively, as in rate-making schemes that only consider distance driven without other telematics, by setting $\gamma = 1$. All remaining telematics covariates, risks scores, and all other transformed features from any preprocessing filtering or machine learning or deep learning methods are included in the third term of (5), possibly with transformation $h(\cdot)$, and define the varying costs to be charged to the policyholder. We explicitly allow telematics features to refer to J telematics covariates in the current period, or L telematics covariates from the previous time interval $(t - 1)$. More lags could be considered. Moreover, note that a multiplicative structure, such as the log-link in the classical Poisson model approach, would be more natural and would guarantee positivity. However, it may be less appealing to inform consumers, who are used to simple arrangements similar to a pay-per-kilometer insurance plan.

Desjardins et al. (2023) propose a count-data model with hierarchical random effects for the posterior insurance rate making of vehicles belonging to a fleet, by allowing random effects for the fleet, the vehicles, and time. Their rate-making formula for posterior premium calculation is easily applicable using telematics variables.

4.1 Pay-as-You-Drive

Pay-as-you-drive (PAYD) is a car insurance policy that calculates the cost of the insurance premium based on the amount of driving done by a policyholder. A scheme of this type corresponds to the first two terms of (5), with $\gamma = 1$, and can be found in markets around the world. To uphold privacy, a telematics device recording distance data does not typically transmit detailed locations, especially when the data collected are subject to strict privacy regulations.

PAYD policies provide incentives for policyholders to drive less, potentially reducing traffic congestion and environmental pollution (see Litman 2005; Boucher et al. 2017; Frees et al. 2014).

4.2 Pay-How-You Drive

Pay-how-you-drive (PHYD) policies calculate premiums based on driving data that typically include information on speed, acceleration, braking, cornering, and other driving behaviors. Rate making is then calculated on the driving behavior of the policyholder, with safe drivers receiving lower premiums than risky drivers, the scheme taking a similar form to (5).

Policyholders receive feedback on their driving behavior in PHYD scheme, which can help them improve their driving habits. Denuit et al. (2019) propose a multivariate credibility modeling for

UBI pricing using behavioral data that takes into consideration an experience rating based on three telematics features (total distance driven, night driving, and urban driving) and the number of claims. They advocate that behavioral telematics information should be considered as part of an experience rating scheme. This means that some telematics features should not be incorporated in the premium score directly, as ordinary a priori observables. Instead, some telematics information can be used ex-post to correct the baseline premium. The actuary can handle a diverse range of scenarios that might arise in insurance practice using this method. Such situations might include new drivers that do not have a telematics record, contracts with varying seniority levels, and drivers who use their vehicles to different extents, resulting in varying amounts of telematics data.

4.3 Near-Miss Telematics

Near-miss telematics insurance is a type of insurance rate making that uses telematics event data to evaluate the risk of an individual driver and which includes a cost per risky event. Near-miss events are defined as incidents in which a driver takes an action that could have resulted, but which did not, in an accident. Guillen et al. (2021a) present examples of near-miss events that capture a driver's behavior, such as speed, acceleration, braking, cornering, and cell phone usage while driving. However, the fact that some risky events may not have been caused by the drivers themselves is a major drawback when including some specific events in the rate-making formula.

Near-miss telematics insurance is a relatively new concept and still has to adequately address a number of challenges, such as defining what constitutes a near-miss event and determining how to measure the risk associated with these events accurately. This is why near-misses for prices may not be accepted by consumers. Nonetheless, it has the potential to improve road safety by encouraging drivers to adopt safer driving behaviors. One interesting feature of near-miss telematics is the way in which the rate-making scheme is presented. If we consider distance driven and exposure in the whole time interval ($\upsilon_{it} = 1$), then under the specification in model (1), where coefficients, β, are percentage increases due to each near-miss event of the base rate $P_{i\,\text{base}}$, $\overline{C}$ is the mean claim cost, that is, an approximate linear pure premium that would penalize each additional near-miss as follows:

$$\overline{C}\exp\left(x_i'\alpha\right)\exp\left(E_{it}'\beta\right) = P_{i\ \text{base}}\exp\left(E_{it}'\beta\right) \cong P_{i\ \text{base}}\left(1 + E_{it}'\beta\right) \leq P_{i\ \text{base}} + E_{it}'\beta_{\max},$$

where $\beta_{\max} = \max_{1\leq i\leq n}\left(\beta P_{i\ \text{base}}\right)$. Note that $\beta_{\max}$ depends on the maximum value of $P_{i\,\text{base}}$. In practice, to determine $\beta_{\max}$, a reasonable threshold for $P_{i\,\text{base}}$ is, for example, three times the average base premium. When considering each additional unit of distance ($D_{it} > 0$), the approximate base premium is:

$$\begin{aligned}\overline{C}T_{it}\exp\left(x_i'\alpha\right)\exp\left(E_{it}'\beta\right) = P_{i\ \text{base}}D_{it}\exp\left(E_{it}'\beta\right) &\cong P_{i\ \text{base}}\left(1 + E_{it}'\beta + \ln\left(D_{it}\right)\right)\\ &\leq P_{i\ \text{base}} + E_{it}'\beta\ p_{\max} + p_{\max}\ln\left(D_{it}\right),\end{aligned}$$

where $p_{\max} = \max_{1\leq i\leq n}\left(P_{i\ \text{base}}\right)$.

5 Further Topics

5.1 Driving Scores

A driving score is based on a combination of factors, such as the driver's speed, acceleration, and braking habits, as well as the frequency and severity of lane departures, distance to the vehicle in front, distractions, and other safety events. The score is updated in real-time and displayed on the vehicle's dashboard.

Scores usually range from 0 to 100, with higher scores indicating safer driving behavior. Drivers can use the score as feedback to improve their driving habits, and some insurance companies may use scores to determine insurance premiums.

Many traditional navigation apps use expert knowledge to score safe driving behavior. Meng et al. (2022b) propose a model that scores driving risk using telematics data and claims. The one-dimensional convolutional neural network takes individual trip data and returns a score from 0 to 1. By averaging each driver's score, the Poisson generalized linear model for insurance claims incorporates telematics-based information. Tesla's driving score (https://www.tesla.com/support/insurance/real-time-insurance) and similar scales introduced by manufacturers, such as Toyota's pulse score,[1] are metrics that evaluate a driver's behavior behind the wheel based on data collected by the vehicle's cameras and sensors. However, their effectiveness may vary depending on various external factors, such as road conditions and weather.

Guillen et al. (2021b) employ quantile regression models to estimate the percentiles of distance traveled at speeds exceeding the legal limit, taking into account the characteristics of drivers and their travel routes. These percentiles, conditional on drivers' characteristics, can be displayed in percentile charts and interpreted as individual risk scores.

5.2 Contextual Data

Factors such as posted speed limit, weather temperature, wind speed, traffic conditions, and road slope are among the top ten most relevant causes of risky events. Masello et al. (2022) report that the driving context is a significant predictor of driving risk. Analyses of how variations in contextual features impact risky events reveal that low-speed limits increase the predicted frequency of speeding and phone unlocking, while high-speed limits decrease harsh accelerations. Low temperatures decrease the expected frequency of harsh maneuvers, and precipitation increases harsh acceleration, harsh braking, and distraction events. Road slope, intersections, and pavement quality are found to be the most critical factors among road layout attributes.

5.3 Regulators and Ethical Concerns About Telematics Data

Telematics data have frequently raised concerns on the grounds that they might seem too intrusive or, even, capture personal information that is ethically questionable or illegal for insurance rate making. In addition, the availability of individual risk information and excessive personalization may challenge principles of risk-pooling and risk-sharing. Taken together, they might be deemed reasons

[1] Toyota's driving score https://support.toyota.com/s/article/What-is-Drive-Pulse-s-10726?language=en_US#:~:text=Answer-,Drive%20Pulse%20score%20uses%20the%20sensor%20data%20from%20the%20connected,a%20scale%20of%200%2D100

to limit the expansion of telematics. Cevolini and Esposito (2020), for example, examine the impact of predictive algorithms on insurance premiums and the potential risks of discrimination and exclusion from coverage. However, a wide gap still exists between the many possibilities identified by analysts for using digital technologies in this field and their actual usage by insurance companies. Cather (2020) recognizes that innovative data allow insurers to remove discriminatory pricing variables from their risk classification systems and to select low-risk profiles, something that is encouraging small insurance companies to enter the race for telematics-based rate-making schemes.

5.4 A Synthetic Telematics Dataset

For various reasons, it is not easy to obtain telematics data from private insurers. First, such data are highly sensitive, containing drivers' personal and, on occasions, location-related information. In this regard, insurers are obviously concerned about protecting their customers' privacy and complying with data protection regulations. Second, insurers may consider the data proprietary and, as such, may not be willing to share them with competitors or third parties. Moreover, the act of both collecting and managing large volumes of telematics data requires significant investment in infrastructure and resources. Finally, insurers may have concerns about the accuracy and quality of telematics data, which may affect the validity of any predictive models built using them. As a result, insurers are often reluctant to share telematics data with researchers, making it challenging to obtain real datasets for research and modeling purposes.

So et al. (2021b) provide a synthetic dataset comprising 100,000 policies that can be used to develop models for assessing risks in UBI. The dataset includes observations on a driver's claims record, in addition to classical and telematics-related risk variables. Generation is based on a three-stage process involving machine learning algorithms. In the first stage, a synthetic portfolio of the space of feature variables is created using an extended synthetic minority oversampling technique (SMOTE) algorithm. The data set includes generated telematics factors. In the second stage, feedforward neural networks are used to simulate values for the number of claims as multiple binary classifications. In the third stage, feedforward neural networks are used to simulate values for the aggregated amount of claims as regression, with the number of claims included as a feature variable. According to So et al. (2021b), the resulting dataset is comparable to real datasets when fitting Poisson and gamma regression models. Visualization and data summarization techniques show a remarkable degree of similarity between the synthetic and real datasets. The synthetic dataset provides an excellent sandbox in which researchers can test new modeling approaches and their application.

Acknowledgments Support was provided by the Spanish Ministry of Science and Innovation, FEDER grant PID2019-105986GB-C21 and NextGenerationEU, grant number TED2021-130187B-I00. MG gratefully thanks the ICREA Academia Program and members of the RISKcenter at the University of Barcelona.

References

Ahn JY, Jeong H, Lu Y (2023) A simple Bayesian state-space approach to the collective risk models. Scand Actuar J 2023(5):509–529

Ayuso M, Guillen M, Perez-Marín AM (2016) Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. Risks 4(2):10

Ayuso M, Guillen M, Nielsen JP (2019) Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. Transportation 46:735–752

Baecke P, Bocca L (2017) The value of vehicle telematics data in insurance risk selection processes. Decis Support Syst 98:69–79

Barry L, Charpentier A (2020) Personalization as a promise: Can Big Data change the practice of insurance? Big Data Soc 7(1):2053951720935143
Boucher JP, Côté S, Guillen M (2017) Exposure as duration and distance in telematics motor insurance using generalized additive models. Risks 5(4):54
Cather DA (2020) Reconsidering insurance discrimination and adverse selection in an era of data analytics. Geneva Pap Risk Insur-Issues Pract 45:426–456
Cevolini A, Esposito E (2020) From pool to profile: social consequences of algorithmic prediction in insurance. Big Data Soc 7(2):2053951720939228
Chamberlain G (2010) Binary response models for panel data: identification and information. Econometrica 78(1):159–168
Denuit M, Trufin J (2023) Model selection with Pearson's correlation, concentration and Lorenz curves under autocalibration. Eur Actuar J 2:1–8. https://doi.org/10.1007/s13385-023-00353-5
Denuit M, Guillen M, Trufin J (2019) Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. Ann Actuar Sci 13(2):378–399
Desjardins D, Dionne G, Lu Y (2023) Hierarchical random-effects model for the insurance pricing of vehicles belonging to a fleet. J Appl Econ 38(2):242–259
Dionne G, Vanasse C (1992) Automobile insurance ratemaking in the presence of asymmetrical information. J Appl Econ 7(2):149–165
Duval F, Boucher JP, Pigeon M (2022) How much telematics information do insurers need for claim classification? N Am Actuar J 26(4):570–590
Eling M, Kraft M (2020) The impact of telematics on the insurability of risks. J Risk Financ 21(2):77–109
Frees EW, Derrig RA, Meyers G (eds) (2014) Predictive modeling applications in actuarial science, vol 1. Cambridge University Press
Frees EW, Bolancé C, Guillen M, Valdez EA (2021) Dependence modeling of multivariate longitudinal hybrid insurance data with dropout. Expert Syst Appl 185:115552
Fung TC, Tzougas G, Wüthrich MV (2022) Mixture composite regression models with multi-type feature selection. N Am Actuar J 27(2):396–428
Gao G, Wüthrich MV (2019) Convolutional neural network classification of telematics car driving data. Risks 7(1):6
Gao G, Meng S, Wüthrich MV (2019) Claims frequency modeling using telematics car driving data. Scand Actuar J 2:143–162
Gao G, Meng S, Wüthrich MV (2022a) What can we learn from telematics car driving data: a survey. Insur: Math Econ 104:185–199
Gao G, Wang H, Wüthrich MV (2022b) Boosting Poisson regression models with telematics car driving data. Mach Learn 111:243–272
Geyer A, Kremslehner D, Muermann A (2020) Asymmetric information in automobile insurance: evidence from driving behavior. J Risk Insur 87(4):969–995
Guillen M, Nielsen JP, Ayuso M, Pérez-Marín AM (2019) The use of telematics devices to improve automobile insurance rates. Risk Anal 39(3):662–672
Guillen M, Nielsen JP, Pérez-Marín AM, Elpidorou V (2020) Can automobile insurance telematics predict the risk of near-miss events? N Am Actuar J 24(1):141–152
Guillen M, Nielsen JP, Pérez-Marín AM (2021a) Near-miss telematics in motor insurance. J Risk Insur 88(3):569–589
Guillen M, Pérez-Marín AM, Alcañiz M (2021b) Percentile charts for speeding based on telematics information. Accid Anal Prev 150:105865
Guillen M, Pérez-Marín AM, Nielsen JP (2024) Pricing weekly motor insurance drivers' with behavioral and contextual telematics data. Heliyon 10(16), e 36501
Henckaerts R, Antonio K (2022) The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. Insur: Math Econ 105:79–95
Holzapfel J, Peter R, Richter A (2023) Mitigating moral hazard with usage-based insurance. Available at SSRN 4325478
Huang Y, Meng S (2019) Automobile insurance classification ratemaking based on telematics driving data. Decis Support Syst 127:113156
Lemaire J, Park SC, Wang KC (2016) The use of annual mileage as a rating variable. ASTIN Bull 46(1):39–69
Lindholm M, Lindskog F, Palmquist J (2023) Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. Scand Actuar J 2023(10):946–973
Litman T (2005) Pay-As-You-Drive Pricing and insurance regulatory objectives. J Insur Regul 23(3):36–53
Litman T (2011) Pay-as-you-drive vehicle insurance in British Columbia. Pacific Institute for Climate Solutions, University of Victoria
Ma YL, Zhu X, Hu X, Chiu YC (2018) The use of context-sensitive insurance telematics data in auto insurance ratemaking. Transp Res A 113:243–258
Maillart A (2021) Toward an explainable machine learning model for claim frequency: a use case in car insurance pricing with telematics data. Eur Actuar J 11:579–617

Masello L, Castignani G, Sheehan B, Murphy F, McDonnell K (2022) On the road safety benefits of advanced driver assistance systems in different driving contexts. Trans Res Interdiscip Persp 15:100670
Masello L, Castignani G, Sheehan B, Guillén M, Murphy F (2023) Using contextual data to predict risky driving events: a novel methodology from explainable artificial intelligence. Accid Anal Prev 184:106997
McDonnell K, Murphy F, Sheehan B, Masello L, Castignani G (2023) Deep learning in insurance: accuracy and model interpretability using TabNet. Exp Syst Appl 217:119543
Meng S, Gao Y, Huang Y (2022a) Actuarial intelligence in auto insurance: claim frequency modeling with driving behavior features and improved boosted trees. Insur: Math Econ 106:115–127
Meng S, Wang H, Shi Y, Gao G (2022b) Improving automobile insurance claims frequency prediction with telematics car driving data. ASTIN Bull 52(2):363–391
Paefgen J, Staake T, Fleisch E (2014) Multivariate exposure modeling of accident risk: insights from Pay-as-you-drive insurance data. Transp Res A Policy Pract 61:27–40
Pérez-Marín AM, Ayuso M, Guillen M (2019) Do young insured drivers slow down after suffering an accident? Transport Res F: Traffic Psychol Behav 62:690–699
Pesantez-Narvaez J, Guillen M, Alcañiz M (2019) Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks 7(2):70
Pinquet J (2020) Poisson models with dynamic random effects and nonnegative credibilities per period. ASTIN Bull 50(2):585–618
Reig Torra J, Guillen M, Pérez-Marín AM, Rey Gámez L, Aguer G (2023) Weather conditions and telematics panel data in monthly motor insurance claim frequency models. Risks 11(3):57
Richman R (2021) AI in actuarial science–a review of recent advances–part 2. Ann Actuar Sci 15(2):230–258
Richman R (2022) Mind the gap–safely incorporating deep learning models into the actuarial toolkit. Br Actuar J 27:e21
Richman R, Wüthrich MV (2022) LASSO regularization within the LocalGLMnet architecture. Adv Data Anal Classif:1–31. https://doi.org/10.1007/s11634-022-00529-z
Richman R, Wüthrich MV (2023) LocalGLMnet: interpretable deep learning for tabular data. Scand Actuar J 2022:71–95
Schanz KU, Sommerrock F (2016) Harnessing technology to narrow the insurance protection Gap. Published by the Geneva Association—'International Association for the Study of Insurance Economics', Zurich
Semykina A, Wooldridge JM (2018) Binary response panel data models with sample selection and self-selection. J Appl Econ 33(2):179–197
So B, Boucher JP, Valdez EA (2021a) Cost-sensitive multi-class AdaBoost for understanding driving behavior based on telematics. ASTIN Bull 51(3):719–751
So B, Boucher JP, Valdez EA (2021b) Synthetic dataset generation of driver telematics. Risks 9(4):58
Stevenson M, Harris A, Wijnands JS, Mortimer D (2021) The effect of telematic based feedback and financial incentives on driving behaviour: a randomised trial. Accid Anal Prev 159:106278
Sun S, Bi J, Guillen M, Pérez-Marín AM (2020) Assessing driving risk using internet of vehicles data: an analysis based on generalized linear models. Sensors 20(9):2712
Tselentis DI, Yannis G, Vlahogianni EI (2017) Innovative motor insurance schemes: a review of current practices and emerging challenges. Accid Anal Prev 98:139–148
Verbelen R, Antonio K, Claeskens G (2018) Unravelling the predictive power of telematics data in car insurance pricing. J R Stat Soc Ser C (Appl Stat) 67(5):1275–1304
Weidner W, Transchel FW, Weidner R (2017) Telematic driving profile classification in car insurance pricing. Ann Actuar Sci 11(2):213–236
Winlaw M, Steiner SH, MacKay RJ, Hilal AR (2019) Using telematics data to find risky driver behaviour. Accid Anal Prev 131:131–136
Wooldridge JM (2019) Correlated random effects models with unbalanced panels. J Econ 211(1):137–150
Wüthrich MV (2017) Covariate selection from telematics car driving data. Eur Actuar J 7:89–108
Wüthrich MV (2023) Model selection with Gini indices under auto-calibration. Eur Actuar J 13:469–477
Wüthrich MV, Buser C (2016) Data analytics for non-life insurance pricing. SSRN Manuscript ID 2870308, Version of September 10, 2020, (page 251)
Wüthrich MV, Buser C (2023) Data analytics for non-life insurance pricing. Swiss Finance Institute Research Paper No 16-68
Wüthrich MV, Merz M (2019) Yes, we CANN! ASTIN Bull 49(1):1–3
Wüthrich MV, Merz M (2023) Statistical foundations of actuarial learning and its applications. Springer Nature
Zhu R, Wüthrich MV (2021) Clustering driving styles via image processing. Ann Actuar Sci 15(2):276–290

Data Science for Insurance Fraud Detection: A Review

Denisa Banulescu-Radu and Yannick Kougblenou

1 Introduction

Insurance fraud[1] is a major issue that affects insurance companies and policyholders all over the world. Each year, the insurance industry loses billions of dollars due to fraudulent activities such as falsified or exaggerated claims, staged accidents, and other types of insurance scams. Insurance companies use nowadays a variety of techniques and methods to detect and prevent fraudulent activities in order to combat insurance fraud, which explains why, in recent years, both academics and professionals have focused on designing, developing, and testing accurate and reliable fraud detection systems.

Most of academic research on financial fraud, in general, is related either to (i) the theoretical literature on the incentive theory à la Laffont and Tirole (1993) oriented through the economic understanding of the phenomenon, or (ii) the empirical applications mainly concerned with prediction and based on econometric and machine learning tools (see the surveys of Adewumi and Akinyelu 2017; Kou et al. 2004; Bolton and Hand 2002, among others). Concerning insurance fraud, first comprehensive investigations studies on the magnitude and the stakes of the phenomenon were conducted in the 1980s, followed by theoretical studies aimed to enhance the effectiveness of auditing strategies and contract design (see Picard 2013, 1996; Bond and Crocker 1997; Crocker and Tennyson 1999; Watt 2003, among others). As highlighted by Derrig (2002), it was in the 1990s that practical models started to be used to effectively filter claims in insurance fraud investigations, including

This research benefits from the financial support of "Data Science and Insurance Fraud Detection" research initiative under the aegis of the Europlace Institut of Finance, the "Fraud detection and Anti-Money Laundering" research project (REDFLAG APR-IA AE 2019-1850) funded by Centre-Val de Loire Region, and the ANR program CaliBank (ANR-19-CE26-0002-02).

[1] According to Insurance Europe, insurance fraud affects all types of insurance, including nonlife, life, and health, and it occurs when someone: (i) provides false or incomplete information during the application process; (ii) submits misleading claims, including exaggerating legitimate losses; or (iii) engages in deceptive actions to gain unwarranted benefits from the insurance contract (Insurance Europe 2019).

D. Banulescu-Radu (✉) · Y. Kougblenou
Laboratory of Economics of Orléans, University of Orléans, Orléans, France
e-mail: georgiana-denisa.radu@univ-orleans.fr; denisa.banulescu-radu@univ-orleans.fr; koug@outlook.fr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_15

database structuring and strategies for selecting suspect claims (Major and Riedinger 1992), fuzzy clustering (Derrig and Ostaszewski 1995), simple regression scoring models (Weisberg and Derrig 1998), and Probit and Logit models (Artıs et al. 1999; Belhadji et al. 2000; Dionne et al. 2009). Nowadays, machine learning algorithms are increasingly being used in fraud detection because they allow insurance companies to analyze massive amounts of data, identify complex and subtle patterns that would be difficult to detect with traditional methods, and detect criminal networks and fraudulent claims more effectively (see, Viaene et al. 2002; Baesens et al. 2015; Nian et al. 2016; Dhieb et al. 2020; Óskarsdóttir et al. 2022, among others). Moreover, real-time automated fraud detection is possible, which can help reduce financial losses associated with fraud.

Three main categories of algorithms, along with the main challenges related to the fraud modeling process, will be discussed in this chapter, with a quasi-exclusive focus on the academic literature on insurance fraud. First, one of the most common approaches is based on *unsupervised learning*, which focuses on the analysis of data and identification of anomalous cases and patterns of fraud using data analytics and machine learning algorithms. For instance, using clustering techniques, one can identify groups of similar claims, and among them groups of potentially fraudulent claims. Anomaly detection, on the other hand, can help to identify claims that are very different from others and that raise suspicions of fraud. Although the application of unsupervised methods cannot guarantee the detection of all cases of insurance fraud, their main value lies in the recognition of new and time-evolving fraud techniques and patterns. They thus play an essential role in the ongoing process of fraud prevention (see Nian et al. 2016; Gomes et al. 2021 and references therein). Another useful trait of unsupervised learning is the absence of labeling, which is often expansive and unreliable in fraud detection.

Secondly, *supervised learning* techniques use historical data to build predictive models that are able to detect fraudulent activity. The underlying principle of supervised machine learning in this context is to use labeled data (fraudulent/non-fraudulent) to train a model to recognize the particularities of fraudulent insurance claims. Subsequently, the objective of the model is to predict whether a new claim is fraudulent or not, based on its inherent characteristics. In this way, insurance companies can quickly and efficiently identify potentially fraudulent claims, thereby limiting the financial losses associated with fraud. However, it is important to note that this approach presupposes that a previous claims dataset has already been scrutinized to classify it as fraudulent or non-fraudulent, a process that is not error proof (Caudill et al. 2005; Warren and Schweitzer 2018). On the other hand, the rarity of fraud cases, in association with unexamined or mislabeled claims, makes the database highly skewed (imbalanced), presenting significant challenges for the construction and evaluation of fraud detection models (Baesens et al. 2003; Brownlee 2020). Third, in many cases, fraudsters operate in organized groups, and the most intuitive way to detect them in this case is by using *social network analysis*. This approach identifies connections between individuals and groups that might be engaged in fraudulent activities by carefully examining the relationships between them and their behavioral patterns.

In light of these considerations, it becomes evident that fraud detection cannot be approached as just an ordinary classification topic by machine learning approaches, mainly due to the intentional, time-evolving and rare nature of the phenomenon. Several additional challenges emerge. Firstly, the availability and quality of data is an area of particular concern, as only a limited number of databases are accessible for obvious confidentiality reasons. Secondly, model interpretability poses a unique challenge in this context, as we need to avoid revealing the workings of the model to potential fraudsters. Finally, the question of fairness can also be a significant point of concern. For instance,

insurance companies are particularly cautious when it comes to developing expert systems that could result in legal action for discriminatory practices.[2]

Importantly, there are two main approaches to evaluating the effectiveness of the fraud detection process. Firstly, it can be assessed by comparing the statistical improvement in predictive performance achieved by machine learning algorithms against standard econometric methods. Secondly, it can be assessed by taking into account the cost dimension. The latter is essential as well, as it becomes imperative to minimize financial losses and efficiently allocate limited resources toward fraud investigation and detection. Consequently, the detection of financial fraud needs to be acknowledged as a cost-sensitive classification problem, as highlighted in previous studies by Viaene et al. (2007), Höppner et al. (2022), Vanderschueren et al. (2022), De Vos et al. (2023). We will empirically illustrate these two aspects by first examining the advantages in terms of predictive performance achieved through the use of the random forest and eXtreme Gradient Boosting (XGBoost) algorithms compared to a standard logistic Lasso regression model. Next, we will look at the financial benefits that could potentially be realized by employing cost-sensitive variants of the three aforementioned methods.

The rest of this chapter is structured as follows: Sect. 2 presents the main technical tools used for both the prevention and detection of insurance fraud. A distinction is made between the analysis of model performance using purely statistical criteria and their financial efficiency. Section 3 focuses on the four most significant challenges that data scientists confront most often in the fraud detection process. An empirical illustration of the gains achieved by using supervised machine learning techniques over standard econometric models, and the benefits of cost-sensitive versus cost-insensitive approaches, is presented in Sect. 4. Finally, concluding remarks and potential directions for future research are provided in Sect. 5.

2 Data Science for Insurance Fraud Analysis

As the volume of data and explanatory variables continues to grow, conventional models are gradually giving way to more advanced machine learning methods that aim to answer to different statistical and financial objectives (e.g., minimize false positives[3] and false negatives, minimize the loss given fraud, etc.). Moreover, machine learning models are also preferred for fraud detection due to their ability to process quickly large amounts of unstructured data, their adaptability to changes in fraudster behavior, the fraud risk assessment in real time and high statistical performance and automation, which offer significant advantages over standard econometric models. The challenge on these models of detecting and, ultimately, preventing fraudulent claims lies in their ability to identify the characteristics that differentiate them from legitimate claims. For financial fraud in general, and insurance fraud in particular, these methods can be categorized into three categories: unsupervised, supervised, and social network learning approaches. The utility of each category will be presented in the following subsection and supported by academic references.

2.1 Insurance Fraud: a Statistical Analysis

Unsupervised learning methods aim at finding unusual anomalous behavior deviating from the average behavior or norm (Bolton et al. 2001). They are known as *prevention* models, designed

[2] In some situations, denied payments may have to be justified in court, requiring explanations before a judge.

[3] False positives generate expenses related to the investigation of the associated alerts.

to detect fraud before it happens, and they do not require labeled data. Indeed, obtaining labels (fraud/no fraud) associated with each claim can, in certain circumstances, be costly and time-consuming, sometimes even unfeasible. In addition, there may be implicit bias in the labels output, which can result either from human intervention or the procedure used to assign them. One of the earliest unsupervised fraud detection techniques was actually proposed in the context of insurance fraud by Brockett et al. (2002), and further extended by Ai et al. (2009). The technique is called PRIDIT (principal component analysis of RIDITs) and is able to calculate claim-specific fraud suspicion score (reflecting the likelihood of fraud for each claim), and that can be used as inputs in subsequent statistical modeling. Later, Ai et al. (2013) proposed PRIDIT-FRE (PRIDIT-based Fraud Rate Estimation) as an estimator of the overall rate of fraud in a population of claims, and robust to the bias contained in the audited sample of data. In 2020, Golden et al. (2020) introduced aPRIDIT, which extends the previous methods by the introduction of asymmetries, by considering that information about target group membership may vary across the levels of categorical inputs.

Intuitively, fraud is treated often as an anomaly problem [i.e., a large collection of data (input variables) related to the type of fraud under analysis are used to detect any new anomalous patterns among the existing ones]. Accordingly, Nian et al. (2016) and Ghorbani and Farzai (2018), among others, show that unsupervised learning techniques have the potential to identify suspicious cases in a timely manner and therefore reduce significantly economic losses, to both the insurers and policy holders. Kose et al. (2015) propose an interactive machine learning approach that allows incorporating expert knowledge in an unsupervised setting. The authors used it to detect fraud and abusive cases in healthcare. In the context of unsupervised learning, a ranking of the degree of relative abnormality can also be provided to evaluate the most suspicious transactions/claims. For instance, Stripling et al. (2018) proposed iForestCAD approach that computes conditional anomaly scores, useful for fraud detection. They demonstrate the usefulness of their technique to uncover workers' compensation fraud. As for Gomes et al. (2021), the authors used unsupervised deep learning approaches to offer pragmatic insights into the main driving factors of fraud, and one of the datasets used was related to automobile insurance claims. Unsupervised learning can indeed be useful for organizations that start doing fraud detection as no labeled historical data on fraud are required. However, Baesens et al. (2015) argue that the usage of unsupervised learning for fraud detection requires extensive follow-up and validation of the identified, suspicious observations.

Supervised learning methods are designed to *detect* fraud cases ex post. In this setting, both fraudulent and non-fraudulent past records are used to design models that allow to assign new observations to one of the two classes. Therefore, the aim of supervised learning is to build an analytical model predicting this target measure of interest (Hastie et al. 2009). The target is generally used to guide the learning process during an optimization procedure. Two types of predictive analysis can be distinguished depending on the type of the target variable: regression and classification. In regression, the target variable is continuous (a typical example in the context of fraud detection is the prediction of the amount of fraud). In classification, the target is categorical, which means that it takes only a limited set of predefined values (mostly binary, i.e., fraud/no fraud). Fraud detection commonly involves classic supervised econometric and machine learning methods such as linear regression methods, logistic regression, decision trees, k-nearest neighbors (KNN), support vector machine (SVM), and so on. However, it is essential to avoid relying too heavily on a single approach or model type, as fraudsters use a range of techniques that require a combination of approaches to detect them. To address this issue, it can be helpful to using ensemble machine learning methods such as bagging, boosting, and random forest. These methods involve estimating multiple analytical learners together, rather than relying on a single one, to improve fraud detection accuracy. Viaene et al. (2002) conducted a benchmarking analysis of multiple machine learning algorithms on an automobile fraud dataset and observed that the variations in ROC AUC scores among the different

algorithms were generally insignificant. For their specific context and dataset, no algorithm stands out as consistently superior in terms of predictive performance across all evaluation scenarios; however, decision tree models exhibit lower efficiency compared to other models. Phua et al. (2004) combined algorithms like backpropagation neural network, Naïve Bayes and C4.5 decision tree, as well as meta-classifiers (bagging), while taking into account the issue of skewed data distribution. New innovative machine learning techniques, such as random forest (Breiman 2001), boosted decision trees (Zhang et al. 2018), XGBoost (Chen and Guestrin 2016), and their extensions, have also been used recently in the context of insurance fraud detection. For instance, Dhieb et al. (2020) recommend XGBoost for the detection of automobile insurance fraud, and Li et al. (2018) adapt random forest algorithm to solve the same issue.

As observed, the literature on insurance fraud detection has developed partly around automobile insurance fraud, and there are few studies on other types of insurance fraud (i.e., household, crop, or healthcare insurance fraud). We can cite Sahni et al. (2020), who propose a solution based on the analysis of aerial images of field fires to investigate the characteristics and magnitude of declared fires. The information collected is then incorporated as inputs in different classification models (e.g., SVM, KNN, Naïve Bayes, etc.) to better detect crop insurance fraud. Banulescu-Radu and Yankol-Schalck (2023) propose a practical guideline to efficiently detect insurance fraud with a study case on a household insurance database.

Social network analysis has also proven its utility for fraud detection. Although fraud can be difficult to detect on the basis of available variables, it is often very useful to analyze the relationships between fraudsters, as fraud is often a carefully organized crime, rather than a standalone phenomenon. Fraudsters tend to cluster together to exchange knowledge about how to commit fraud, use the same resources, and collaborate on fraudulent activities (the phenomenon is known as *homophily* in social sciences).[4] This assumes that the probability of someone committing fraud depends on the people (s)he is connected to (situation known as *guilt-by-associations*) (Baesens et al. 2015), and exploiting relational information might provide interesting insights in criminal patterns and activities. More precisely, a social network consists of nodes, which stand in for people or organizations, and edges, which show the connections or interactions among them. Social network analysis can identify potentially fraudulent actors or suspicious patterns of behavior that may indicate fraudulent activity by analyzing the structure and characteristics of the network.[5] Social network analysis can also be used to identify important network nodes, such as influential individuals or key players in a fraudulent scheme. Furthermore, fraud detection systems can target their investigations and allocate resources more effectively to prevent fraudulent activity by focusing on these nodes. Šubelj et al. (2011) are among the first to use network analysis for insurance fraud detection. More recently, Óskarsdóttir et al. (2022) have proposed the use of bipartite network analysis to detect fraud in automobile insurance claims as well. The authors link first all relevant parties (i.e., policyholders, brokers, adjusters, and garages) in a network, compute a fraud score for each claim by using the Bi-Rank algorithm, and extract then features related to both the fraud score and the neighborhood structure of the claims. Together with standard claim features, the newly created features are included in a logistic-based model to detect fraud.[6] Additionally, Tumminello et al. (2023) use social bipartite

[4] In the context of insurance fraud detection, homophily can occur when fraudsters are more likely to collaborate with others who have similar characteristics or are in the same social network.

[5] If a fraudulent actor operates in a social network, they may be linked to other nodes who are also engaging in suspicious behavior. Social network analysis can identify clusters or subgroups of nodes that are more likely to be involved in fraudulent activity by identifying the connections and relationships between nodes.

[6] Lismont et al. (2018) use similar approaches in the fiscal fraud environment, and Van Vlasselaer et al. (2017) for the detection of social security contribution fraud.

network analysis to detect communities of fraudsters in the insurance fraud framework, by using microlevel data of subjects and vehicles involved in the same accidents.

2.2 Insurance Fraud: A Cost-Sensitive Analysis

As highlighted by Höppner et al. (2022), the task of detecting fraudulent transactions is essentially a problem of binary classification from the perspective of a machine learning algorithm. And, classification models are commonly trained and evaluated solely in terms of statistical performance measures. In other words, fraud detection models often focus on minimizing error rate rather than on the cost of classification (Derrig 2002). This can have important implications, as the purely statistical approach fails to address two of the key challenges in banking and insurance industries: i) how to minimize financial losses due to fraud and ii) how to optimally allocate limited resources to investigate fraud.

Typically, the cases that are misclassified lead to different costs. For instance, false negatives are especially expensive because they allow fraudulent claims to go undetected, resulting in significant financial losses for insurance companies. False positives involve the cost of investigation. Additionally, they can also imply a hidden cost, because they result in legitimate claims being denied, which can lead to a deterioration of the insurance company's reputation. Thus, in order to minimize fraud costs, cost-sensitive learning approaches can be used to assign different weights to false positives and false negatives.[7] In the academic literature, we can actually distinguish between class-dependent and instance-dependent cost-sensitive learning methods. Class-dependent cost-sensitive learning assigns fixed cost weights to all instances, optimizing the model to minimize the total cost of errors across the dataset (see Dal Pozzolo et al. 2014; Sahin et al. 2013; Elkan 2001; Philip and Chan 1998, and references therein). However, in real-world applications, the costs of errors can vary widely between instances, which suggests the need for instance-dependent cost-sensitive learning. This approach allows the model to assign different cost weights to each instance, based on its specific characteristics, and leads to more accurate predictions and a reduction in the overall costs associated with errors.[8]

Lately, a number of methods that account for instance-dependent misclassification costs have been proposed, with application in the field of credit card fraud detection. For instance, Bahnsen et al. (2013) proposed the Bayes Minimum Risk (BMR) approach, Bahnsen et al. (2014a) introduced the cost-sensitive logistic regression, and Bahnsen et al. (2015b) proposed the cost-sensitive decision tree (CSDT), which was extended to cost-sensitive random forest (CSRF) in Bahnsen et al. (2015a). More recently, Höppner et al. (2022) introduced the cost-sensitive Lasso-regularized logistic classifier (CSLogit) and CSboost classifier, a gradient boosting incorporating instance-specific cost matrix to improve fraud detection. Certainly, the fraud management process is both proactive, which involves continuous monitoring of new fraud opportunities and of new patterns and emerging trends, and agile in deploying revised controls to combat fraud. Nevertheless, it has associated costs and relies on limited financial and human investigative resources (Viaene and Dedene 2004). In practice, a manager can only assign a small subset of transactions or claims (identified by the model as suspicious) to the

[7] Specifically, the cost of a false negative should be set to be much higher than the cost of a false positive, reflecting the higher cost of failing to detect fraud.

[8] It should be noted that instance-dependent cost-sensitive learning may be more computationally demanding than class-dependent cost-sensitive learning. In addition, a larger dataset may be required to accurately capture the specific costs associated with each instance. Therefore, the choice between these two approaches should be based on the specifics of the problem to be solved as well as the resources available.

fraud team for investigation. As a solution, Vanderschueren et al. (2022) propose an approach that optimizes directly the expected profit of the assignment given a limited stochastic capacity.

Although mainly proposed in the context of credit card fraud, the aforementioned techniques can definitely be extended to develop a better data-driven insurance fraud detection system as well, by aligning the predictive model with the decision-making context. This can be done by optimizing the processing of positive signals (alerts) emitted by the model and scaling them to the control capacities and the different costs accompanying the decisions. In Sect. 4, we provide an in-depth discussion and apply four instance-dependent cost-sensitive methods, namely CSLogit, CSRF, CSBoost, and BMR, in the context of insurance fraud detection. We assess their statistical and economic performance relative to that of conventional classifiers.

Table 1 summarizes and completes the (non-exhaustive) list of references on insurance fraud detection discussed above. For each study, we specify the type of insurance analyzed, the main methods applied, and if it makes the object or not of a cost-sensitive analysis. Most studies are done in the context of automobile insurance fraud, which is certainly related to data availability, and as observed, there is still room in the fraud insurance context for cost-sensitive techniques to be exploited.

2.3 Data Science for Insurance Fraud: Practical and Policy Implications

Artificial intelligence (AI) is nowadays widely adopted by the insurance sector to perform a variety of functions, such as understanding policyholder preferences, facilitating claims processing, and improving fraud detection (Insurance Europe 2023). According to Bauer et al. (2021), the integration of digitization in association with big data offers valuable opportunities to improve this industry. In the context of insurance fraud detection, this is primarily achieved by leveraging increased data availability, real-time monitoring capabilities, advanced analytics, and automation. Some of the benefits are discussed below. Firstly, insurers can develop more accurate models to identify suspicious patterns and behavior, enabling them to flag potential fraud cases more effectively. Secondly, machine learning algorithms enable real-time monitoring of insurance transactions and claims. This proactive approach can help insurers detect fraud as soon as it occurs, thus preventing the payment of fraudulent claims which, once realized, become impossible to recover. Thirdly, by analyzing historical data and model results, insurers can optimize resource allocation. They can focus their investigative efforts on high-risk cases, reducing the time and costs associated with fraud investigations. Improved fraud detection also translates into significant cost savings for insurance companies, which in turn can lead to lower premiums for policyholders and strengthen their confidence in the insurance sector.

Naturally, digital transformation also raises a number of questions and concerns among regulators and supervisory authorities. Nonetheless, when insurance companies and intermediaries use AI, they operate within a regulated framework, with respect to the regulations applicable at international, European, and national levels. In Europe, these regulations encompass primary EU legislation and secondary EU law, such as the Insurance Distribution Directive, the Solvency II framework or the GDPR. In April 2021, the European Commission also presented the first EU regulatory framework for AI, namely the AI Act, which sets out a system for assessing and categorizing AI systems used in various applications according to the risks they entail for users. These different levels of risk determine the extent of the regulatory measures applied, with higher risks justifying stricter regulation (Madiega 2021, 2023). EIOPA embraces the legislative proposal and supports the objectives and principles of the AI Act, designed to promote the ethical and trustworthy use of AI.

Table 1 Non-exhaustive list of econometric and machine learning models in insurance fraud management

Research paper	Fraud type	Methods applied	Cost sensitive
Panel A—Supervised econometric and machine learning techniques			
Weisberg and Derrig (1998)	Automobile	Regression, tobit model	
Artıs et al. (1999)	Automobile	Multinomial logit	
Belhadji et al. (2000)	Automobile	Probit model	
Artís et al. (2002)	Automobile	Logistic regression	
Viaene et al. (2002)	Automobile	Logistic regression, KNN, Bayesian neural network, Naïve Bayes, SVM, decision tree	
Dionne et al. (2003)	Automobile	Auditing strategies, logistic regression	✓
Phua et al. (2004)	Automobile	Naïve Bayesian algorithm, C4.5 decision tree, backpropagation neural networks, Bagging	
Caudill et al. (2005)	Automobile	Multinomial logit	
Jin et al. (2005)	Crop	Logistic regression, probit	
Viaene et al. (2005)	Automobile	Bayesian neural network	
Atwood et al. (2006)	Crop	Yield-switching model	
Viaene et al. (2007)	Automobile	Cost-sensitive logit	✓
Pinquet et al. (2007)	Automobile	Bivariate probit model	
Bermúdez et al. (2008)	Automobile	Bayesian skewed logit	
Dionne et al. (2009)	Automobile	Logistic regression	✓
Caldeira et al. (2015)	Automobile	Artificial neural network	
Sundarkumar and Ravi (2015)	Automobile	Logistic regression, SVM	
Bayerstadler et al. (2016)	Healthcare	Bayesian multinomial latent variable	
Vassiljeva et al. (2017)	Automobile	Artificial neural network	
Roy and George (2017)	Automobile	Decision tree, random forest, Naïve Bayes	
Kowshalya and Nandhini (2018)	Automobile	Decision tree, Naïve Bayes, random forest	
Li et al. (2018)	Automobile	Random forest	
Wang and Xu (2018)	Automobile	SVM, random forest, deep neural network	
Majhi et al. (2019)	Automobile	Fuzzy c-means, random forest, logistic regression, XGBoost	
Padhi and Panigrahi (2019)	Automobile	SVM, multilayer perceptron, k-nearest neighbor	
Dhieb et al. (2020)	Automobile	XGboost, decision tree, KNN, Naïve Bayes	
Sahni et al. (2020)	Crop	Logistic regression, SVM, KNN, Naïve Bayes	
Severino and Peng (2021)	Household	Logistic regression, GBM, KNN, deep neural network, SVM, Naïve Bayes, random forest	
Błaszczyński et al. (2021)	Automobile	Random forest, SVM	
Yankol-Schalck (2022)	Automobile	Gradient boosting, neural networks	
Amponsah et al. (2022)	Healthcare	Decision tree, blockchain	
Aslam et al. (2022)	Automobile	Logistic regression, SVM	
		Naïve Bayes	
Banulescu-Radu and Yankol-Schalck (2023)	Household	Logistic regression, random forest, XGBoost, etc.	

Table 1 (continued)

Research paper	Fraud type	Methods applied	Cost sensitive
Panel B—Unsupervised techniques			
Derrig and Ostaszewski (1995)	Automobile	Fuzzy clustering	
Brockett et al. (2002)	Automobile	PRIDIT	
Yang and Hwang (2006)	Healthcare	Association rules	
Ai et al. (2009)	Automobile	PRIDIT	
Kirlidog and Asuk (2012)	Healthcare	Data mining, SVM	
Ai et al. (2013)	Automobile	PRIDIT-FRE	
Kose et al. (2015)	Healthcare	Interactive machine learning	
Nian et al. (2016)	Automobile	Unsupervised spectral ranking	
Wang and Xu (2018)	Automobile	Latent Dirichlet allocation	
Sun et al. (2018)	Healthcare	Patient cluster divergence	
Ghorbani and Farzai (2018)	Automobile	k-means clustering	
Hainaut (2019)	Nonlife insurance	Self-organizing maps (SOM)	
Golden et al. (2020)	Automobile	aPRIDIT	
Błaszczyński et al. (2021)	Automobile	Dominance-based rough set-balanced rule ensemble	
Gomes et al. (2021)	Automobile	Unsupervised deep learning	
Settipalli and Gangadharan (2023)	Healthcare	Weighted multitree—density-based clustering	
Panel C—Social network analysis			
Šubelj et al. (2011)	Automobile	Network and graph theory	
Pourhabibi et al. (2020)	Automobile	Social networks, graph-based anomaly detection	
Tumminello et al. (2023)	Automobile	Bipartite networks	
Óskarsdóttir et al. (2022)	Automobile	Bipartite networks	

3 Main Challenges for Modeling Insurance Fraud

There are many challenges in detecting financial fraud using machine learning. In this section, we will explore four of the most significant of these challenges, particularly relevant to the domain of insurance fraud modeling as well. This list is obviously not exhaustive and leaves open the possibilities of discussion and research.

3.1 Fraud as a Complex, Time-Evolving, and Deliberate Phenomenon

Fraud is a sophisticated, time-evolving, and intentional phenomenon. Thus, a methodological reflection on insurance fraud detection algorithms, as well as on their stability and governance, is highly required. First, insurance fraud can take a wide variety of forms (e.g., automobile, household, healthcare, crop insurance fraud, etc.), and the solution to each problem is domain-specific (Baesens

et al. 2015, 2021). Fraud schemes can also be highly sophisticated and involve multiple parties, including policyholders, medical providers, and legal representatives. Additionally, for a same type of fraud, fraudsters can be very creative and use a multitude of ways to perpetrate their illegal activities. Therefore, two transactions flagged as fraud cases may result from different actions and require specific detection models. This can also make it difficult for traditional models to detect new forms of fraud that have not yet been identified and highlights the importance of approaches based on model stacking (keeping together old models that fit an old fraud process and new models associated with new fraud methods) and model aggregation.

Second, fraud is not a static phenomenon, i.e., fraudsters have a highly adaptive behavior that evolves over time. They are aware of the dynamic of business environment itself and the constantly evolving fraud detection techniques, which encourages them to change their behavior as well to avoid detection. Moreover, fraudsters learn from the mistakes of their predecessors and/or the detection methods used by banks/insurance companies, and correct them fast. However, it is not reasonable to assume that a known fraudster will not get audited when submitting a new claim. Information on known fraudsters is shared among insurance companies, so attempting to deceive a different company is unlikely to result in less scrutiny. Fraud techniques become hence time evolving, making fraud a time-dependent phenomenon. Additionally, traditional rule-based approaches to fraud detection are nowadays relatively easy to comprehend and reverse-engineer by fraudsters Baesens et al. (2015). Consequently, insurance companies aim at developing and implementing combined, automated, and real-time fraud detection models that are updated regularly to remain effective.

Third, other difficulty encountered by modelers is represented by the fact that fraud is an intentional or deliberate action. Therefore, the *i.i.d.* assumption generally used in statistics is likely to be violated, since fraudsters often collaborate to create complex fraudulent networks. As seen previously, this characteristic of fraud can be addressed by incorporating social network analysis and clustering techniques into fraud detection models. By analyzing behavioral patterns and connections between individuals, it is possible to identify groups of individuals who are more likely to engage in fraudulent activities together. In addition, incorporating features related to social connections and network structure into fraud detection models could enhance their predictive power and sensitivity to fraudulent behavior associated with homophily.

3.2 Limited Access to Data and Data Quality

In order to produce high-quality research, the use of original and mainly proprietary datasets is highly required. In the specific context of fraud detection, acquiring such datasets represents a considerable challenge, which may also explain the limited academic literature on the subject. Companies are often hesitant to collaborate with academics on such topics for confidentiality reasons, and they are particularly reluctant to disclose data related to fraud cases or the strategies used for fraud detection. Consequently, few publicly available datasets on fraud cases exist, and the available datasets are typically outdated or have already been extensively studied.

Furthermore, the quality of available data is not straightforward due to the multiple data sources that have to be consistently merged, the incompleteness and inconsistencies in data, or the data misrepresentation.[9] Misrepresentation can appear either at the level of the target variable (i.e.,

[9] Typically, a complete dataset on fraud contains internal systems data on the customer, behavioral data, unstructured data embedded in text documents, qualitative expert-based data, publicly available data related to the general context of the economy, as well as social media data (Baesens et al. 2015). For example, Banulescu-Radu and Yankol-Schalck (2023) combine input variables from internal and external databases. Internal databases include data on claims, insureds, and property characteristics. These data are generally taken from claims reports or police files. External data, on the

mislabeled data) or of the inputs, which can lead to a dramatic model performance decrease. For instance, independent variables can be affected by completion bias: Suspicious claims might have more attributes completed because the insurance company's detection process flagged something unusual about the claim and prompted a more thorough investigation. Additionally, time dependency plays a critical role, as once payments are disbursed, it is improbable that they will ever be recovered. Furthermore, claims follow different life cycles and do not progress at the same speed. It is therefore essential to analyze the significant milestones in a claim's lifetime that symbolize reaching a "normal" level of information shared by all claims when they reach such milestones. At these junctures, it is imperative to predict fraud using consistent models that are applied uniformly across all claims.

Another major criticism when using supervised learning for fraud detection is that the models are trained on a variable of interest whose labeling process may have been imperfect or incomplete, in which case the classification model reproduces the phenomenon. In other words, if a fraud is not correctly labeled, the model will not look for it. The sources of mislabelization are various. First, among all claims, only a fraction is proposed by the model as suspicious (i.e., alerts) after the screening process. Clearly, no statistical technique is able to perfectly predict fraud, which means that fraudulent claims not detected by the model are labeled automatically as honest. Second, the special investigation units (SIUs) of insurance companies are limited in terms of financial and human resources, which means that internal and/or external fraud investigators can only investigate a limited number of alerts, the others going undetected and naturally considered as non-fraudulent. Finally, the investigation process is based unequivocally on factual evidence, but also on personal judgment and human understanding of the fraud phenomenon, which is not error free (i.e., data points labeled as non-fraudulent by fraud investigators may in fact be fraudulent, and vice versa). Needless to add, few fraud cases are finally brought to court, as insurers fear being involved in costly lawsuits and proceedings that could potentially damage their reputation. We propose an illustration of the mislabelization process described above in Fig. 1.

Importantly, this labeling problem is specific to the fraud domain and is not common to other classification problems (e.g., medicine, credit scoring, etc.). For example, the default of a customer in the banking sector or the occurrence or not of a disease in the medical field is observed objectively. To the best of our knowledge and despite the importance of the subject, the academic literature in this area is limited. For instance, Artís et al. (2002) examine the issue of data quality in the context of supervised machine learning models applied to fraud detection and propose a methodology based on logistic regression, which corrects the model to account for target variable misspecification. Caudill et al. (2005) extend the study of Artís et al. (2002) by using a multinomial logit model with missing information. More recently, Li et al. (2021) introduced a class of nonparametric methods to study the inputs misrepresentation issue in insurance applications that can be used to capture behavioral fraud (e.g., intentionally declaring untrue statements to alter insurance eligibility and/or lower the insurance premiums).

3.3 Fraud as a Rare Event

Beyond the problem of mislabeling, fraud is also inherently a rare event, which means that there are relatively few true fraudulent claims compared to the total number of claims processed by insurance companies (Baesens et al. 2015). Given these issues, the datasets used for fraud detection are generally huge, but highly imbalanced. They contain a huge number of legitimate cases and only few fraudulent

other hand, come from the French National Statistics Institute (INSEE) and include socio-demographic and geographical data, at various levels of granularity.

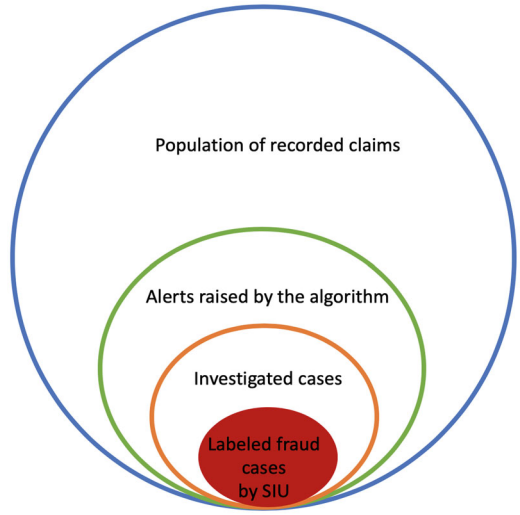


Fig. 1 Illustration of the labeling process

cases to learn from, which makes the detection process even more complicated.[10] It has been largely documented that the imbalanced class issue heavily compromises both the model learning process and the evaluation of its accuracy (see Fernández et al. 2018). Moreover, there is no consensus in the literature regarding the threshold for which a dataset is qualified as highly imbalanced (these thresholds vary between less than 1% and 5%) or the optimal resampling ratio to use to get it less skewed. Typically, the imbalance issue poses difficulties for learning algorithms, as they will be biased toward the majority group. However, in many real-world situations (medicine, fraud detection, churn analysis, etc.), the minority class is the class of interest because, despite its rarity, it can carry pertinent and valuable knowledge. Consequently, when dealing with skewed distributions, it is necessary to design intelligent systems capable of addressing such a challenge.

The topic of imbalanced learning has been widely discussed in the last two decades, and a significant number of surveys and research articles present recent advances on the subject (see Brownlee 2020; Kovács 2019; Fernández et al. 2018; Zhu et al. 2018, 2017, and references therein). Several solutions have been proposed to address class imbalance problems, and they are generally grouped into three main categories: (i) data sampling, which involves modifying the training instances to produce a more balanced distribution of classes; (ii) algorithmic modification, which involves adapting basic learning methods to better deal with class imbalance problems; (iii) cost-sensitive learning, which takes into account the higher costs of misclassifying instances of the minority class relative to instances of the majority class. Nevertheless, this problem is minimally addressed in the context of fraud detection in general, and almost nonexistent in the context of insurance fraud detection. Banulescu-Radu and Yankol-Schalck (2023) and Sundarkumar and Ravi (2015) are among the first to use data resampling methods in the context of insurance fraud detection and verify if the statistical performance of the models is improving.

[10] For instance, many fraud detection studies report a fraud ratio of less than 1%.

3.4 Difficulties in Evaluating the Performances of Fraud Models

The evaluation of fraud prevention/detection models is also critical and raises many questions. Starting with the assessment of *supervised learning* techniques, we will delve deeper into two main decisions that must be addressed when evaluating these models, particularly in the context of fraud detection. The first decision concerns the partitioning of the dataset (into a train dataset, i.e., the part of the data on which the model will be trained, and a test dataset, i.e., the part of the data on which the model's performance will be assessed).[11] The second decision is related to the choice of metrics used to assess either statistically the predictive performances of models or their financial efficiency. We start by discussing the statistical metrics. Although widely used in the classification environment, it is important to note that some evaluation measures are not well suited to fraud data (due to their skewness).[12] Consider an imbalanced dataset with a 1:100 ratio of minority class to majority class. Based on the percentage of good classifications, the model will be considered very efficient even if it classifies all individuals in the majority class, since a 99% good classification is excellent at first sight (i.e., the "accuracy paradox"). Conversely, when focusing on the "cost" aspect, it is essential to compute specific metrics that integrate the financial dimension of the phenomenon. In the following, we can classify these metrics into two broad categories: purely statistical metrics and economic metrics.

Typically, when conducting a classification task, a confusion matrix is generated, as shown in Table 2. This matrix gives information on the number of fraudulent/honest claims that were incorrectly classified by the model (false negatives, FN, and false positives, FP, respectively), as well as on the correct classifications (true positives, TP, and true negatives, TN, respectively). Moreover, different financial costs can be associated with incorrect classifications and benefits with correctly classified instances. A classification problem is considered cost insensitive when the costs associated with the two types of errors, FP and FN, are identical. It becomes class dependent cost sensitive when these costs differ but remain constant. Lastly, we refer to an example-dependent cost-sensitive classification problem when the cost matrix varies across all instances. Let $C_i(\hat{y}|y)$ be the cost of predicting class $\hat{y}$ for an instance i when the true class is y (i.e., $y, \hat{y} \in \{0, 1\}$). The costs associated to the four cells in the confusion matrix are different and have different sources. Following Hand et al. (2008), false and true positive cases are associated with the administrative/investigation cost ($C_i(1|0) = c_f$) that is fixed. In the case of a false negative, the cost is defined by the amount of the transaction, which will correspond to the amount of the reported claim i in the case of insurance fraud ($C_i(0|1) = A_i$). No cost is associated to a true negative. The instance-dependent costs are also summarized in Table 2, and a non-exhaustive list of performance indicators tailored to the general context of financial fraud is given in Table 3. Additionally, several clustering performance metrics are suggested for application in the context of *unsupervised learning* methods.

[11] In the case of large databases, the data is initially split into training and validation sets. Subsequently, cross-validation is performed on the training set to identify the optimal hyperparameters. Finally, the model's performance is evaluated using the validation dataset that was set aside initially. However, with small databases, typically featuring a limited number of positive cases, creating a separate validation set is often impractical. In such instances, cross-validation is applied directly to the entire database.

[12] "*In the case of class imbalances, the problem is even more acute because the default, relatively robust procedures used for unskewed data can break down miserably when the data is skewed*" (He and Ma 2013). Indeed, the more imbalanced the dataset, the harder it is for the model not to be wrong in predicting the minority class. The model will tend, according to classical performance measures, to classify all individuals in the majority class.

Table 2 Confusion/cost matrix

	Predicted fraudulent: $\hat{y} = 1$	Predicted legitimate: $\hat{y} = 0$
True fraudulent: $y = 1$	TP $(C(1\|1) = c_f)$	FN $(C(0\|1) = A_i)$
True legitimate: $y = 0$	FP $(C(1\|0) = c_f)$	TN $(C(0\|0) = 0)$

Table 3 Evaluation metrics for supervised and unsupervised learning

Type	Measure	Formula
Panel A—Supervised learning		
Panel A.1—Statistical metrics		
Threshold metrics	**Sensitivity** or **true positive rate (TPR)** or **recall** or **hit rate** describes how good the model is at predicting the positive class when the actual outcome is positive, or how many of the fraudsters are correctly labeled by the model as a fraudster.	$Sensitivity = Recall = Hit\ rate = \frac{TP}{TP+FN}$
	Precision indicates how many of the predicted fraudsters are actually fraudsters.	$Precision = \frac{TP}{TP+FP}$
	F-measure (F1) focuses on the analysis of positive class by combining precision and recall measures.	$F1 = 2 \times (Precison \times Recall)/(Precision + Recall)$
	False discovery rate (FDR) is the fraction of non-fraudsters incorrectly predicted as fraudsters out of all predicted fraudsters.	$FDR = \frac{FP}{TP+FP}$
	False negative rate (FNR) is the fraction of fraudsters incorrectly predicted as non-fraudsters out of all actual fraudsters.	$FNR = \frac{FN}{TP+FN}$
Ranking metrics	**AUC-ROC** (Bradley 1997; Hanley and McNeil 1982) represents the area under the ROC (receiver operating characteristic) curve, which plots the sensitivity versus 1-specificity. Compared to the previous measures, it is not dependent on the cut-off evaluating hence the overall discriminatory performance of a model or classifier. AUC-ROC is very close to 1 for a better model.	
	AUC-PR represents the area under the precision–recall curve.	
Probabilistic metrics	**Log-loss** or **cross-entropy** is a measure of dissimilarity between the predicted probability of fraud for individual i (π_i) and the actual label $(y_i \in \{1 \text{ if default, } 0 \text{ otherwise}\})$. The smaller the better.	$CE = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i))$
	Brier score (BS) is a measure of dissimilarity between the predicted probability of fraud and the actual label (similar to the mean square error in the regression analysis), and it is bounded between 0 and 1. Lower values indicate better discrimination.	$BS = \frac{1}{N}\sum_{i=1}^{N}(\pi_i - y_i)^2$
	Brier skill score (BSS) The differences in Brier score for different classifiers can be very small. In order to address this problem, the score can be scaled against a reference score, such as the score from a no skill classifier. A zero BSS represents no skill, worse than no skill results in BSS negative, and the perfect skill is represented by 1.	$BSS = 1 - \frac{Brier\,Score}{Brier\,Score_{ref}}$

(continued)

Table 3 (continued)

Type	Measure	Formula
Panel A.2—Economic metrics		
Cost metrics	The total **Cost** is the sum of the amounts of the undetected fraudulent transactions plus the administrative cost incurred (with c_f the investigation cost and A_i the claim amount) (Bahnsen et al. 2016).	$Cost = \sum_{i=1}^{N} (y_i(1-\hat{y}_i)A_i + \hat{y}_i c_f)$
	Average expected cost (AEC) is independent of any threshold value because it relies on the estimated probabilities, π_i, instead of the predicted classes, $\hat{y}_i$ (Höppner et al. 2022).	$AEC = \frac{1}{N} \sum_{i=1}^{N} (y_i(1-\pi_i)A_i + \pi_i c_f)$
Profit metrics	The **Savings** metric is the sum of the amounts of detected fraudulent transactions minus the administrative costs incurred in detecting them, divided by the total amount of the fraudulent transactions (or the cost of using an algorithm that classifies all claims as honest) (Höppner et al. 2022).	$Savings = \frac{\sum_{i=1}^{N} (y_i \hat{y}_i A_i - \hat{y}_i c_f)}{\sum_{i=1}^{N} y_i A_i}$
	Expected Savings metric is similarly computed to savings, but based on the predicted probability of fraud (Höppner et al. 2022).	$Expected\ Savings = \frac{\sum_{i=1}^{N} (y_i \pi_i A_i - \pi_i c_f)}{\sum_{i=1}^{N} y_i A_i}$
Panel B—Unsupervised learning		
	Silhouette score (Ss) measures the similarity of data within the cluster compared to the other cluster. A value of 1 indicates that each cluster is entirely distinct from the others, while a value of -1 suggests that all the data has been incorrectly assigned to clusters. A value of 0 signifies the absence of any meaningful clusters within the data. Ss is computed using intra- and intercluster distances (a_i and b_i, respectively).	$Ss_i = \frac{b_i - a_i}{max(a_i, b_i)}$
	Calinski–Harabasz Index (I_{CH}) (also known as the variance ratio criterion) quantifies the separation between clusters by considering the ratio of between-cluster dispersion (B) to within-cluster dispersion (W). It ranges from 0 to infinity, the higher the score, the better the performance. Consider k the number of clusters and N the sample size.	$I_{CH} = \frac{B}{W} \frac{(N-k)}{(k-1)}$
	Davies–Bouldin Index (I_{DB}) is the average of the maximum ratio between the distance from a point to the center of its group and the distance between two group centers. We denote $\bar{\delta}_k$ the average distance between each point and the center of its cluster (μ_k). I_{DB} ranges from 0 to infinity, with a lower value indicating better clustering.	$I_{DB} = \frac{1}{K} \sum_{k=1}^{K} \underset{k' \neq k}{max} (\frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(\mu_k, \mu_{k'})})$

Table 4 Descriptive statistics and evolution of fraud

Year	2017	2018	2019	2020	Q1–Q3 2021	All
Number of claims	32,487	32,399	32,300	25,733	19,213	142,132
Fraud rate	0.70%	0.81%	0.82%	0.84%	0.79%	0.79%
% of fraud in the total amount of claims	6.13%	0.58%	0.68%	0.97%	0.51%	1.97%

4 Empirical Illustration

4.1 Dataset

The empirical application is based on an original proprietary dataset on automobile insurance claims provided by a French insurance company. The database contains exclusive information on collision claims, covering the time period from January 2017 to August 2021. We dispose of 142,132 observations, out of which 1,121 cases were confirmed as fraud (0.79%). The data before 2021, which includes 122,923 observations with 970 fraud cases, is used as training set, while the data after 2021, which includes 11,478 observations with 92 fraud cases, is used as test set. The study considers 67 explanatory variables that are related to insurance contracts, claim statements, or police reports. They contain information on the client (e.g., policy numbers, history of previous claims, etc.), the claim (e.g., place, date, coverage, etc.), and the car (e.g., model, number of km, etc.).[13]

As we will illustrate thereafter, we employ a combination of data analysis, machine learning, and model supervision to explain and detect the deliberate nature and the temporal evolution of fraud over time. As shown by Baesens et al. (2015), these are dominant features that relate to various forms of fraudulent activity, including insurance fraud. First, we conduct thorough analyses of historical data. For instance, Table 4 shows the temporal evolution of claims volume over the period of analysis, along with the fraud rate and the percentage of the fraud amount in the total claims. As expected, we observe a reduction in the total number of claims during 2020 (the year with two lockdowns in France due to COVID-19). Nevertheless, the fraud rate and the relative amount of fraud increased, which could be explained by reduced monitoring and law enforcement during lockdown (e.g., drivers involved in car accidents during the pandemic may have considered themselves less likely to be caught for insurance fraud because of reduced surveillance). Overall, the fraud rate is relatively stable over the period under analysis, and the percentage of fraud amount as well (except for 2017, when the insurance company had to deal with two isolated cases of extreme fraud).

Second, we incorporate time-related features into our models. For example, we investigate how fraudulent behavior varies on weekdays versus weekends or during specific seasons. Third, the machine learning models used are specifically designed to learn from historical data, enabling them to adapt to changing fraud patterns as they assimilate new data and adjust their predictions accordingly. Lastly, in a practical fraud detection system, we emphasize the importance of regular updates of the fraud detection methods. For example, if the analyst observes a decline in model accuracy or an increase in false positives, these may be early indicators of evolving fraud behavior, prompting him to adjust the models accordingly.

4.2 Technical Tools and Materials

In the remainder of the chapter, we place ourselves within the framework of supervised learning for fraud detection and proceed to investigate: (i) the gains in predictive performance achieved by using machine learning methods over standard econometric methods; (ii) the usefulness of resampling methods used to deal with the skewed class distribution; and (iii) the financial benefits of using cost-sensitive versus cost-insensitive approaches.

[13] It is important to emphasize that the use of proprietary data must be handled with great care and in accordance with ethical standards. We have treated data with confidentiality, and all analyses have been conducted in close collaboration with the insurance company. For confidentiality reasons, all results are presented in relative terms, and the names of the variables used are not disclosed.

To illustrate the aforementioned points, we use logistic Lasso-based regression model, random forest, and XGBoost algorithms, as well as their cost-sensitive counterparts.[14] Let us consider $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$ the training set with input vectors $\boldsymbol{x}_i \in \mathbf{R}^p$ and target labels $y_i \in \{0, 1\}$, with $y_i = 1$ when claim i is fraudulent, and $y_i = 0$ otherwise. A binary classification model provides a probability score $\pi_i : \pi(\boldsymbol{x}_i) \in [0, 1]$ for each instance i based on the features $\boldsymbol{x}_i$. According to a classification threshold, π_i is then converted into a predicted class $\hat{y} \in \{0, 1\}$.

Logistic Lasso Regression and Cost-Sensitive Logistic Regression

Logistic regression (Cox 1958) is a commonly utilized model across a wide range of fields, in which the binary response is represented by the probability of the response success:

$$\mathbf{P}(y_i = 1|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}}}, \tag{1}$$

with $\boldsymbol{\beta}$ the column vector of the regression coefficients. The parameter estimates are obtained by maximizing the log-likelihood function:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}[y_i log(\pi_i) + (1 - y_i)log(1 - \pi_i)] = \sum_{i=1}^{N}[y_i\boldsymbol{x}_i\boldsymbol{\beta} - log(1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}})]. \tag{2}$$

When the number of inputs is important, fitting logistic models becomes challenging. Hopefully, several procedures can be used to select the best set of explanatory variables. In this chapter we opt for the Lasso technique (Tibshirani 1996). Originally proposed for linear regression models, Lasso has become nowadays a popular selection and shrinkage estimation method in many other settings. The logistic Lasso estimator depends on the choice of a tuning parameter $\lambda \geq 0$ to be determined separately. The estimators are thus obtained by maximizing a penalized version of the log-likelihood function taking the following form:

$$l_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{N}[y_i\boldsymbol{x}_i\boldsymbol{\beta} - log(1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}})] - \lambda\sum_{j=1}^{p}|\beta_j|. \tag{3}$$

When λ is sufficiently large, the method forces some of the coefficient estimates to be exactly equal to zero. The final model involves only a subset of predictors, improving thus its interpretability.

Let us focus now on the instance-dependent **Cost-Sensitive Logistic (CSLogit)** model with Lasso regularization as proposed by Bahnsen et al. (2014a) and Höppner et al. (2022). It is worthy to note that, omitting the regularization term, the objective function in the logistic Lasso regression assigns the same weight to both false positives and false negatives, and this is not the case in many real-world applications. Nonetheless, as suggested by Höppner et al. (2022), the weight assigned to each observation can be associated to its relative cost of misclassification. The authors propose thus

[14] In selecting classifiers, we decided to choose models that have demonstrated their effectiveness in comparable scenarios. We did, however, explore also deep learning methods, notably artificial neural networks (ANNs), but their performance proved mediocre. These results are in line with those of similar research. For example, studies by Banulescu-Radu and Yankol-Schalck (2023), Yankol-Schalck (2022), Gunnarsson et al. (2021) have also indicated that deep learning algorithms may not be suitable models for tasks such as credit scoring and fraud detection, generally yielding inferior predictive performance.

to *minimize* instead the financial losses due to fraud, which can be included in the CSLogit Lasso objective function as follows:[15]

$$l_{\lambda}^{c}(\boldsymbol{\beta}) = AEC(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4}$$

$AEC(.)$ corresponds here to the average expected cost of using a classification model of score $\pi(.)$ on a set $\mathcal{D}$, based on the conditional expected value of $\hat{y}$ (i.e., $\mathbf{E}(\hat{y}|\boldsymbol{x}_i) = \mathbf{P}(\hat{y}_i = 1|\boldsymbol{x}_i) \approx \pi(\boldsymbol{x}_i)$), and given by

$$AEC(\pi(\mathcal{D})) = \frac{1}{N}\mathbf{E}[Cost(\pi(\mathcal{D}))|\boldsymbol{X}] = \frac{1}{N}\sum_{i=1}^{N}[y_i(1-\pi_i)A_i + \pi_i c_f], \tag{5}$$

where $Cost(.)$ is the cost of using a classification model of score $\pi(.)$ on a set $\mathcal{D}$ defined in Bahnsen et al. (2016) as

$$\begin{aligned} Cost(\pi(\mathcal{D})) &= \sum_{i=1}^{N}(y_i[\hat{y}_i C_i(1|1) + (1-\hat{y}_i)C_i(0|1)] \\ &\quad + (1-y_i)[\hat{y}_i C_i(1|0) + (1-\hat{y}_i)C_i(0|0)]) \\ &= \sum_{i=1}^{N}[y_i(1-\hat{y}_i)A_i + \hat{y}_i c_f]. \end{aligned} \tag{6}$$

Based on the previous cost definition, Bahnsen et al. (2016) also introduce the Savings and Expected Savings metrics presented in Table 3. The Savings correspond to the relative improvement of the cost of the newly adopted model (i.e., $Cost(\pi(\mathcal{D}))$), with respect to the cost of using a model that assigns all instances to a single class ($Cost_r(\pi(\mathcal{D})) = min\{Cost(\pi_0(\mathcal{D})), Cost(\pi_1(\mathcal{D}))\}$, with π_0 a classifier that predicts all claims in $\mathcal{D}$ as honest and π_1 as frauds):

$$\begin{aligned} Savings(\pi(\mathcal{D})) &= 1 - \frac{Cost(\pi(\mathcal{D}))}{Cost_r(\pi(\mathcal{D}))} \\ &= \frac{\sum_{i=1}^{N}(y_i\hat{y}_i A_i - \hat{y}_i c_f)}{\sum_{i=1}^{N} y_i A_i}. \end{aligned} \tag{7}$$

Expected Savings are similarly computed, but based on the predicted probability of fraud:

$$Expected\ Savings(\pi(\mathcal{D})) = \frac{\sum_{i=1}^{N}(y_i\pi_i A_i - \pi_i c_f)}{\sum_{i=1}^{N} y_i A_i}. \tag{8}$$

[15] The regularization parameter λ cannot be directly estimated from data and must be determined using hyperparameter optimization strategies. As in Höppner et al. (2022), we use a combination of grid search and cross-validation to find the optimal value. As for the CSLogit algorithm, it uses the gradient-based optimization method from Kraft (1994) to sequentially minimize the regularized average expected cost.

Random Forest and Cost-Sensitive Random Forest

Random forest (RF) method, originally introduced by Breiman (2001), aims to enhance the decision tree method by combining the concepts of decision trees and bagging to create a forest of decision trees. During the training phase, the random forest algorithm generates a large number of decision trees (or base classifiers) (T). Each base classifier ($B(\mathcal{D}_j)$) is built from a random subset ($\mathcal{D}_j$) of the training data ($\mathcal{D}$) and employs a random subset of features (m). To simplify, we denote $B_j \equiv B(\mathcal{D}_j)$ for $j = 1, \ldots, T$, and $\mathcal{B} = \{B_j\}_{j=1}^{T}$ the set of base classifiers. This randomness contributes to tree decorrelation and reduces overfitting, making the ensemble more robust and accurate. To achieve optimal performance, several hyperparameters need to be carefully adjusted, including the number of trees, the maximum depth of the trees, and the number of features considered at each split.[16]

The prediction for a new observation y_n is given by majority voting, which consists of gathering the predictions of each base classifier and selecting the decision with the highest number of votes, as following:

$$f_{y_n}(\mathcal{D}, \mathcal{B}) = \underset{c \in \{0,1\}}{argmax} \sum_{j=1}^{T} \mathbf{1}_c(B_j(\mathcal{D})). \tag{9}$$

In the case of **cost-sensitive random forest (CSRF)**, the base classifiers are combined using cost-sensitive weighted voting, which supposes giving a weight ω_j to each classifier B_j during the voting phase as follows:

$$f_{y_n}(\mathcal{D}, \mathcal{B}, \omega) = \underset{c \in \{0,1\}}{argmax} \sum_{j=1}^{T} \omega_j \mathbf{1}_c(B_j(\mathcal{D})), \tag{10}$$

with $\omega = \{\omega_j\}_{j=1}^{T}$. The weight is hence related to the performance of each classifier and defined as the normalized $Savings$ of the base classifier B_j in the out of bag set $\mathcal{D}_j^{oob} = \mathcal{D} - \mathcal{D}_j$(Bahnsen et al. 2015b). Using 7, we introduce

$$\omega_j = \frac{Savings(B_j(\mathcal{D}_j^{oob}))}{\sum_{j=1}^{T} Savings(B_j(\mathcal{D}_j^{oob}))}. \tag{11}$$

This implies that the base classifiers contributing to a higher increase in savings hold greater importance within the ensemble.[17]

eXtreme Gradient Boosting and Cost-Sensitive Boosting

Going further, **eXtreme Gradient Boosting (XGBoost)** (Chen and Guestrin 2016) is a machine learning framework that uses gradient boosting (Friedman 2001) to iteratively improve the accuracy of predictions. The algorithm is highly optimized and designed for efficiency, flexibility, and portability. It is an extension of the gradient boosting machine algorithm with significant improvements, making it a popular choice for machine learning and data mining challenges in various fields.[18]

Let us consider $\hat{y}_i$ the predicted value of the entry i, as defined by XGBoost:

[16] For more technical details on the main steps of the random forest method, see Hastie et al. (2009).

[17] The source code is publicly available as part of the https://github.com/albahnsen/CostSensitiveClassification library.

[18] Boosting is an algorithm that combines multiple weak learners into a stronger learner and corrects the errors made by the existing models. New learners are trained on the errors of the previous ones in order to increase the predictive power.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \ \ f_k \in \mathcal{F}, \tag{12}$$

with f_k an independent tree in the space of regression trees $\mathcal{F}$, and $f_k(\boldsymbol{x}_i)$ the predicted score given by the k-th tree of the i-th sample. To learn the set of functions f_k, XGBoost minimizes the following regularized objective function:

$$\mathcal{L} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \tag{13}$$

with

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||\omega||^2. \tag{14}$$

$l(.)$ herein is a training loss function, which measures the difference between the predictive value $\hat{y}_i$ and the target y_i. The second term, $\Omega(.)$, is the penalization factor, which aims at avoiding the overfitting of the model. γ and λ are two regularization parameters, and T and ω are the numbers of leaves and the score of each leaf, respectively. The tree ensemble model is then trained in an additive manner by greedily adding for each instance i at each iteration t the function of the regression tree, f_t, that mostly improves the previous prediction $\hat{y}_i^{(t-1)}$:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)) + \Omega(f_t). \tag{15}$$

After using the second-order Taylor expansion and removing the constant term, the objective function to minimize becomes

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{N} [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \Omega(f_t), \tag{16}$$

where g_i and h_i stand for the first- and second-order gradients of the loss function $l(.)$.[19]

In order to further achieve an algorithm for instance-dependent **Cost-Sensitive gradient tree Boosting (CSBoost)**, Höppner et al. (2022) substitute the gradient and second-order gradient of the Average Expected Cost seen previously.[20]

Bayes Minimum Risk

The cost-sensitive approaches outlined above involve modifying the learning algorithm to accommodate the costs associated with different decisions. Another option is to use the costs post-estimation, as basis for determining the threshold employed to convert estimated probabilities into binary outcomes (either fraudulent or non-fraudulent cases). The **Bayes Minimum Risk (BMR)** classifier (Ghosh et al. 2006) is such a decision support tool that assesses the trade-offs between different decisions by considering estimated probabilities (derived from different models) and the costs associated to each

[19] XGBoost is implemented as an open-source package, available in popular languages such as Python, R, Julia, C++, Scala. See https://github.com/dmlc/xgboost.

[20] The CSBoost and CSLogit algorithms are included in the R package *csboost* and *cslogit*, respectively, and available at github.com/SebastiaanHoppner/CostSensitiveLearning.

decision. Considering the probability estimates π_i (regardless of the algorithm used to calculate them), we define the risk associated with predicting a claim as legitimate as $R_i(0|x_i) = C_i(0|0)(1-\pi_i) + C_i(0|1)\pi_i$, while the risk when predicting it as fraudulent is $R_i(1|x_i) = C_i(1|1)\pi_i + C_i(1|0)(1-\pi_i)$. If $R_i(0|x_i) \leq R_i(1|x_i)$, then the example i is classified as legitimate (see Bahnsen et al. 2013, 2014b). In the subsequent, BMR is adapted to suit the three classification algorithms, namely logistic regression, random forest, and XGBoost.

4.3 Experimental Framework Design

The empirical analysis was conducted in several steps summarized in Fig. 2:

(a) Data Pre-processing and Feature Engineering It involved data cleaning (for inconsistencies and missing values) and transformation (e.g., standardization of variables, categorization, etc.), as well as the creation of new variables to better distinguish between honest and fraudulent behavior.

(b) Splitting of Data The period of analysis was divided on a period of training (before 2021, 86% of observations) and a period of test (starting from January 2021, 14% of observations). After verification, this division led us to a quasi-stratified splitting with about 0.8% of fraud rate in the two datasets.

(c) Model Estimation and Evaluation We trained logistic Lasso, random forest, and XGBoost classifiers, as well as their cost-sensitive counterparts (CSLogit Lasso, CS-RF, and CSBoost), on the training set and evaluated their performances on the test set of data. To find the optimal hyper-

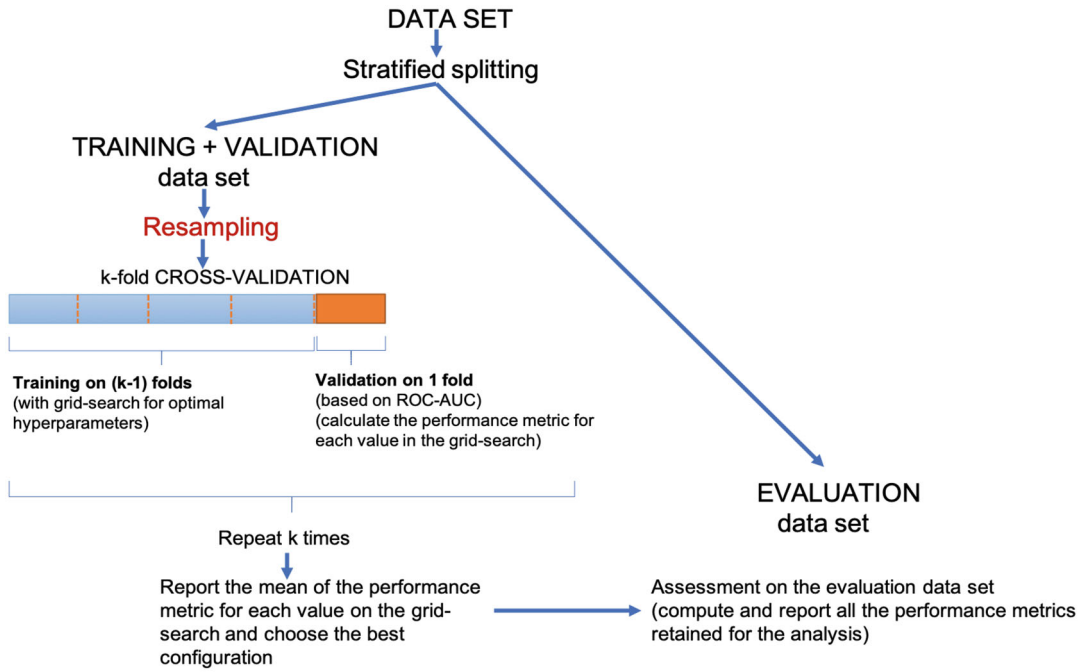


Fig. 2 Experimental framework design

parameters (e.g., λ for logistic Lasso, etc.), a four-fold cross-validation approach was implemented during the training. The performance metrics used to evaluate the models are: ROC-AUC, ROC-PR, log-loss, BSS, F-measure, FDR, FNR, Savings, and Expected savings (see Table 3 for their formulas).

(d) Dealing with Imbalanced Data Given the actual imbalance in data (0.08%/99.02%), we applied four resampling techniques, at different resampling rates (e.g., 15%/85%, 10%/90%, 5%/95%) on the training set:[21]

- *Random undersampling*, an undersampling method that works by randomly selecting a subset of the majority class with a number of instances equal or close to that of the minority class.
- *Synthetic Minority Oversampling TEchnique (SMOTE)* (Chawla et al. 2002), one of the most popular and successful oversampling methods. It selects at random one or more nearest neighbors of a minority class instance and produces new instances based on the linear interpolations between them.
- *ADAptive SYNthetic sampling* (ADASYN) (He et al. 2008), an improved version of SMOTE that creates a number of synthetic data based on the assumption that minority class instances that are close to majority class instances are more difficult to classify than those that are further away.
- *SMOTE-Tomek* Batista et al. (2004), a hybrid method that combines both SMOTE and Tomek methods. Thus, it is able to oversample the minority class with synthetic data while removing instances of the majority class that are misclassified.

Logistic Lasso regression, random forest, and XGBoost were then trained on this new less skewed data and evaluated, as previously, on the unmodified test dataset. When applying different resampling methods in the context of imbalanced data, the estimated probabilities may be biased or poorly calibrated. There is no problem if the fraud analyst is only interested in ranking the observations in terms of their fraud risk, but if well calibrated fraud probabilities are needed (for instance for the computation of probabilistic performance metrics such as log-loss or Brier score), then the posterior probabilities need to be adjusted. We opted for the calibration proposed by Baesens et al. (2015):

$$\pi_c(G_i|\boldsymbol{X}) = \frac{\frac{p(G_i)}{p_r(G_i)}\pi_r(G_i|\boldsymbol{X})}{\sum_{j=1}^{2}\frac{p(G_j)}{p_r(G_j)}\pi_r(G_j|\boldsymbol{X})},$$

with G_i the class/group i (e.g., class 1 for fraudsters and 2 for non-fraudsters), $p(.)$ the prior probability,[22] $p_r(.)$ the resampled prior probability due to the resampling procedure,[23] and $\pi_r(.|\boldsymbol{X})$ the posterior probability for observation i as calculated by the machine learning technique using the resampled data.

As for the cost-sensitive methods, they already address the problem of class imbalance through the cost-weighted system adopted, so there is no interest to reestimate them on the resampled data.

[21] Resampling techniques can be grouped into three families: (i) Undersampling method: A subset of the original dataset is created by deleting instances; (ii) Oversampling methods: A superset of the original dataset is created by replicating some instances or creating new instances from the existing ones; (iii) Hybrids methods: Combine both sampling approaches.

[22] For example, $p(G_1) = 0.79\%$ and $p(G_2) = 99.21\%$, in our case.

[23] For example, $p_r(G_1) = 20\%$ and $p_r(G_2) = 80\%$.

4.4 Main Results

As outlined previously, the empirical analysis and the interpretation of results are structured around three scenarios, each addressing a specific research question:

(a) Does the use of machine learning techniques improve the statistical predictive performances over conventional insurance fraud models?
(b) How effective are resampling techniques in addressing imbalanced datasets in the context of insurance fraud detection?
(c) What are the financial gains of employing cost-sensitive versus cost-insensitive approaches in the context of insurance fraud detection?[24]

Scenario (a) and Scenario (b) Table 5 displays the results in terms of predictive performance for logistic Lasso, random forest, and XGBoost, respectively, as well as their counterparts obtained by the reestimation of the models on rebalanced training datasets. The seven metrics used assess exclusively the statistical predictive performance of the models. Results show that XGBoost and its variants are globally the best-performing approaches (followed by random forest), combining the highest AUC, PR, and lowest entropy (log-loss). XGBoost minimizes the false negative rate, while logistic regression and random forest do better in terms of false discovery rate. Globally, for this specific dataset, the results obtained provide evidence to justify the use of machine learning algorithms over traditional econometric methods, in terms of statistical performance. As for the usefulness of resampling techniques, we observe that they might improve the performance of the models, but the degree of improvement depends on the resampling rate and the algorithm chosen. Moreover, for our dataset, the improvement of XGBoost, random forest, and logistic regression when using resampling techniques is marginal.

Scenario (c) Thereafter, we compare cost-sensitive (CSLogit, CSRF, and CSBoost) to cost-insensitive approaches (logistic Lasso, random forest, and XGBoost), in terms of both statistical and financial performances. The results are in Table 6. Financial performance is measured using specific metrics, namely Savings and Expected savings. The former is calculated based on the binary outcomes generated by each model, while the latter relies on the estimated probabilities. Thus, to compute the Savings measure, we consider two different thresholding approaches to convert predicted fraud probabilities into binary outcomes (i.e., fraud or non-fraud). These approaches include one calculated by maximizing sensitivity and specificity measures (column *Optimal cut-off Savings*), and a second one applying the Bayes Minimum Risk principle (as explained in Sect. 4.2, and referred to as *BMR cut-off Savings* in Table 6). The Cost and Average Expected Cost (AEC) have been computed as well, but they are not displayed due to confidentiality concerns.

In terms of economic performance, CSBoost displays the best results in terms of expected savings, followed by XGBoost and random forest. The same trend is observed in terms of Average Expected Cost (AEC) (i.e., CSBoost exhibits the lowest AEC, followed by XGBoost). The average difference in AEC between XGBoost and CSBoost is 5.94€, while the difference between logistic Lasso and CSLogit is less than 1€. Regarding the performance metrics that are influenced by the thresholding applied to transform estimated fraud probabilities into binary decisions (namely Savings and Cost), the top-performing model is XGBoost-BMR, with BMR-based random forest and logistic models following behind. It is worth noting that the groups of models based on the statistical optimal

[24] It is important to note that the answers to these questions are specific to the dataset used in the analysis and may not be generalizable to other contexts.

Table 5 Results for logistic Lasso regression, random forest, and XGBoost and their counterparts on resampled data

Model	AUC	PR	Log-loss	BSS	F1	FDR	FNR
Logistic Lasso	87.31%	0.20	0.08	−0.42	0.96	85.14%	21.19%
RandomUnderSampler_LR_5%	87.32%	0.20	0.07	−0.17	0.96	85.25%	21.85%
RandomUnderSampler_LR_10%	87.32%	0.20	0.07	−0.19	0.96	85.14%	21.19%
RandomUnderSampler_LR_15%	87.33%	0.20	0.07	−0.13	0.96	85.14%	21.19%
ADASYN_LR_5%	87.32%	0.20	0.06	−0.06	0.96	85.14%	21.19%
ADASYN_LR_10%	87.33%	0.20	0.06	−0.11	0.96	85.14%	21.19%
ADASYN_LR_15%	87.32%	0.20	0.07	−0.13	0.96	85.16%	21.85%
SMOTE_LR_5%	87.33%	0.20	0.05	−0.03	0.96	85.14%	21.19%
SMOTE_LR_10%	87.33%	0.20	0.06	−0.05	0.96	85.14%	21.19%
SMOTE_LR_15%	87.35%	0.21	0.05	−0.02	0.96	85.25%	21.85%
SMOTE_TomekLinks_LR_5%	87.34%	0.21	0.05	0.02	0.96	85.25%	21.85%
SMOTE_TomekLinks_LR_10%	87.36%	0.21	0.05	0.02	0.96	85.25%	21.85%
SMOTE_TomekLinks_LR_15%	87.35%	0.21	0.05	0.00	0.96	85.14%	21.19%
Random forest	93.14%	0.57	**0.02**	0.38	**0.97**	**83.26%**	23.18%
RandomUnderSampler_RF_5%	93.11%	0.58	**0.02**	0.28	0.89	94.06%	13.91%
RandomUnderSampler_RF_10%	93.39%	0.57	0.03	0.20	0.87	95.05%	11.26%
RandomUnderSampler_RF_15%	93.52%	0.54	**0.02**	0.30	0.88	94.49%	11.26%
SMOTE_RF_5%	92.78%	0.59	0.03	0.23	0.92	92.41%	12.58%
SMOTE_RF_10%	93.01%	0.57	0.03	0.14	0.93	91.66%	15.89%
SMOTE_RF_15%	92.03%	0.56	0.03	0.11	0.95	88.03%	19.21%
ADASYN_RF_5%	93.76%	0.59	0.03	0.22	0.92	92.21%	14.57%
ADASYN_RF_10%	93.25%	0.58	0.03	0.13	0.91	93.27%	14.57%
ADASYN_RF_15%	92.87%	0.58	0.03	0.09	0.95	88.14%	19.21%
SMOTE_TomekLinks_RF_5%	92.60%	0.58	0.03	0.17	0.91	92.97%	15.89%
SMOTE_TomekLinks_RF_10%	93.38%	0.58	0.03	0.15	0.95	89.18%	19.21%
SMOTE_TomekLinks_RF_15%	92.58%	0.59	0.03	0.08	0.93	91.54%	16.56%
XGBoost	93.67%	0.56	**0.02**	**0.45**	0.84	95.59%	**7.95%**
RandomUnderSampler_XGB_5%	**93.77%**	0.55	**0.02**	0.44	0.90	93.78%	14.57%
RandomUnderSampler_XGB_10%	92.13%	0.52	0.04	0.21	0.96	87.18%	23.84%
RandomUnderSampler_XGB_15%	93.14%	**0.60**	0.03	0.17	0.83	96.03%	9.27%
SMOTE_XGB_5%	93.16%	0.56	0.03	0.29	0.90	93.91%	15.89%
SMOTE_XGB_10%	93.04%	0.53	0.03	0.21	0.93	91.19%	20.53%
SMOTE_XGB_15%	92.92%	0.55	0.04	0.18	0.94	90.47%	20.53%
ADASYN_XGB_5%	92.55%	0.55	0.03	0.29	0.87	95.07%	13.91%
ADASYN_XGB_10%	92.95%	0.53	0.03	0.21	0.93	90.99%	17.22%
ADASYN_XGB_15%	92.02%	0.54	0.04	0.16	0.93	91.56%	18.54%
SMOTE_TomekLinks_XGB_5%	92.97%	0.57	0.03	0.30	0.94	90.15%	19.21%
SMOTE_TomekLinks_XGB_10%	92.83%	0.56	0.03	0.21	0.92	92.36%	18.54%
SMOTE_TomekLinks_XGB_15%	93.05%	0.55	0.04	0.18	0.93	91.67%	17.22%

thresholding method display comparatively reduced financial efficiency, especially when applied to cost-insensitive algorithms and even more for cost-sensitive ones.

As for the statistical predictive performance, the results show that cost-insensitive methodologies outperform their cost-sensitive counterparts in terms of statistical performance as measured by AUC,

Table 6 Cost-sensitive versus cost-insensitive approaches

Model	AUC	PR	Log-loss	BSS	Expected savings	Optimal cut-off Savings	BMR cut-off Savings
Logistic regression	87.31%	0.20	0.08	−0.42	6.09%	56.24%	73.84%
Random forest	93.14%	0.57	0.02	0.38	20.46%	56.42%	73.89%
XGBoost	**93.69%**	**0.57**	**0.02**	0.44	31.08%	**69.65%**	**77.39%**
CSLogit	87.02%	0.06	0.04	0.04	3.58%	64.28%	63.43%
CSRF	82.23%	0.04	0.04	0.02	1.06%	52.59%	62.17%
CSBoost	89.60%	0.24	0.09	−1.69	**49.81%**	52.92%	57.09%

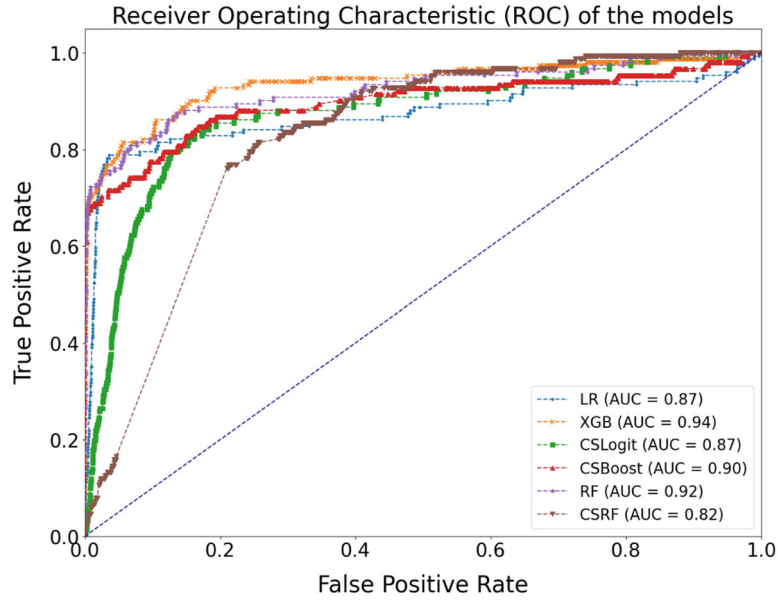


Fig. 3 AUC-ROC curves for logistic Lasso, random forest, XGBoost, and their cost-sensitive counterparts

PR, entropy, and BSS (with XGBoost performing the best again). The ROC-AUC of the four models is represented graphically in Fig. 3. To statistically assess the differences between them, we use the statistical test proposed by DeLong et al. (1988). Widely adopted for its simplicity of implementation, this approach tests under the null hypothesis the equality of two or more correlated areas under ROC curves. The results of the test are available in Table 7 and show that there are not statistical significant differences between CSLogit and logistic Lasso, logistic Lasso and CSBoost, random forest and CSBoost, CSLogit and CSBoost, as well as between CSBoost and XGBoost, in terms of AUC-ROC. However, XGBoost outperforms significantly logistic Lasso, random forest, and their cost-sensitive counterparts.

The main findings of this study can be summarized as follows. Firstly, machine learning algorithms demonstrate superior statistical predictive performance. Secondly, when considering financial efficiency measures that directly use estimated probabilities (such as Expected Savings and Average Expected Cost), the best-performing models are those using cost-sensitive machine learning approaches, CSBoost being the best. However, if the aim is to use binary outcomes decisions generated by different models, it is advisable to favor the BMR thresholding approach in combination

Table 7 P-values of DeLong et al. (1988) test of model comparison in terms of AUC-ROC

	Logistic regression	Random forest	XGBoost	CSLogit	CSRF	CSBoost
Logistic regression	1.00					
Random forest	0.00	1.00				
XGBoost	0.00	0.03	1.00			
CSLogit	0.87	0.00	0.00	1.00		
CSRF	0.00	0.00	0.00	0.00	1.00	
CSBoost	0.17	0.48	0.17	0.14	0.00	1.00

with cost-insensitive models. Finally, it can be affirmed that the machine learning algorithms used in this illustration outperform conventional econometric models, both statistically and financially.

5 Conclusion

This chapter provides an overview of the main models used for insurance fraud prevention and detection, encompassing supervised, unsupervised, and social network methods, along with the main challenges that modelers face in this context. The chapter's emphasis is two-fold. Firstly, it assesses the improvements in statistical predictive performance achieved through the use of machine learning models compared with traditional econometric models. Secondly, the chapter assesses the financial benefits associated with the transition from conventional methodologies to cost-sensitive approaches.

The inclusion of different methods allows for performance comparison and can aid in determining the most effective approach for a given dataset. It is important to note that the success of the machine learning techniques for detecting insurance fraud is dependent on the quality of the data, as well as on the specific context of the insurance industry. Furthermore, just as standard approaches, these techniques are not perfect, and they may also not detect all instances of fraud and could generate false positives, leading to costs for the insurance company and delays for legitimate claims. However, results show that machine learning classifiers perform better (both statistically and financially) than standard econometric approaches.

Moreover, machine learning algorithms can be used to assess fraud risks in real time, which reduces the time it takes to process claims and improves customer satisfaction. They can be adapted to cope with constantly changing fraud situations, identifying new fraud patterns and incorporating them into fraud detection models. Finally, paired with expert knowledge and human assessment to improve their effectiveness, machine learning algorithms can help insurance companies to improve the efficiency of their fraud detection process, reduce financial losses, and protect honest policyholders from fraudulent behavior.

Summarizing the main findings of this study, several points are worth highlighting. Firstly, our research demonstrates the superior predictive statistical performance of machine learning algorithms over conventional econometric models. Secondly, when using threshold-insensitive measures to assess the financial efficiency, such as Expected Savings and Average Expected Cost, we find that the best-performing models are those using cost-sensitive methodologies, with CSBoost emerging as the most effective among them. However, when the aim is to work with binary outcomes decisions produced by different models, it is advisable to give priority to the BMR thresholding approach in conjunction with cost-insensitive models. Finally, it should be noted that these results are specific to our database and that, in further research, it would be interesting to explore their generalization using other databases. This future exploration could provide valuable insights into the wider applicability of our results in various insurance fraud detection scenarios.

References

Adewumi AO, Akinyelu AA (2017) A survey of machine-learning and nature-inspired based credit card fraud detection techniques. Int J Syst Assurance Eng Manag 8:937–953

Ai J, Brockett PL, Golden LL (2009) Assessing consumer fraud risk in insurance claims: an unsupervised learning technique using discrete and continuous predictor variables. North Am Actuarial J 13(4):438–458

Ai J, Brockett PL, Golden LL et al (2013) A robust unsupervised method for fraud rate estimation. J Risk Insur 80(1):121–143

Amponsah AA, Adekoya AF, Weyori BA (2022) A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology. Decis Anal J 4:100122

Artís M, Ayuso M, Guillén M (2002) Detection of automobile insurance fraud with discrete choice models and misclassified claims. J Risk Insur 69(3):325–340

Artıs M, Ayuso M, Guillen M (1999) Modelling different types of automobile insurance fraud behaviour in the Spanish market. Insur Math Econ 24(1–2):67–81

Aslam F, Hunjra AI, Ftiti Z et al (2022) Insurance fraud detection: evidence from artificial intelligence and machine learning. Res Int Bus Financ 62:101744

Atwood JA, Robison-Cox JF, Shaik S (2006) Estimating the prevalence and cost of yield-switching fraud in the federal crop insurance program. Am J Agric Econ 88(2):365–381

Baesens B, Setiono R, Mues C et al (2003) Using neural network rule extraction and decision tables for credit-risk evaluation. Manag Sci 49(3):312–329

Baesens B, Van Vlasselaer V, Verbeke W (2015) Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons, London

Baesens B, Höppner S, Verdonck T (2021) Data engineering for fraud detection. Decis Support Syst 150:113492

Bahnsen AC, Stojanovic A, Aouada D et al (2013) Cost sensitive credit card fraud detection using Bayes minimum risk. In: 2013 12th international conference on machine learning and applications. IEEE, Piscataway, pp 333–338

Bahnsen AC, Aouada D, Stojanovic A et al (2016) Feature engineering strategies for credit card fraud detection. Expert Syst Appl 51:134–142

Bahnsen AC, Aouada D, Ottersten B (2014a) Example-dependent cost-sensitive logistic regression for credit scoring. In: 2014 13th International conference on machine learning and applications. IEEE, Piscataway, pp 263–269

Bahnsen AC, Stojanovic A, Aouada D et al (2014b) Improving credit card fraud detection with calibrated probabilities. In: Proceedings of the 2014 SIAM international conference on data mining. SIAM, pp 677–685

Bahnsen AC, Aouada D, Ottersten B (2015a) Ensemble of example-dependent cost-sensitive decision trees. arXiv e-prints, pp arXiv–1505

Bahnsen AC, Aouada D, Ottersten B (2015b) Example-dependent cost-sensitive decision trees. Expert Syst Appl 42(19):6609–6619

Banulescu-Radu D, Yankol-Schalck M (2023) Practical guideline to efficiently detect insurance fraud in the era of machine learning: a household insurance case. Working paper

Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newslett 6(1):20–29

Bauer D, Tyler Leverty J, Schmit J et al (2021) Symposium on insure-tech, digitalization, and big-data techniques in risk management and insurance. J Risk Insur 88(3):525–528

Bayerstadler A, van Dijk L, Winter F (2016) Bayesian multinomial latent variable modeling for fraud and abuse detection in health insurance. Insur Math Econ 71:244–252

Belhadji EB, Dionne G, Tarkhani F (2000) A model for the detection of insurance fraud. Geneva Papers Risk Insur-Issues Practice 25(4):517–538

Bermúdez L, Pérez JM, Ayuso M et al (2008) A Bayesian dichotomous model with asymmetric link for fraud in insurance. Insur Math Econ 42(2):779–786

Bolton RJ, Hand DJ (2002) Statistical fraud detection: a review. Stat Sci 17(3):235–255

Bolton RJ, Hand DJ et al (2001) Unsupervised profiling methods for fraud detection. Credit scoring and credit control VII, pp 235–255

Bond EW, Crocker KJ (1997) Hardball and the soft touch: the economics of optimal insurance contracts with costly state verification and endogenous monitoring costs. J Public Econ 63(2):239–264

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brockett PL, Derrig RA, Golden LL et al (2002) Fraud classification using principal component analysis of RIDITs. J Risk Insur 69(3):341–371

Brownlee J (2020) Imbalanced classification with python: better metrics, balance skewed classes, cost-sensitive learning. machine learning mastery

Błaszczyński J, de Almeida Filho AT, Matuszyk A et al (2021) Auto loan fraud detection using dominance-based rough set approach versus machine learning methods. Expert Syst Appl 163:113740
Caldeira AM, Gassenferth W, Machado MAS et al (2015) Auditing vehicles claims using neural networks. Proc Comput Sci 55:62–71
Caudill SB, Ayuso M, Guillén M (2005) Fraud detection using a multinomial logit model with missing information. J Risk Insur 72(4):539–550
Chawla NV, Bowyer KW, Hall LO et al (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
Cox DR (1958) The regression analysis of binary sequences. J Roy Stat Soc Ser B (Methodological) 20(2):215–232
Crocker KJ, Tennyson S (1999) Costly state falsification or verification? Theory and evidence from bodily injury liability claims. In: Automobile insurance: road safety, new drivers, risks, insurance fraud and regulation, pp 119–130
Dal Pozzolo A, Caelen O, Le Borgne YA et al (2014) Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst Appl 41(10):4915–4928
DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 837–845
Derrig RA (2002) Insurance fraud. J Risk Insur 69(3):271–287
Derrig RA, Ostaszewski KM (1995) Fuzzy techniques of pattern recognition in risk and claim classification. J Risk Insur 447–482
De Vos S, Vanderschueren T, Verdonck T et al (2023) Robust instance-dependent cost-sensitive classification. In: Advances in data analysis and classification, pp 1–23
Dhieb N, Ghazzai H, Besbes H et al (2020) A secure AI-driven architecture for automated insurance systems: fraud detection and risk measurement. IEEE Access 8:58546–58558
Dionne G, Giuliano F, Picard P (2003) Optimal auditing for insurance fraud. Working paper, HEC Montréal
Dionne G, Giuliano F, Picard P (2009) Optimal auditing with scoring: theory and application to insurance fraud. Manag Sci 55(1):58–70
Elkan C (2001) The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence, Lawrence Erlbaum Associates, pp 973–978
Fernández A, García S, Galar M et al (2018) Learning from imbalanced data sets. Springer, Berlin
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 1189–1232
Ghorbani A, Farzai S (2018) Fraud detection in automobile insurance using a data mining based approach. Int J Mechatron Electr Comput Technol 8(27):3764–3771
Ghosh JK, Delampady M, Samanta T (2006) Bayesian inference and decision theory. An introduction to Bayesian analysis: theory and methods, pp 29–63
Golden LL, Brockett PL, Guillén M et al (2020) APRIDIT unsupervised classification with asymmetric valuation of variable discriminatory worth. Multivariate Behav Res 55(5):685–703
Gomes C, Jin Z, Yang H (2021) Insurance fraud detection with unsupervised deep learning. J Risk Insur 88(3):591–624
Gunnarsson BR, Vanden Broucke S, Baesens B et al (2021) Deep learning for credit scoring: do or don't? Eur J Oper Res 295(1):292–305
Hainaut D (2019) A self-organizing predictive map for non-life insurance. Eur Actuarial J 9(1):173–207
Hand DJ, Whitrow C, Adams NM et al (2008) Performance criteria for plastic card fraud detection tools. J Oper Res Soc 59(7):956–962
Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, Berlin
He H, Ma Y (2013) Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, London
He H, Bai Y, Garcia EA et al (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, Piscataway, pp 1322–1328
Höppner S, Baesens B, Verbeke W et al (2022) Instance-dependent cost-sensitive learning for detecting transfer fraud. Eur J Oper Res 297(1):291–300
Insurance Europe (2019) Insurance fraud—not a victimless crime. Booklet Insurance Europe
Insurance Europe (2023) Annual report 2022–2023. Insurance Europe
Jin Y, Rejesus RM, Little BB (2005) Binary choice models for rare events data: a crop insurance fraud application. Appl Econ 37(7):841–848
Kirlidog M, Asuk C (2012) A fraud detection approach with data mining in health insurance. Proc-Soc Behav Sci 62:989–994

Kose I, Gokturk M, Kilic K (2015) An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Appl Soft Comput 36:283–299
Kou Y, Lu CT, Sirwongwattana S et al (2004) Survey of fraud detection techniques. In: IEEE international conference on networking, sensing and control. IEEE, Piscataway, pp 749–754
Kovács G (2019) An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Appl Soft Comput 83:105662
Kowshalya G, Nandhini M (2018) Predicting fraudulent claims in automobile insurance. In: 2018 second international conference on inventive communication and computational technologies (ICICCT). IEEE, Piscataway, pp 1338–1343
Kraft D (1994) Algorithm 733: TOMP–Fortran modules for optimal control calculations. ACM Trans Math Softw 20(3):262–281
Laffont JJ, Tirole J (1993) A theory of incentives in procurement and regulation. MIT Press, Cambridge
Li Y, Yan C, Liu W et al (2018) A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Appl Soft Comput 70:1000–1009
Li H, Song Q, Su J (2021) Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. J Risk Insur 88(3):625–663
Lismont J, Ram S, Vanthienen J et al (2018) Predicting interpurchase time in a retail environment using customer-product networks: an empirical study and evaluation. Expert Syst Appl 104:22–32
Madiega TA (2021) Artificial intelligence act. European Parliament: European Parliamentary Research Service
Madiega TA (2023) Artificial intelligence act. European Parliament: European Parliamentary Research Service
Majhi SK, Bhatachharya S, Pradhan R et al (2019) Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection. J Intell Fuzzy Syst 36(3):2333–2344
Major JA, Riedinger DR (1992) EFD: a hybrid knowledge/statistical-based system for the detection of fraud. Int J Intell Syst 7(7):687–703
Nian K, Zhang H, Tayal A et al (2016) Auto insurance fraud detection using unsupervised spectral ranking for anomaly. J Financ Data Sci 2(1):58–75
Padhi S, Panigrahi S (2019) Decision templates based ensemble classifiers for automobile insurance fraud detection. In: 2019 Global Conference for Advancement in Technology (GCAT). IEEE, Piscataway, pp 1–5
Philip K, Chan S (1998) Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In: Proceeding of the fourth international conference on knowledge discovery and data mining. AAAI Press, Manchester, pp 164–168
Phua C, Alahakoon D, Lee V (2004) Minority report in fraud detection: classification of skewed data. ACM SIGKDD Explor Newslett 6(1):50–59
Picard P (1996) Auditing claims in the insurance market with fraud: the credibility issue. J Public Econ 63(1):27–56
Picard P (2013) Economic analysis of insurance fraud. Handbook of insurance, pp 349–395
Pinquet J, Ayuso M, Guillén M (2007) Selection bias and auditing policies for insurance claims. J Risk Insur 74(2):425–440
Pourhabibi T, Ong KL, Kam BH et al (2020) Fraud detection: a systematic literature review of graph-based anomaly detection approaches. Decis Support Syst 133:113303
Roy R, George KT (2017) Detecting insurance claims fraud using machine learning techniques. In: 2017 international conference on circuit, power and computing technologies (ICCPCT). IEEE, Piscataway, pp 1–6
Sahin Y, Bulkan S, Duman E (2013) A cost-sensitive decision tree approach for fraud detection. Expert Syst Appl 40(15):5916–5923
Sahni S, Mittal A, Kidwai F et al (2020) Insurance fraud identification using computer vision and IoT: a study of field fires. Proc Comput Sci 173:56–63
Settipalli L, Gangadharan G (2023) WMTDBC: an unsupervised multivariate analysis model for fraud detection in health insurance claims. Expert Syst Appl 215:119259
Severino MK, Peng Y (2021) Machine learning algorithms for fraud prediction in property insurance: empirical evidence using real-world microdata. Mach Learn Appl 5:100074
Stripling E, Baesens B, Chizi B et al (2018) Isolation-based conditional anomaly detection on mixed-attribute data to uncover workers' compensation fraud. Decis Support Syst 111:13–26
Sun C, Li Q, Li H et al (2018) Patient cluster divergence based healthcare insurance fraudster detection. IEEE Access 7:14162–14170
Sundarkumar GG, Ravi V (2015) A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Eng Appl Artif Intell 37:368–377
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodological) 58(1):267–288
Tumminello M, Consiglio A, Vassallo P et al (2023) Insurance fraud detection: a statistically validated network approach. J Risk Insur 90(2):381–419
Vanderschueren T, Baesens B, Verdonck T et al (2022) A new perspective on classification: optimally allocating limited resources to uncertain tasks. arXiv preprint arXiv:220204369

Van Vlasselaer V, Eliassi-Rad T, Akoglu L et al (2017) Gotcha! network-based fraud detection for social security fraud. Manag Sci 63(9):3090–3110
Vassiljeva K, Tepljakov A, Petlenkov E et al (2017) Computational intelligence approach for estimation of vehicle insurance risk level. In: 2017 international joint conference on neural networks (IJCNN). IEEE, Piscataway, pp 4073–4078
Viaene S, Dedene G (2004) Insurance fraud: issues and challenges. Geneva Papers Risk Insur Issues Practice 29:313–333
Viaene S, Derrig RA, Baesens B et al (2002) A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. J Risk Insur 69(3):373–421
Viaene S, Dedene G, Derrig RA (2005) Auto claim fraud detection using Bayesian learning neural networks. Expert Syst Appl 29(3):653–666
Viaene S, Ayuso M, Guillen M et al (2007) Strategies for detecting fraudulent claims in the automobile insurance industry. Eur J Oper Res 176(1):565–583
Wang Y, Xu W (2018) Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis Support Syst 105:87–95
Warren DE, Schweitzer ME (2018) When lying does not pay: how experts detect insurance fraud. J Bus Ethics 150(3):711–726
Watt R (2003) Curtailing ex-post fraud in risk sharing arrangements. Eur J Law Econ 16(2):247–263
Weisberg HI, Derrig RA (1998) Quantitative methods for detecting fraudulent automobile bodily injury claims. Risques 35(July–September):75–99
Yang WS, Hwang SY (2006) A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31(1):56–68
Yankol-Schalck M (2022) The value of cross-data set analysis for automobile insurance fraud detection. Res Int Bus Financ 63:101769
Zhang J, Vukotic I, Gardner R (2018) Anomaly detection in wide area network mesh using two machine learning anomaly detection algorithms. arXiv preprint arXiv:180110094
Zhu B, Baesens B, vanden Broucke SK (2017) An empirical comparison of techniques for the class imbalance problem in churn prediction. Inform Sci 408:84–99
Zhu B, Baesens B, Backiel A et al (2018) Benchmarking sampling techniques for imbalance learning in churn prediction. J Oper Res Soc 69(1):49–65
Óskarsdóttir M, Ahmed W, Antonio K et al (2022) Social network analytics for supervised fraud detection in insurance. Risk Anal 42(8):1872–1890
Šubelj L, Furlan Š, Bajec M (2011) An expert system for detecting automobile insurance fraud using social network analysis. Expert Syst Appl 38(1):1039–1052

Nonnegative Second-Order Semiparametric Analysis and Experience Rating in Non-life Insurance

Jean Pinquet

1 Introduction

Experience rating in nonlife insurance can be justified with two arguments:

- The first argument is actuarial neutrality. For nonlife insurance data, a claimless period usually implies a reduction in frequency risk for the next periods, whereas an accident triggers an increase in risk. Hence, bonus–malus systems (i.e., no-claim discounts and increases in premium after a claim) can be justified with an actuarial neutrality argument.
- The second argument are the incentives to risk prevention created by experience rating. There is a short-term efficiency of effort in reducing nonlife insurance risks,[1] and experience rating creates such incentives. Things are different for health and life risks. These risks are related to a capital whose depletion is partly irreversible. Prevention efforts are inefficient in the short run, and there is a reclassification risk which makes experience rating very uncommon.[2]

Automobile insurance rating is mostly deregulated today, but bonus–malus scales are still in force in various countries. These scales are discussed in Sect. 2. They consist of a finite number of levels, each with its own relativity (also termed as bonus–malus coefficient). New policyholders are assigned to a specified level of the scale. After each year, the policy moves up (if claims are reported) or down, according to transition rules. These scales do not enforce experience rating, and their incentive properties are not clear-cut. However, bonus–malus scales provide an information available to all the

I thank the editor and the referees for their comments.

[1] Risk reduction applies to frequency rather than severity in most of the economic literature. Hence prevention is of the "self-protection" rather than of the "self-insurance" type, with the Ehrlich and Becker (1972) terminology. Self-insurance is also referred to as mitigation.

[2] Hendel and Lizzeri (2003) mention however term life insurance contracts in the USA that offer state contingent prices, where low premiums are contingent on the insured showing he is still in good health.

J. Pinquet (✉)
UFR SEGMI, Université Paris Nanterre, Nanterre Cedex, France
e-mail: jpinquet@free.fr; pinquet@parisnanterre.fr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_16

competitors in the market. Reducing the information rents owned by the insurers is the main role of bonus–malus scales.

The incentives to safe driving are the main motivation for point-record driving licenses based on traffic violations. These incentives increase when the threshold of a license suspension gets closer. This is studied theoretically in Bourgeon and Picard (2007), and both theoretically and empirically in Dionne et al. (2011), Dionne et al. (2013). In Sect. 2, it is argued that current point-record driving licenses are detrimental to drivers with a high risk exposure. However, this might change in the future, if the driver's behavior was recorded in real time.

The information rent owned by an insurer on a policyholder depends on the seniority of the latter in the portfolio, but also on the memory in the data. Sect. 3 presents the bulk of the paper. It is devoted to the statistical analysis of the memory in nonlife insurance data, with a semiparametric approach. The risk distributions are partially specified from moment constraints. Second-order stationary time series are defined in a semiparametric framework, in which the length of memory is rigorously defined. In the following sections, the stationary time series are dynamic random effects, which are used to mix the risk distributions. Then, affine predictions are derived in mixture models. This approach is usually referred to as "linear credibility" if the dependent variables refer to nonlife insurance risks (e.g., the number of accidents, or their total cost: see Bühlmann and Gisler 2006). Risk variables are nonnegative in nonlife insurance, and positive properties on the autocovariance functions of the random effects are required in order to ensure the positivity of the affine predictors. This subject has been treated comprehensively by the author in recent publications (Pinquet 2020a,b, 2022b), and the main results are recalled in this chapter. The solutions include the AR(p) models with nonnegative and decreasing filtering coefficients, as well as the ARFIMA$(0, d, 0)$ specifications. These ergodic specifications have a short and a long memory, respectively. The positivity of the affine predictors is maintained if a time-invariant random effect is placed side by side with a dynamic random effect that follows an ergodic specification specified as above.

These autocovariance functions should be compatible with positive time series, as they are applied to positive random effects. This is the case for all the autocovariance functions that ensure positive affine predictors in the mixture model (see Pinquet 2022a). A case study is presented in Sect. 4, from a panel data set of automobile insurance contracts.

By increasing length of memory, we have the following ranking: bonus–malus scales, real-world rating structures, and frequency risks estimated by statistical models.

2 Bonus–Malus Scales, Real-World Rating Structures, and Point-Record Driving Licenses: Memory and Incentive Properties

2.1 Bonus–Malus Scales

Automobile insurance rating is mostly deregulated today, but bonus–malus scales are still in force in various countries (either compulsory as in France or in Italy, or at least used by most of the competitors as in Belgium: see Lemaire and Zi 1994; Lemaire 1995). Bonus–malus scales consist of a finite number of levels, each with its own relativity (also termed as bonus–malus coefficient). A bonus–malus scale summarizes an event history, where events are most often claims at fault. New policyholders are assigned to a specified level. After each year, the policy moves up (if claims are reported) or down, according to transition rules. The insurance premium is the product of the bonus–malus coefficient and of a basic premium. A bonus–malus scale enforces the experience rating policy if the basic premium does not depend on the individual history. This is not the case any more in most European countries, but the bonus–malus scale provides an information available to all the

competitors in the market. Reducing the information rents owned by the insurers is now the role of bonus–malus scales, more than enforcing experience rating rules.[3]

The informational content of the bonus–malus scale is discussed hereafter. The main characteristics of these scales are the following:

- The bonus–malus scales are unbalanced. Indeed, the number of claimless years that are necessary to offset a claim at fault ranges usually between two and five years in these scales. This must be paralleled with the average frequency of events per year, which is usually close to six or seven percent (i.e., one claim at fault on average every 15 years). As a result, senior policyholders cluster at the lowest level of the bonus–malus scale. Another consequence is that the bonus–malus coefficients are positively correlated with the premiums to which they apply multiplicatively (Taylor 1997; Brouhns et al. 2003). There is no such correlation in the statistical models with random effects that are presented in Sect. 3.7.
- A key characteristic is the duration of a claimless history that is necessary for a new policyholder to reach the lowest level of the bonus–malus scale. This duration equals 13 years in France or in Italy and is usually lower.
- A minor characteristic is the ratio between the bonus–malus coefficient related to the highest level of the scale, and the coefficient attributed to new policyholders. This ratio relates to the affordability of insurance for drivers with a poor history, although the bonus–malus scale usually does not enforce experience rating. This ratio equals 1.74 ($= 2/1.15$) in Italy and 3.5 in France.

The informational content of the bonus–malus scale increases with the seniority of the policyholder up to the duration mentioned in the second item. Afterward, this informational content decreases, and this all the more since the bonus–malus scale is unbalanced.

2.2 Memory in Bonus–Malus Scales and in Real-World Rating Structures

Section 3 focuses on nonlife insurance data where the frequency of claims is low, and with a long memory.[4] Accident data—and especially automobile insurance data—have these characteristics. The turnover of the portfolio must be low enough to enable long individual histories, which necessitates insurance contracts that are renewable by tacit agreement.

Memory in bonus–malus scales and in real-world rating structures is related to the magnitude of no-claim discounts, if the seniority is controlled for. The bonus–malus scales mentioned in Sect. 2.1 do not have a long memory. The bonus–malus coefficient of a new policyholder strongly decreases with time in the absence of claims, but the scales are unbalanced and senior policyholders cluster at the lowest level. In France, the maximum discount is 50 percent (with respect to a new driver), but the average discount of senior drivers is close to 40 percent. Hence, the discount of drivers at the lowest level of the scale is close to 15 percent with respect to their siblings in terms of seniority.[5] Insurers that use bonus–malus scales usually increase the no-claim discount. However, these discounts rarely exceed a third in real-world rating structures, if the seniority in the portfolio is controlled for.

[3] Twenty years ago, the European Commission sued France, arguing that the bonus–malus system distorted competition. As an answer, French authorities argued that the bonus–malus system did not enforce experience rating. They finally won the case.

[4] A rigorous definition of the length of memory will be given later from the theory of stationary time series.

[5] A multiplicative bonus–malus coefficient that decreases from $1 - 0.4 = 0.6$ to 0.5 is associated to a relative discount of $1/6$.

Statistical models for frequency risks (see Sect. 3) lead to higher values of the no-claim discounts. The discrepancy with real-world rating structures is explained by a longer term horizon in the derivation of commercial premiums, and by the information asymmetry between the incumbent insurer and its competitors. Regarding the first point, the insurance premiums can be derived by models with a long-term horizon, if the premium is seen as a control variable in a dynamic programming on the value of the policyholder (Taylor 1986; Emms and Haberman 2005). Regarding the second point, the assessment of information rents necessitates a thorough analysis of the memory in nonlife insurance data, which is the subject of Sect. 3.

2.3 Point-Record Driving Licenses

The incentive to careful driving of a bonus–malus scale is not clear-cut, because experience rating is not really enforced. At the opposite, the incentives are the main motivation for the point-record driving licenses based on traffic violations. In many countries, each convicted traffic offense is filed with a specific number of demerit points. When the accumulated demerit points exceed a given threshold, the driver's license is suspended or revoked. This mechanism is in force in North American jurisdictions. In many European countries, demerit points deplete a given capital (e.g., France, Italy). Redemption clauses are added so that the license suspension can be avoided in the long run. The redemption of a convicted traffic offense may occur if a seniority threshold is reached, as is the case in North American states and provinces. In Québec for instance, the seniority threshold equals two years, and driver's license suspensions depend on the demerit points recorded during the last two years. In many European countries, all the demerit points are redeemed (and the driver is reinstated with the initial capital) after a given period without any convicted traffic violation.

The incentives to drive carefully increase when the threshold of a license suspension gets closer. This is studied theoretically in Bourgeon and Picard (2007), and both theoretically and empirically in Dionne et al. (2011), Dionne et al. (2013). Since the 1970s, fatality rates due to road-traffic accidents have decreased steadily in OECD countries despite increased risk exposure. One major reason for these results has been the development of incentives for safe driving, including the point-record driving licenses.

Up to now, traffic violations are mostly recorded by radars, and only a tiny proportion of the committed traffic offences are registered by the road -traffic authorities. The frequency of convicted traffic violations is usually greater than the frequency of accidents, but their magnitude is similar. This could change drastically if cars were connected to the Internet of Things, a situation that might happen in a not so distant future. Then, the incumbent insurer could have access to the driver's behavior in continuous time (see Wüthrich 2017; Denuit et al. 2019, and Gao et al. 2019 for an analysis of this type of data). This information could also be available to the authorities, in the case of a point-record driving license.

The information asymmetry between the driver on the one hand, and the insurers and the authorities on the other hand, would then be reduced drastically. Risk exposure (the annual mileage or the number of trips) could be taken into account in the point-record mechanism. This is not the case today, and point-record driving licenses are very detrimental to drivers with an intensive use of their car for two reasons. On the one hand, the elasticity of the license removal risk with respect to traffic violation risk is much greater than one.[6] Hence, the license removal risk per kilometer increases with risk exposure.

[6] This elasticity can be derived with Markov chain arguments. It is lower than the average number of claims that trigger a license suspension (six in Quebec, five in France) and is lowered by the incentive effects of the point-record driving license. However, this elasticity remains much greater than one.

On the other hand, the private value of a license removal is much higher for drivers with an intensive use of their car (e.g., with daily commuting). Integrating risk exposure in point-record driving licenses could be an upside of a recording in real time of the driver's behavior.

3 Statistical Models for Panel Nonlife Insurance Data, with a Focus on Semiparametric Analysis

3.1 Statistical Models for Longitudinal Count Data

Numerous fully specified models handle panel data with constraints on the support of the dependent variable: See Davis et al. (2021) for a survey of count time series. Some of these models will now be listed:

- Models with thinning operators: An autoregressive specification is retained on the count variable Y, e.g.,
$$Y_t = \sum_{h=1}^{p} (\varphi_h \circ Y_{t-h}) + I_t^Y, \quad \varphi_h \geq 0 \ \forall h = 1, \ldots, p; \quad \sum_{h=1}^{p} \varphi_h < 1. \tag{1}$$
The product $\varphi \circ Y$ is defined as $\sum_{j=1}^{Y} K_j$, where the $\left(K_j\right)_{j=1,\ldots,Y}$ are i.i.d. nonnegative integer variables with a distribution that depends on the expectation φ. The innovations $\left(I_t^Y\right)_{t>p}$ are i.i.d. The Bernoulli distribution $B(1, \varphi)$ corresponds to binomial thinning. The distribution of the innovation can be chosen from a given parametric family such as Poisson or negative binomial: This leads to INAR(p) models (Al-Osh and Alzaid 1987; Jin-Guan and Yuan 1991; Gourieroux and Jasiak 2004). Relaxing the Bernoulli assumption on the $\left(K_j\right)_{j=1,\ldots,Y}$ leads to GINAR(p) models (Joe 2019). A downside of INAR(p) models is that at least p years of observation are needed in order to conduct pricing.
- "State-space" models: The causal scheme of the Kalman filter type is the following
$$\begin{array}{ccccc} \ldots & U_{t-1} & \longrightarrow & U_t & \ldots \\ & \downarrow & & \downarrow & \\ \ldots & Y_{t-1} & & Y_t & \ldots \end{array}.$$
In the next sections, the state variable U is a dynamic random effect, and Y is defined in a mixture model. Fully specified state-space models on longitudinal count data are found in Yang et al. (2015) and Lu (2018). A partly specified model based on first- and second-order moments (i.e., of the second-order semiparametric type) is presented in the following section.

3.2 Mixture Models with Dynamic Random Effects: a Specification for Nonnegative Dependent Variables

Single equation models with dynamic random effects for nonlife insurance risks are considered in this section. Three sequences define a mixture model on panel data:

- The dependent variable Y is nonnegative and square integrable.
- The expectation of Y splits into a function of covariates and a random effect:

$$E(Y_t \mid x_t,\ U_t) = \eta(x_t)\, U_t\ \forall t \in \mathbb{N}^*, \tag{2}$$

where $\mathbb{N}^*$ is the set of positive integers. The individual index is removed in this section, as cross-section analysis is used for estimation only. The covariates for each period t are represented by a vector x_t. A dynamic and positive random effect U, with an expectation equal to one, is plugged into the expectation of Y. The assumption $E(U_t) = 1\ \forall t \in \mathbb{N}^*$ is retained for identification purposes because the positive function $x \to \eta(x)$ includes a scale parameter. As $E(Y_t \mid x_t) = \eta(x_t) \overset{\text{def}}{=} \eta_t$ for all $t \in \mathbb{N}^*$, η is the marginal risk level. The link: $x \to E(Y \mid x) = \eta(x)$ can be estimated ad libitum (e.g., log-linear specification, index model, neural network, regression tree, or random forest). Equation (2) can be retained if the expectation of Y ranges in $]0, +\infty[$ without restriction. This is the case in nonlife insurance for the number of claims, the cost per claim, or the total cost of claims per period (see Pinquet 2022b for a case study with the total cost).

- The dynamic random effect U is assumed to be second-order stationary. The autocovariance function is denoted by γ_U.

3.3 Affine Prediction (Credibility Models) in a Second-Order Semiparametric Framework

Affine predictors of Y are obtained with a probabilistic regression in the model with random effects. They are denoted by $L(Y_{T+1} \mid Y_1, \ldots, Y_T)$. The predictor of Y_{T+1} resulting from linear filtering of Y at time T is split into the predicted marginal risk level and the history as follows:

$$L(Y_{T+1} \mid Y_1, \ldots, Y_T) = \eta_{T+1}\, L(U_{T+1} \mid Y_1, \ldots, Y_T). \tag{3}$$

This result is obtained if the following conditions hold (Pinquet 2022b)

$$E(Y_{t_1} \mid U_{t_1},\ U_{t_2}) = E(Y_{t_1} \mid U_{t_1})\ \left(= \eta_{t_1}\, U_{t_1}\right)\ \forall t_1, t_2 \in \mathbb{N}^*; \tag{4}$$

$$Cov(Y_{t_1},\ Y_{t_2} \mid U_{t_1},\ U_{t_2}) = 0\ \forall t_1, t_2 \in \mathbb{N}^*,\ t_1 \neq t_2. \tag{5}$$

Equation (5) corresponds to uncorrelated variables in a model with fixed effects: $Cov(Y_{t_1},\ Y_{t_2} \mid U_{t_1} = u_1,\ U_{t_2} = u_2) = 0\ \forall u_1, u_2 > 0,\ \forall t_1, t_2 \in \mathbb{N}^*,\ t_1 \neq t_2$. Together with (5), the assumptions made in (4) allow calculations with the total variance and covariance formulas.[7]

A simple example is detailed before the presentation of a generic result. Suppose that: a) $T = 1$ in (3); b) The random effect U is time invariant; c) Y is a count variable, with $V(Y_t \mid U) = E(Y_t \mid U) = \eta_t\, U\ \forall t = 1, 2$. This is the case if the distribution of Y_t conditional on $U = u$ is of the Poisson type $\forall t,\ \forall u > 0$. It is easily seen that

$$L(U \mid Y_1) = (1 - \alpha_1) + \alpha_1 \frac{Y_1}{\eta_1},\ \text{with}\ \alpha_1 = \frac{\eta_1 V(U)}{1 + \eta_1 V(U)}.$$

As $E(U) = 1$ is assumed, the weight α_1 is the no-claim discount. Indeed, the predictor derived from (3) equals $L(Y_2 \mid Y_1) = \eta_2\, L(U \mid Y_1) = \eta_2\ (1 - \alpha_1)$, if $Y_1 = 0$. The coefficient α_1 is usually referred to as a credibility weight.

[7] Alternative stories can be told from a mixture model. If the individual distributions belong to the model with random effects rather than to the model with fixed effects, the dynamic on the data becomes of the "self-exciting" type rather than of the revelation type. Disentangling the two effects is intricate (see Abbring et al. 2003).

In the general case, the affine probabilistic regression of U_{T+1} with respect to $Y_1, \ldots, Y_T$ is

$$L(U_{T+1} \mid Y_1, \ldots, Y_T) = \left(1 - \sum_{h=1}^{T} \alpha_{T,h}\right) + \left(\sum_{h=1}^{T} \alpha_{T,h} \frac{Y_{T+1-h}}{\eta_{T+1-h}}\right). \tag{6}$$

The index h denotes the lag between the forecast period $T+1$ and the risk exposure period $T+1-h$. The corresponding credibility weight is denoted by $\alpha_{T,h}$. From Eq. (3), the credibility weights are interpreted as a relative no-claim discount. They are related to the memory in the statistical model, as defined in Sect. 2.2.

The variance–covariance matrix of the random effects $U_1, \ldots, U_T$ is denoted by V_T. The vector of stacked credibilities per period is denoted by $v_\alpha^T = \underset{1 \le h \le T}{vec} \left(\alpha_{T,h}\right)$. The corresponding vector of autocovariances of U is denoted by $v_\gamma^T = \underset{1 \le h \le T}{vec} \left(\gamma_U(h)\right)$. Then, the credibility weights of Eq. (6) are the solution of the linear system (Pinquet 2022b)

$$\left[V_T + D_T\right] v_\alpha^T = v_\gamma^T, \text{ with : } D_T = \underset{1 \le h \le T}{diag} \left(d_{T+1-h}\right); \tag{7}$$

$$d_t = \frac{E(V(Y_t \mid U_t))}{\eta_t^2} = V\left(\frac{Y_t}{\eta_t}\right) - \gamma_U(0), \text{ with } d_t > 0 \, \forall t = 1, \ldots, T. \tag{8}$$

In the case of Poisson mixtures, Eq. (8) becomes

$$V(Y_t \mid U_t) = E(Y_t \mid U_t) = \eta_t U_t \Rightarrow d_t = 1/\eta_t \; \forall t. \tag{9}$$

An affine predictor of the nonnegative variable Y must be nonnegative. Sect. 3.4 defines positivity properties of the autocovariance function γ_U that ensure the positivity of the affine predictor regardless of the risk exposure.

3.4 Positivity Levels for Variance–Covariance Matrices and for the Autocovariance Functions of Stationary Time Series

Positivity Levels for Variance–Covariance Matrices

The symmetric positive semidefinite matrices are referred to as variance–covariance matrices in what follows. Definitions and results related to the theory of nonnegative matrices are recalled hereafter: See Berman and Plemmons (1994), Bernstein (2009) for a survey:

- A square matrix is said to be a Z-matrix if its off-diagonal entries are nonpositive.
- A square matrix is diagonally dominant if the absolute value of each diagonal entry is greater than or equal to the sum of the absolute values of the other entries in the same row (or column).
- The inverse of a nonsingular variance–covariance matrix V is called a precision matrix.

If a variance–covariance matrix V of order T is nonsingular and if the precision matrix $P = V^{-1}$ is a Z-matrix, then it can be shown that

$$P = s \times (I_T - B), \text{ with: } s > 0; \; B \ge 0; \; \rho(B) < 1, \tag{10}$$

where s is a scalar, I_T is the identity matrix of order T, and ρ is the spectral radius function (i.e., the maximum modulus of the eigenvalues of a square matrix). The inequalities applied to the vectors and matrices in the paper are assumed to hold entrywise. The spectral radius of an entrywise nonnegative matrix B is called its Perron root, which corresponds to a nonnegative eigenvalue and to an eigenvector that can be chosen to be entrywise nonnegative.

Equation (10) defines the class of invertible $M-$matrices.[8] Because $P^{-1} = (1/s) \times \sum_{n=0}^{+\infty} B^n \geq 0$ in (10), the inverse of a nonsingular M-matrix is entrywise nonnegative. The inverses of nonsingular M-matrices are usually called $\mathcal{IM}$-matrices (Johnson and Smith 2011). Nonsingular matrices whose inverses are diagonally dominant and $M-$matrices are said to be potentials (Dellacherie et al. 2016).

Three positivity levels have been recalled for variance–covariance matrices. By increasing order of positivity, these levels are: nonnegativity entrywise, $\mathcal{IM}$-type, and potential type. Autocovariance functions that generate variance–covariance matrices with various positivity levels are detailed in Sect. 3.4.

Positivity Levels for Autocovariance Functions

Three positivity levels of the autocovariance functions of random effects are considered. The lowest level refers to the filtering equations derived from the autocovariance function. If U is a positive and stationary random effect with an expectation equal to one, γ_U is also the autocovariance function of the centered stationary sequence $U^c = U - 1$. The filtering coefficients are obtained from the linear probabilistic regressions:

$$U_t^c = \sum_{h=1}^{T} \varphi_{T,h}\, U_{t-h}^c + E_t^T \ \forall T \in \mathbb{N}^*,\ \forall t > T. \tag{11}$$

The residual E_t^T is uncorrelated with the regressors in (11), and the filtering coefficients $\varphi_{T,h}$ are derived with the Levinson–Durbin recursion:

1. The lowest positivity level is associated to nonnegative linear filtering ($i.e.,\ \varphi_{T,h} \geq 0\, \forall T, h$). From the Levinson–Durbin recursion, nonnegative linear filtering implies nonnegative autocovariances (Pinquet 2020a). The autocovariance functions with nonnegative linear filtering are compatible with positive time series. This result is proven in Pinquet (2022a). Hence, the autocovariance functions specified later are compatible with sequences of stationary and positive random effects.
2. The $\mathcal{IM}$ positivity level is reached if the off-diagonal entries of the precision matrices $\left(V_T^{-1}\right)_{T\in\mathbb{N}^*}$ are nonpositive. The $\left(V_T^{-1}\right)_{T\in\mathbb{N}^*}$ are M-matrices (see (10)). Hence, the variance–covariance matrices $(V_T)_{T\in\mathbb{N}^*}$ are of the $\mathcal{IM}$-type. This positivity level implies nonnegative linear filtering (Pinquet 2020a). The $\mathcal{IM}$ positivity level is strictly stronger than nonnegative linear filtering: This point is clarified later for AR(p) specifications. The $\mathcal{IM}$ positivity level on the random effects implies nonnegative linear filtering of the dependent variables in mixture models (see Sect. 3.5).
3. The highest positivity level is associated with potential matrices (i.e., diagonal dominance is a supplementary property of the precision matrices $\left(V_T^{-1}\right)_{T\in\mathbb{N}^*}$). This level implies positive intercepts for the affine predictors of nonnegative dependent variables in mixture models (Pinquet 2022b). Hence, these affine predictors are positive. Variance–covariance matrices of the $\mathcal{IM}$-type are not necessarily potentials, but the examples known by the author are not Toeplitz. The autocovariance

[8] The M-matrix and Z-matrix denominations were introduced by Alexander Ostrowski as a tribute to Hermann Minkowski and to Ernst Zermelo. However, the Z-matrix denomination is also used in computational chemistry and refers to internal coordinates of atoms in molecules.

functions discussed below have two positivity levels (either nonnegative linear filtering or the highest level).

3.5 Positive Affine Prediction of Risk Variables and Positivity Properties of the Autocovariance Function of the Random Effects

We consider the affine prediction problem of Sect. 3.3. The highest positivity level is assumed for γ_U. The nonnegativity of the credibility weights is reached from the positivity properties of variance–covariance matrices of the random effects. The proof starts from (7):

$$v_\alpha^T = [V_T + D_T]^{-1} v_\gamma^T = [I_T + (P_T\, D_T)]^{-1} v_\varphi^T, \text{ with}$$

$$P_T = V_T^{-1};\ v_\varphi^T = \underset{1\le h\le T}{vec}(\varphi_{T,h}) \text{ (see (11))} = V_T^{-1} v_\gamma^T. \text{ Then,}$$

$$v_\alpha^T = [I_T + (P_T\, D_T)]^{-1} v_\varphi^T = D_T^{-1}[D_T^{-1} + P_T]^{-1} v_\varphi^T. \tag{12}$$

From the positivity assumptions, the vector of filtering coefficients v_φ^T is nonnegative and P_T is an M-matrix. The symmetric positive definite matrix $A = D_T^{-1} + P_T$ is a Z-matrix (its off-diagonal entries are nonpositive). Hence, A is a nonsingular M -matrix (see (10)). The inverse of a nonsingular M -matrix is entrywise nonnegative (see Sect. 3.4). Hence, $[D_T^{-1} + P_T]^{-1} \geq 0$: from (12) and $v_\varphi^T \geq 0$, and we obtain $v_\alpha^T \geq 0$.

The credibility weights are then nonnegative if V_T is a potential matrix for all $T \in \mathbb{N}^*$. This assumption also implies a positive intercept of the affine predictor given in (6). The proof (Pinquet 2022b) uses the diagonal dominance property of the precision matrix associated to a potential variance–covariance matrix. Hence, the affine predictors defined by (3) and (6) are positive regardless of the individual history.

The set of potential matrices is closed under extraction of principal submatrices (Dellacherie et al. 2014). This hereditarity result implies positive predictors at any horizon.

3.6 Autocovariance Functions with Positivity Properties and That Reach the Three Levels for the Length of Memory

The autocovariance function γ_U of the stationary random effect U is defined by $\gamma_U(h) = Cov(U_t,\, U_{t-h})\ \forall h \in \mathbb{N}^*,\ \forall t > h$: The invariance of the covariance with respect to time translations reflects the stationarity. Contract time is usually the set of positive integers: However, the time domain must be extended to all the integers in order to use representations of autocovariance functions by linear filtering. Linear filtering of the random effects is unambiguous if the variance–covariance matrices $(V_T)_{T\in\mathbb{N}^*}$ are nonsingular (and hence positive definite). Three levels of memory can be reached by the autocovariance function of the stationary random effects.[9]

[9] The three levels of memory are defined in the time domain in what follows, from the autocovariance function. These levels are also obtained with a duality argument. From the Fourier transform, the time domain $\mathbb{Z}$ is associated to the frequency domain $\mathbb{R}/2\pi\mathbb{Z} \simeq [-\pi, \pi[$. The autocovariance function is the Fourier transform of a nonnegative spectral measure, defined on the frequency domain. The length of memory is related to the regularity at zero (in the frequency domain) of the cumulant of the spectral measure. This function is increasing, since the spectral measure is nonnegative. The three regularity levels at zero of this increasing function are related to the length of memory as follows: (a) differentiability: short memory; (b) continuity, without differentiability: long memory; (c) discontinuity:

- Short memory corresponds to absolutely summable autocovariance functions (i.e., $\sum_{h\in\mathbb{Z}} |\gamma_U(h)| < +\infty$). This is the case for the stationary autoregressive specifications. Short memory implies ergodicity, i.e., $\lim_{h\longrightarrow+\infty} \gamma_U(h) = 0$. The AR($p$) specifications with nonnegative linear filtering are of the type

$$U_t^c = \sum_{h=1}^{p} \varphi_h \, U_{t-h}^c + I_t^U, \; \varphi_h \geq 0 \; \forall h = 1, \ldots, p; \; \sum_{h=1}^{p} \varphi_h < 1, \tag{13}$$

  where $U^c = U - 1$ is the centered random effect. The stationary AR(p) specifications that reach the highest positivity level (i.e., the M-matrix and the diagonal dominance properties for the precision matrices) are characterized by (13) and by the supplementary condition (Pinquet 2020a)

$$\varphi_h \geq \sum_{1\leq\tau\leq p-h} \varphi_\tau \, \varphi_{\tau+h}, \quad \forall h = 1, \ldots, p-1. \tag{14}$$

  The affine predictors of Y defined in (3) are then positive regardless of the risk exposure. From (13), the condition (14) is fulfilled if φ is decreasing.
  In Eq. (6), the total credibility $\sum_{h=1}^{T} \alpha_{T,h}$ is the relative no-claim discount in the frequency risk model (see (3)). As such, total credibility is related to the informal definition given in the abstract for the memory in the data. The limit credibility (i.e., the limit of the total credibility when the length of the history T goes to infinity) is less than one if the random effects have a short memory (Pinquet 2020a).
- Long memory corresponds to non-summable and ergodic autocovariance functions. This is the case for the ARFIMA$(0, d, 0)$ specifications (Granger and Joyeux 1980), which are defined by the equation $(I - L)^d \, U^c = I^U$. The lag operator is denoted by L, and the fractional differencing parameter d belongs to the interval $\left]0, \frac{1}{2}\right[$. The filtering equation is

$$U_t^c = \sum_{h=1}^{+\infty} \varphi_h \, U_{t-h}^c + I_t^U, \qquad \text{with } \varphi_1 = d; \; \frac{\varphi_{h+1}}{\varphi_h} = \frac{h-d}{h+1} \; \forall h \in \mathbb{N}^*. \tag{15}$$

  The filtering coefficients are positive and decreasing. The equality $\sum_{h=1}^{+\infty} \varphi_h = 1$ reflects the long memory of the ARFIMA$(0, d, 0)$ specifications, which reach the highest positivity level (Pinquet 2020b).
- The highest level of memory corresponds to non-ergodic autocovariance functions. We write

$$U_t = P \times Q_t \Rightarrow \gamma_U = \sigma_P^2 + \left[(1 + \sigma_P^2) \, \gamma_Q\right], \tag{16}$$

  where γ_Q vanishes at infinity. Equation (16) is obtained if: a) the variables P and $(Q_t)_{t\in\mathbb{N}^*}$ are uncorrelated; b) $E(P) = E(Q_t) = 1 \, \forall t \in \mathbb{N}^*$; c) P^2 and $Q_t \, Q_{t+h}$ are uncorrelated for all t, h. Since γ_Q vanishes at infinity, the non-ergodic specifications correspond to $\sigma_P^2 > 0$. If γ_Q reaches the highest positivity level defined in Sect. 3.4, then this also holds for γ_U: see Pinquet (2022b). Hence, autocovariance functions that ensure positive affine predictors in the mixture model and that reach the three levels of memory have been obtained in this section.
  The limit credibility equals one if the random effects have a long or the highest level of memory (Pinquet 2020a).

highest level of memory. In that case, the jump of the cumulant is associated to a Dirac mass of the spectral measure located at zero, and to the variance of the time-invariant component of the random effect (see (16)):

3.7 Fully Specified vs. Semiparametric Mixture Models

This section discusses the pros and cons of fully specified vs. semiparametric models for nonnegative panel data. The starting point is different: the data generating process for the former approach, and the estimating functions for the latter. In favor of fully specified models is their number and the shape of the prediction obtained by conditional expectation: The predictor is not bound to be an affine function of the dependent variable. At the opposite, a semiparametric analysis of mixture models with dynamic random effects is difficult to undertake beyond the second order.[10] Hence, the affine constraint on the shape of the predictor is natural. However, numerous fully specified models lead to predictors with an affine shape.

The estimators and predictors are difficult to derive if the likelihood of the fully specified model does not have a closed form, whereas the semiparametric approach uses estimating functions that are easily tractable. In addition, there is an identification issue with the fully specified models. The output of the analysis (length of memory, shape of the predictors) is less profuse than the numerous specifications. This identification issue is often referred to as a "Rashōmon effect."[11]

3.8 Estimation of Panel Count Data with the GMM

We will now give in detail the estimation of the semiparametric model of Sect. 3.2. Estimating functions must be chosen first, whose number is greater than or equal to the number of parameters to estimate (i.e., overidentified or just identified approach). The individual index is $i = 1, \ldots, I$. The vector of parameters θ splits into first- and second-order components, with $\theta = (\theta_1, \theta_2)$. On an unbalanced panel data set (i.e., the length T_i of the time series varies with i), the moment equations derived from the model of Sect. 3.5 become

$$E(Y_{i,t}) = \eta_{i,t}(\theta_1);\ Cov(Y_{i,t}, Y_{i,t-h}) = \eta_{i,t}(\theta_1)\,\eta_{i,t-h}(\theta_1)\,\gamma_{\theta_2}(h)\ (0 < h < t \leq T_i). \tag{18}$$

The expectation $\eta_{i,t}(\theta_1)$ varies within the time series, and Y is not stationary at the individual level. The covariances are obtained from (2) and (5). The variances are $V(Y_{i,t}) = V(E(Y_{i,t} \mid U_{i,t})) + E(V(Y_{i,t} \mid U_{i,t}))$, with $V(E(Y_{i,t} \mid U_{i,t})) = V(\eta_{i,t}(\theta_1)\,U_{i,t}) = \eta_{i,t}^2(\theta_1)\,\gamma_{\theta_2}(0)$. The conditional variance

[10] Semiparametric analysis beyond the second order is well documented on cross-section count data. For instance, moments of any order for Poisson distributions are obtained from

$$Y \sim P(\lambda) \Rightarrow E\,[Y(Y-1)\ldots(Y-k+1)] = \lambda^k\ \forall k \in \mathbb{N}^*.$$

Then, moments of any order for time-invariant random effects in Poisson mixtures are obtained from

$$Y \mid U = u \sim P(\lambda\,u),\ E(U) = 1 \Rightarrow E(U^k) = \frac{E\,[Y(Y-1)\ldots(Y-k+1)]}{[E(Y)]^k}\ \forall k \in \mathbb{N}^*. \tag{17}$$

The mixing distribution is asymptotically identified from a sequence of moments of increasing order. Then, nonparametric results are obtained from a semiparametric analysis (e.g., see Zhang 1990). Hence, results for semiparametric analysis at various orders are available for time-invariant random effects in mixture models. To the best of our knowledge, there are no such results for dynamic random effects beyond the second order.

[11] There is a crime scene in Kurosawa's movie "Rashōmon": A samourai has been murdered, and his wife has been raped. The wife, a bandit, a woodcutter, and even the murdered samourai (through a Shinto psychic) testify at a trial. The four stories are different, but they are all consistent with the crime scene. The "Rashōmon effect" occurs when different explanations exist for the same phenomenon. This effect is quoted by Breiman (2001). Breiman despises fully specified models in statistical analysis, but the alternative is the unsupervised learning approach rather than the partly specified models investigated in this chapter.

$V(Y_{i,t} \mid U_{i,t})$ must be specified in order to perform the estimation. In the case of Poisson mixtures: $V(Y_{i,t} \mid U_{i,t}) = E(Y_{i,t} \mid U_{i,t}) = \eta_{i,t}(\theta_1)\, U_{i,t} \Rightarrow E(V(Y_{i,t} \mid U_{i,t})) = \eta_{i,t}(\theta_1)$. The case where Y is the total cost of claims per period is treated in Pinquet (2022b).

The estimating functions for Poisson mixtures are denoted as F_1 and F_2 for the first and second orders, respectively. The first-order parameters are obtained from F_1, with $(F_1)_{i,t} = \left(Y_{i,t} - \eta_{i,t}(\theta_1)\right) x_{i,t}$, and where $x_{i,t}$ is a column vector of covariates. Setting equal to zero the sample mean of F_1 (denoted by $\overline{F_1}$) leads to the maximum likelihood estimation of θ_1 in a Poisson model if $\eta_{i,t}(\theta_1) = \exp(\theta_1 x_{i,t})$.

The first-order estimators are consistent in the mixture model, because the expectation of the dependent variable in the mixture model is not modified by the second-order parameters. Hence, the estimators of θ_1 and of the marginal frequency risks are considered as given in the estimation of the second-order parameters. The simple case of cross-section count data is treated first, in order to clarify the choice of the second-order estimating functions. A moment-based estimator of the variance σ^2 of a random effect is

$$\widehat{\sigma^2} = \frac{\sum_i (Y_i - \widehat{\eta_i})^2 - \widehat{\eta_i}}{\sum_i \widehat{\eta_i}^2} = \frac{\sum_i N_{0,i}}{\sum_i D_{0,i}} = \frac{\overline{N_0}}{\overline{D_0}}, \tag{19}$$

with $(\widehat{\eta_i} = \eta_i(\widehat{\theta_1}))_{i=1,\ldots,I}$. This estimator of σ^2 is unconstrained, and $\widehat{\sigma^2} < 0$ if there is underdispersion in the data. The GMM estimator derived from the estimating function $F_2 = N_0 - \sigma^2 D_0$ is

$$\widehat{\sigma^2}^{GMM} = \arg\min_{\sigma^2 \geq 0} \overline{F_2}^2 = \max(0, \widehat{\sigma^2}) = \left[\widehat{\sigma^2}\right]^+. \tag{20}$$

The GMM estimator equals zero if there is underdispersion in the data. The same result holds for a maximum likelihood estimator of σ^2 a fully specified mixture model, because $\sum_i N_{0,i}$ in (19) is of the same sign as the Lagrangian in Rao's score test for the nullity of σ^2.

In the panel data case, we write: $\widehat{\eta_{i,t}} = \eta_{i,t}(\widehat{\theta_1})$; $\widehat{\eta_{i,\circ}} = \sum_t \widehat{\eta_{i,t}}$; $Y_{i,\circ} = \sum_t Y_{i,t}\ \forall i, t$. The estimator of the variance of a time-invariant random effect in the mixture model, which is derived from Eq. (19), appears as a weighted average of unconstrained estimators of the autocovariances of a stationary random effect U, i.e.,

$$\widehat{\sigma^2} = \frac{\sum_i \left(Y_{i,\circ} - \widehat{\eta_{i,\circ}}\right)^2 - \widehat{\eta_{i,\circ}}}{\sum_i \widehat{\eta_{i,\circ}}^2} = \sum_{0 \leq h < T_{\max}} \alpha_h \widehat{\gamma_U(h)};\ T_{\max} = \max_{1 \leq i \leq I} T_i; \tag{21}$$

$$\widehat{\gamma_U(0)} = \frac{\sum_{i,t} \left(Y_{i,t} - \widehat{\eta_{i,t}}\right)^2 - \widehat{\eta_{i,t}}}{\sum_{i,t} \widehat{\eta_{i,t}}^2};\ \underset{0<h<T_{\max}}{\widehat{\gamma_U(h)}} = \frac{\sum_{i,\, T_i > h} \sum_{h < t \leq T_i} \left(Y_{i,t} - \widehat{\eta_{i,t}}\right)\left(Y_{i,t-h} - \widehat{\eta_{i,t-h}}\right)}{\sum_{i,\, T_i > h} \sum_{h < t \leq T_i} \widehat{\eta_{i,t}}\, \widehat{\eta_{i,t-h}}}. \tag{22}$$

Equation (18) and the splitting with respect to the lag of the numerator and of the denominator in (21) lead to these results.

The GMM estimation of the second-order parameters (e.g., $\theta_2 = (\sigma_P^2, \gamma_Q(0), d)$ if the random effect is defined in (16) with an ARFIMA$(0, d, 0)$ specification for the ergodic component) generalizes the strategy retained in (20).

A natural choice for the second-order estimating functions is to start from the unconstrained estimators of $(\gamma_U(h))_{0 \leq h < T_{\max}}$. If $\widehat{\gamma_U(h)}$ is written as a ratio of sample means (i.e., $\widehat{\gamma_U(h)} = \overline{N_h}/\overline{D_h}$), then $F_{2,h} = N_h - \gamma_{\theta_2}(h) D_h$ is an estimating function related to the lag h, $0 \leq h < T_{\max}$. The sample means are derived on the pairs (i, t). For instance, $(D_h)_{i,t} = 1_{[h<t\leq T_i]}\, \widehat{\eta_{i,t}}\, \widehat{\eta_{i,t-h}}$.

The number of estimating functions equals $T_{\max}$, i.e., the maximum length of the time series. This number must be greater than or equal to the number of parameters to estimate, in order to follow an

overidentified or a just identified estimation strategy. The dimension of θ_2 is 2 or 3, and the identifying condition is easily reached from panel data.

The generalized method of moments (Hansen 1982) uses the sample mean of the estimating functions. The sample mean maps the parameter set onto a differentiable manifold, which is embedded in the space related to the estimating functions. The parameters are estimated by the GMM together with a metric that is used to minimize the distance between the origin of the space and the manifold.

The vector of estimating functions that extends $F_2(\sigma^2) = N_0 - \sigma^2 D_0$ in the cross-section case is $F_2(\theta_2) = \underset{0\leq h<T_{\max}}{vec}(N_h - \gamma_{\theta_2}(h)D_h) = v_N - (v_{\gamma_{\theta_2}} \odot v_D)$, with $v_N = \underset{0\leq h<T_{\max}}{vec}(N_h)$, etc. The Hadamard product (i.e., the entrywise product) is denoted by $\odot$. The unconstrained estimates of the autocovariances given in (22) are such that $\overline{F_2} = 0$ if $v_{\gamma_{\theta_2}}$ is replaced by $v_{\widehat{\gamma}} = \underset{0\leq h<T_{\max}}{vec}(\widehat{\gamma_U(h)}) = \overline{v_N}/\overline{v_D}$. An example is provided in Table 1, Sect. 4.

A GMM estimator of θ_2 is of the type

$$\widehat{\theta_2}^{GMM} = \arg\min_{\theta_2} \left\| \overline{F_2}(\theta_2) = \overline{v_N} - (v_{\gamma_{\theta_2}} \odot \overline{v_D}) \right\|_M^2 = \arg\min_{\theta_2} \left\| v_{\widehat{\gamma}} - v_{\gamma_{\theta_2}} \right\|^2_{M\odot(\overline{v_D}\,\overline{v_D}')}, \tag{23}$$

where M is an Euclidean metric. The last equality results from $\|a \odot b\|_M^2 = \|a\|^2_{M\odot(bb')}$ $\forall a, b, M$. From Hansen (1982), an optimal metric that varies with θ_2 is $M = \left[\overline{F_2(\theta_2)\, F_2'(\theta_2)}\right]^{-1}$. The unconstrained estimates of the autocovariances are a raw material used to fit autocovariance functions with positivity properties, and that have various lengths of memory.

Unconstrained estimating functions are used for the second-order parameters, but the GMM estimator belongs to the parameter space. This parameter space has a boundary, and the GMM estimation is obtained at the boundary if the data reject the model totally or partially. In this respect, the GMM estimator behaves exactly like a maximum likelihood estimator derived from a fully specified model.

4 A Case Study

This case study applies the results of the previous sections to the semiparametric analysis of panel count data. Some of the autocovariance functions that are detailed in Sect. 3.6 are fitted to the unconstrained estimations derived from (22). The database used in this case study represents all the risk exposure of a Spanish portfolio of automobile insurance contracts with renewal dates ranging between 1991 and 1997 (first and last years included). The last period is observed in full (see Pinquet 2020a for more details). With the notation of Sect. 3.7:

- The count variable Y is the number of claims at fault.
- Twenty indicator variables (including the intercept) are used as regression components. Hence, $x_{i,t} \in \{0, 1\}^{20}$ $\forall i, t$.
- The length of the individual time series is denoted by T_i and varies between 1 and 7 in the database. Hence, $T_{\max} = 7$, and there are 27 estimating functions.
- Each period t corresponds to a one-year risk exposure. The number of insurance contracts equals $I = 269388$. The estimates of the autocovariances derived from (22) are given in Table 1.

The strongly decreasing shape of the unconstrained estimations in Table 1 justifies specifications with dynamic random effects. The overidentified estimation strategy of the GMM lets the data speak,

Table 1 Estimates of $\gamma_U(h)$ $(0 \leq h < 7$; $\widehat{\gamma_U}$: unconstrained function)

h(lag)	0	1	2	3	4	5	6
$\widehat{\gamma_U(h)}$	1.269	0.802	0.615	0.586	0.553	0.457	0.442

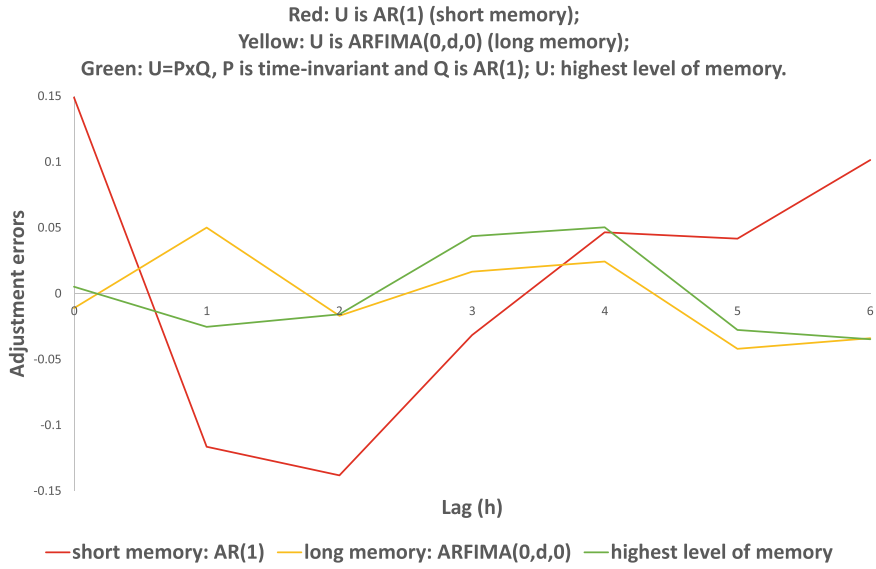


Fig. 1 Adjustment errors (unconstrained - fitted autocovariances of random effects)

regarding the length of memory in the mixing distribution. Some of the autocovariance functions that were discussed in Sect. 3.6 are fitted to the values of Table 1 with a least squares criterion. This estimation is a rough version of the GMM. It does not necessitate supplementary information and corresponds to the metric $M = \underset{0 \leq h < T_{\max}}{diag}\ (1/\overline{D_h}^2)$ (see (23)). Various metrics in this context are discussed in Pinquet (2020a).

Results are given for the three levels of memory in the mixing distribution. The notations are those of Eq. (16), i.e., $U_t = P\,Q_t\ \forall t \in \mathbb{N}^*$; $\gamma_U = \sigma_P^2 + (1 + \sigma_P^2)\gamma_Q$, where γ_Q vanishes at infinity:

1. Short memory: $U = Q$, where γ_Q follows an AR(1) specification ($\gamma_Q(h) = \gamma_Q(0) \times \varphi_1^h \forall h \in \mathbb{N},\ \varphi_1 \in [0, 1[$). The rough GMM estimation of $\theta_2 = (\gamma_Q(0),\ \varphi_1)$ is: $\widehat{\theta_2} = (1.12,\ 0.82)$.
2. Long memory: $U = Q$, where γ_Q follows an ARFIMA$(0, d, 0)$ specification. Hence, $\gamma_Q(h + 1) = \frac{h+d}{h+1-d} \times \gamma_Q(h)\ \forall h \in \mathbb{N}$ (Granger and Joyeux 1980). The rough GMM estimation of $\theta_2 = (\gamma_Q(0),\ d)$ is $\widehat{\theta_2} = (1.28,\ 0.37)$.
3. Highest level of memory: $U = P\,Q$, with $\sigma_P^2 \geq 0$, and where Q follows an AR(1) specification. The rough GMM estimation of $\theta_2 = (\sigma_P^2,\ \gamma_Q(0),\ \varphi_1)$ is $\widehat{\theta_2} = (0.47,\ 0.54,\ 0.45)$. As $\widehat{\sigma_P^2} > 0$, the fitted specification is not ergodic.

The adjustment errors for the three estimations are displayed in Fig. 1.

The short memory specification provides a poor fit to the unconstrained estimations, as compared to the two other specifications (the sums of squared adjustment errors are equal to 0.0700, 0.0067, and 0.0073, respectively). An exponential decay of the autocovariance function provides a good fit to the raw data of Table 1 only if the ergodicity assumption is relaxed. The ARFIMA$(0, d, 0)$ specification is by far the best of the two ergodic choices for the autocovariance function. Besides, adding a time-

invariant component P to an ARFIMA$(0, d, 0)$ specification of Q does not modify the estimation $(\widehat{\sigma_P^2} = 0)$.

Positive affine prediction on frequency risks is obtained from (3) and (6), with credibility weights derived from (7). From (6), the total credibility $\sum_{h=1}^{T} \alpha_{T,h}$ is the no-claim discount. The intercept is always positive: Hence, the total credibility is less than one.

After $T = 40$ years and for the average frequency risk (nine percent per year), the total credibility equals 26.6%, 52.6% and 60.9%, respectively, if the random effects follow the three specifications sorted by increasing length of memory.

The no-claim discounts rarely exceed one-third in the real-world rating of automobile insurance, if the seniority in the portfolio is controlled for. Among the three models, the total credibility obtained with a short memory is the closest to real-life values of no-claim discounts. This result is in contrast to the goodness-of-fit results of Fig. 1. This contradiction is explained by the subsidies from senior policyholders to newcomers in nonlife insurance portfolios. The incumbent insurer collects information on the policyholders which is not available to the competitors. This information rent explains these cross-subsidies, and why real-world rating structures have a lower memory than frequency risks estimated by statistical models.

References

Abbring JH, Heckman JJ, Chiappori P-A, Pinquet J (2003) Adverse selection and moral hazard in insurance: can dynamic data help to distinguish? J Eur Econ Assoc **1**(2–3):512–521

Al-Osh M, Alzaid AA (1987) First-order integer-valued autoregressive (INAR(1)) process. J Time Series Anal 8(3):261–275

Berman A, Plemmons RJ (1994). Nonnegative matrices in the mathematical sciences. SIAM, Philadelphia

Bernstein DS (2009) Matrix mathematics. Princeton University Press, Princeton

Bourgeon J-M, Picard P (2007) Point-record driving licence and road safety: an economic approach. J Public Econ 91(1–2):235–258

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231

Brouhns N, Guillén M, Denuit M, Pinquet J (2003) Bonus-malus scales in segmented tariffs with stochastic migration between segments. J Risk Insur 70(4):577–599

Bühlmann H, Gisler A (2006) A course in credibility theory and its applications. Springer, Heidelberg

Davis RA, Fokianos K, Holan SH, Joe H, Livsey J, Lund R, Pipiras V, Ravishanker N (2021) Count time series: a methodological review. J Am Stat Assoc 116(535):1533–1547

Dellacherie C, Martinez S, San Martin J (2014) Inverse M-matrices and ultrametric matrices, vol 2118. Springer, Berlin

Dellacherie C, Martinez S, San Martin J (2016) Additive representation of symmetric inverse M-matrices and potentials. Adv Appl Math 81:13–39

Denuit M, Guillen M, Trufin J (2019) Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. Ann Actuarial Sci 13:1–22

Dionne G, Michaud P-C, Pinquet J (2013) A review of recent theoretical and empirical analyses of asymmetric information in road safety and automobile insurance. Res Transp Econ 43(1):85–97

Dionne G, Pinquet J, Maurice M, Vanasse C: Incentive mechanisms for safe driving: a comparative analysis with dynamic data. Rev Econ Stat 93(1):218–227

Ehrlich I, Becker G (1972) Market insurance, self-protection, and self-insurance. J Polit Econ 80(4):623–648

Emms P, Haberman S (2005) Pricing general insurance using optimal control theory. ASTIN Bull 35(2):427–453

Gao G, Meng S, Wüthrich MV (2019) Claims frequency modeling using telematics car driving data. Scand. Actuarial J 2019(2):143–162

Gourieroux C, Jasiak J (2004) Heterogeneous INAR (1) model with application to car insurance. Insur Math Econ 34(2):177–192

Granger CW, Joyeux R (1980) An introduction to long-memory time series models and fractional differencing. J Time Ser Anal 1(1):15–29

Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50(4):1029–1054

Hendel I, Lizzeri A (2003) The role of commitment in dynamic contracts: Evidence from life insurance. Q J Econ 118(1):299–328
Jin-Guan D, Yuan L (1991) The integer-valued autoregressive (INAR(p)) model. J Time Ser Anal 12(2):129–142
Joe H (2019) Likelihood inference for generalized integer autoregressive time series models. Econometrics 7(4):43.
Johnson CR, Smith RL (2011) Inverse M-matrices, II. Linear Algebra Appl 435(5):953–983
Lemaire J (1995) Bonus-malus systems in automobile insurance, vol 19. Springer, Berlin
Lemaire J, Zi H (1994) A comparative analysis of 30 bonus-malus systems. ASTIN Bull 24(2):287–309
Lu Y (2018) Dynamic frailty count process in insurance: a unified framework for estimation, pricing, and forecasting. J Risk Insur 85(4):1083–1102
Pinquet J (2020a) Poisson models with dynamic random effects and nonnegative credibilities per period ASTIN Bull 50(2):585–618
Pinquet J (2020b) Positivity properties of the ARFIMA$(0, d, 0)$ specifications and credibility analysis of frequency risks. Insur Math Econ 95(6):159–165
Pinquet J (2022a) Autocovariance functions with nonnegative linear filtering are compatible with positive time series. https://doi.org/10.13140/RG.2.2.33966.69446
Pinquet J (2022b) Hereditarity of potential matrices and positive affine prediction of nonnegative risks from mixture models. Scand Actuarial J 2022(8):659–681
Taylor G (1997) Setting a bonus-malus scale in the presence of other rating factors. ASTIN Bull 27(2):319–327
Taylor GC (1986) Underwriting strategy in a competitive insurance environment. Insur Math Econ 5(1):59–77
Wüthrich MV (2017) Covariate selection from telematics car driving data. Eur Actuarial J 7(1):89–108
Yang M, Cavanaugh JE, Zamba GK (2015) State-space models for count time series with excess zeros. Stat Model 15(1):70–90
Zhang C-H (1990) Fourier methods for estimating mixing densities and distributions. Ann Stat 18(2):806–831



Efficiency and Productivity in the Insurance Industry Using Frontier Methodologies

María Rubio-Misas

1 Introduction

The last four decades have witnessed the growth in research on efficiency and productivity using frontier methodologies. These methods constitute benchmarking ways for estimating efficiency and productivity and have become mainstream in measuring the performance of firms and other types of decision-making units using accounting data, to the point that the development of frontier methodologies constitutes an important advancement in modern economics. This growth has had main implications for insurance economics, as numerous studies require the comparison of insurers with other companies in the industry or the comparison of branches inside a firm. The chapters by Cummins and Weiss (2000, 2013) in the previous editions of this book provided the foundations for insurance economists to adapt their research to incorporate the frontier efficiency approach. They also provided guidance in using methodology and defining inputs and outputs as well as surveys of research on efficiency and productivity in the insurance industry using frontier efficiency methods.[1] Nevertheless, after the last survey conducted by Cummins and Weiss (2013), recent work in this field has refined methodologies, addressed new topics, and extended geographic coverage, particularly in terms of multicountry analyses. The aim of this chapter is to provide a comprehensive survey of research on efficiency and productivity in the insurance industry utilizing frontier methodologies that have been published from 2012 to 2022.

Efficiency concepts were developed in the 1950s by Farrel (1957), building on the early work of Debreu (1951). Traditional microeconomic theory maintains that all companies minimize costs and maximize profits and that companies that fail in getting these objectives are not of interest

[1] While Cummins and Weiss (2000) covered the studies on efficiency and productivity in the insurance industry using frontier methods conducted until 1999, Cummins and Weiss (2013) surveyed the studies in this field over the period 2000–2011.

M. Rubio-Misas (✉)
Finance and Accounting Department, Universidad de Málaga, Malaga, Spain
e-mail: mrubiom@uma.es

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2_17

because they will not survive (Cummins and Weiss 2013). Modern frontier efficiency analysis began to build a theoretical framework to evaluate companies that fail in optimization and are consequently not fully efficient (Farrel 1957). With this framework, efficiency is analyzed by comparing firms to "best practice" efficient frontiers formed by the most efficient firms in the reference set. Posteriorly, methodologies for estimating efficiency were developed in the 1970s. The most important contributions were the development of stochastic frontier analysis by Aigner et al. (1977), Battese and Corra (1977), and Meeusen and van den Broeck (1977) and the development of nonparametric mathematical programming frontiers by Charnes et al. (1978).

The two most important groups of approaches for estimating frontiers are econometric and nonparametric methods. Econometric methods, including stochastic frontier analysis (SFA) (see Greene 2008), estimate efficiency using econometric techniques and usually make assumptions about the functional form of the function and error term distribution. However, nonparametric methods do not make assumptions about functional forms and estimate efficiency utilizing linear programming and other nonparametric techniques (see Cooper et al. 2011). Data envelopment analysis (DEA) constitutes the most outstanding nonparametric method. There is no consensus on the best method for deciding the best-practice frontier against which the relative efficiencies of firms are calculated. While econometric methods have the advantage that firms are allowed to deviate from the frontier due to error as well as inefficiency, nonparametric methods measure all departures from the frontier as inefficiency. However, the econometric methods fail in that if wrong assumptions are made about the functional form and error term distributions, efficiency estimates can be confused by specification errors. Nevertheless, despite the advantages and disadvantages of these two groups of approaches, empirical research showed that they generally produced consistent results (e.g., Cummins and Zi 1998).

The information collected from using frontier analysis to evaluate the performance of insurance firms can be utilized in three main ways. One way is to enhance managerial performance by recognizing "best practices" and "worst practices" and informing management about the effects of new strategies and technologies. A second application of using this information is to provide guidance to regulators and policymakers regarding the appropriate response to problems and developments in an industry or the overall economy. Testing economic hypotheses represents the third main way of using this information, contributing particularly to advancing knowledge in insurance economics (see Cummins and Weiss 2013; Rubio-Misas 2022a).

After this introduction, the rest of the chapter is structured as follows: Sect. 2 shows a general discussion of the issues related to the frontier efficiency and productivity of insurance firms; Sect. 3 presents an overview of the research contributions on efficiency and productivity in the insurance industry using frontier efficiency and productivity methods in the period 2012–2022; Sect. 4 discusses the analyzed issues/tested hypotheses as well as the most important findings from these empirical studies to 12 different application areas as well as trends and gaps; and Sect. 5 provides a summary and conclusions.

2 Evaluating the Performance of Insurance Firms Using Frontier Efficiency and Productivity Analysis[2]

Evaluating the performance of insurers using frontier efficiency and productivity analysis involves separating those production units that perform well from those that perform poorly. This is done by

[2] This section is an updated version (focused on insurers) of the section "Evaluating performance of financial institutions using frontier efficiency and productivity analysis" conducted by Rubio-Misas (2022a).

applying parametric or nonparametric techniques to firms inside the industry or to branches inside a firm. Frontier analysis constitutes an advanced way to "benchmark" the relative performance of production units. It allows selecting "best practice" firms inside the insurance industry (or "best practice" branches inside the insurer) to give numerical efficiency values, recognize areas of input overuse and/or output underproduction, and connect these findings to questions of government policy or academic research (see Berger and Humphrey 1997).

In general, studies on the frontier efficiency of financial institutions depend on accounting measures of outputs, inputs, costs, revenues, or profits to assign efficiency relative to the best practices inside the available sample, since engineering information on the technology of these institutions is not available (Berger and Humphrey 1997). There are three main methods of choosing the outputs that financial institutions produce: financial intermediation, user cost, and value-added approaches (Berger and Humphrey 1992). Under the financial intermediation approach, financial institutions are treated only as financial intermediaries between liability holders and those who receive funds from financial institutions. The user cost method decides whether a financial product is an input or output by its net contribution to the revenues of financial institutions. If a financial product yields a return that exceeds the opportunity cost of funds or if the financial cost of a liability is less than the opportunity cost, then the product is considered to be a financial output. Otherwise, it is considered a financial input (Hancock 1985). The value-added approach describes output as those activities that have significant value added as judged using operating cost allocations (see, e.g., Berger et al. 1987).

Regarding these approaches to the insurance industry, the intermediation approach views insurers as pure financial intermediaries that borrow funds from policyholders, invest them in capital markets, and pay out claims, taxes, and costs. As Cummins and Weiss (2013) claim, this approach is not likely to be suitable for either nonlife or life insurers because it could ignore other services (e.g., insurance services) apart from intermediation services. Concerning the user cost approach, Cummins and Weiss (2013) discuss that this method is theoretically sound, but it is difficult to use in practice since insurance policies bundle many services that are priced implicitly together. They consider that a modified version of the value-added approach is the most suitable method for studying insurance efficiency and is also the most utilized method.[3]

As stated above, there are two main approaches in efficient frontier analysis: econometric approaches and mathematical programming approaches. Econometric approaches specify a production, cost, revenue, or profit function with a specific shape and usually make assumptions about the distributions of the inefficiency and error terms. There are three principal types of econometric frontier approaches: the stochastic frontier approach (SFA), the distribution-free approach (DFA), and the thick frontier approach (TFA). The SFA assumes a composed error model where inefficiencies follow an asymmetric distribution (e.g., half-normal, exponential, or gamma), and the random error term follows a symmetric distribution, which is usually normal. The DFA makes fewer specific assumptions but requires several years of data. The efficiency of each company is assumed to be stable over time, and the random noise averages out to zero. The TFA does not make any distributional assumptions for the random error and inefficiency terms but assumes that inefficiencies differ between the highest and lowest quartile firms (see, e.g., Kumbhakar and Lovell 2000). The mathematical programming approaches put significantly less structure on the specification of the efficient frontier and do not decompose the efficiency and error terms. Data envelopment analysis (DEA) is the most used mathematical programming approach, which employs linear programming to measure the relationship of produced outputs to assigned inputs and determines the efficiency score as an optimization result. The free disposal hull (FDH) approach is a special configuration of DEA where

[3] Leverty and Grace (2010) show that the value-added approach is closely related to traditional measures of insurer performance and that efficient value-added approach firms are less likely to go insolvent.

the convexity assumption on the efficient frontier is relaxed. DEA models for estimating technical efficiency can be specified under the assumptions of constant returns to scale (CRS), variable returns to scale (VRS), and nondecreasing returns to scale (NDRS). These models can be used to decompose technical efficiency into pure technical efficiency and scale efficiency as well as to estimate if a firm operates with constant, increasing, or decreasing returns to scale. The stochastic frontier approach (SFA) is the most used econometric method.[4]

Both methodologies (SFA and DEA) have strengths and limitations. The main strength of the SFA is that firms are allowed to deviate from the frontier due to random error as well as inefficiency, whereas DEA measures all departures from the frontier as inefficiency. However, the SFA has the limitation that if wrong assumptions are made about the functional form and error term distributions, efficiency estimates can be confused by specification errors. Since DEA does not incorporate a random error term in the efficiency estimation model, it is especially sensitive to measurement error, making it less efficient in dealing with outliers than SFA. That is, with the DEA approach, if one or a few inputs of operating units are understated or their outputs are overstated, those units may become outliers that alter the shape of the frontier and decrease the efficiencies of nearby units (Cummins and Weiss 2013). Nevertheless, the DEA approach has the advantage that it is individual firm based, making it easy to decompose efficiency by firms, which is particularly convenient for studying several issues in the insurance industry, such as scope economies (see Cummins et al. 2010). This is because, while DEA solves the optimization problem separately for each firm in the sample (Ray 2004), SFA optimizes over the sample as a whole. The latter implies that the estimated function is assumed to apply to all units of the sample, with all the differences among firms captured through the estimated residuals. Additionally, DEA provides an appropriate way to decompose cost and revenue efficiency into pure, technical, scale and allocative efficiency, allowing a detailed analysis of the causes of inefficiency.

Both methodologies have attractive statistical properties. This is very well known for SFA, but relatively recently known for DEA. However, DEA is equivalent to maximum likelihood estimation (Banker 1993), and its estimators are consistent and converge faster than estimators from other frontier methods (Grosskopf 1996). Banker and Natarajan (2008) demonstrate that DEA is a nonparametric stochastic methodology that performs better than parametric procedures in the estimation of individual decision-making unit productivity. Additionally, Banker and Natarajan (2008) and Banker et al. (2019) show that the two-stage procedure utilized in many DEA applications, where DEA efficiency estimates are regressed on firm characteristics and other covariates, yields consistent estimates. They also demonstrate that the two-stage DEA regression procedure performs as well as the best econometric method.

Nevertheless, several characteristics of the data may affect the efficiency estimates in both methods, as happens with sample size and imprecision in the measurement of inputs, outputs, and prices. For both SFA and DEA, a small sample size can inflate efficiency estimates if relatively efficient firms have been omitted from the sample. Nevertheless, the precision of DEA estimates is not affected by sample size although efficiency scores tend to decline as the sample size increases because there are more units that can enter a given reference set. In the case of SFA, a small sample size seems to be more problematic. As the sample size diminishes, coefficient estimates become less efficient and consequently statistically less reliable. The imprecision in the measurement of inputs, outputs, and prices also influences the efficiency estimates of both methods. This may happen because of data reporting or measurement errors but usually occurs because insurers' data are reported at a high level of aggregation (for instance, in many countries, data are available for life and nonlife insurance but

[4] Cummins and Weiss (2013) provide an extensive discussion on the main frontier methodologies that have been developed to measure efficiency and productivity change over time.

not for more finely gradated definitions). Yet, it is advisable to measure insurance output at the line of business level or aggregate lines with similar characteristics to prevent, for instance, a firm operating in nonlife lines of business with relatively high operating expenses from appearing inefficient because it is benchmarked with all firms operating in the nonlife insurance segment and not with the firms that operate in the nonlife lines with high operating expenses.

Different frontiers can be estimated, considering that the frontier is formed by the most efficient units of the reference set. The production frontier is the most basic efficient frontier. For its estimation, only the output and input quantities are needed. However, if data on input (output) prices are available, one may estimate the cost (revenue and profit) frontier. While in estimating production and cost frontiers, an input orientation is usually adopted, in estimating revenue and profit frontiers, an output orientation is normally followed. For instance, the production frontier is generally estimated based on the assumption that the firm is minimizing input use, conditional on output levels. The cost frontier is normally estimated based on the assumption of cost minimization and involves choosing the optimal quantities of inputs to produce a given output vector. However, the revenue frontier is usually determined based on the assumption that the firm is maximizing revenues, conditional on inputs, and involves choosing the optimal quantities of outputs, conditional on the input vector. The profit frontier is also usually determined based on the assumption that the firm is maximizing profits and involves the optimal choice of inputs and outputs, conditional on output and input prices. The choice of input versus output orientation is based on the microeconomic theory establishing the firm's objective in maximizing profits by minimizing costs and maximizing revenues (Cummins and Weiss 2013).

Efficiency is measured relative to the estimated *best practice frontiers.* It generally refers to the success of the firm in minimizing costs, maximizing revenues, or maximizing profits. The efficiency score is usually standardized between zero and one, with an efficiency score of one indicating a fully efficient firm. However, profit efficiency is usually not constrained between zero and one. There are several concepts of efficiency. Technical efficiency is defined as the ratio of the input usage of a fully efficient firm producing the same output vector to the input usage by the analyzed firm. The cost efficiency for a specific firm is calculated as the ratio of the costs of a fully efficient firm with the same output quantities and input prices to the specific firm's actual costs. Cost efficiency is the product of technical and input allocative efficiency. Thus, input allocative efficiency is the ratio of cost efficiency to technical efficiency and gives information on whether the firm uses the optimal mix of inputs. Revenue efficiency is defined as the ratio of the revenues of a specific firm to the revenues of a fully efficient firm with the same input vector and the same output prices. Revenue efficiency is the product of output technical efficiency and output allocative efficiency. Therefore, the output allocative efficiency can be calculated by the ratio of revenue efficiency to output technical efficiency and gives information on whether the firm uses the optimal combination of outputs. Profit efficiency measures how close a firm gets to generating the maximum possible profits, given the input prices and output prices and their comparison with the best practice frontier (Cummins and Weiss 2013).

In addition to efficiency, total factor productivity (TFP) is often analyzed in frontier efficiency studies to provide a dynamic analysis of performance. TFP is defined as an index of the total quantity of outputs produced divided by an index of total inputs used (Fare et al. 2008). As stated in Cummins and Weiss (2013), the concepts of TFP and efficiency are related. TFP is determined by the optimal production technology available to be employed in producing outputs as well as the efficiency with which firms use the technology. Consequently, TFP change has two major components: technical change and efficiency change. While technical change is given by a shift in the production frontier, efficiency change is given by an index of a firm's efficiency relative to the present and past frontiers. In general, TFP change can be estimated using both econometric approaches and mathematical

programming approaches, the most frequent being the Malmquist index approach, which is estimated using DEA.[5]

3 Survey of Research on Frontier Efficiency and Productivity in the Insurance Industry from 2012 to 2022

This section includes a survey of research on efficiency and productivity in the insurance industry using frontier methods that has been published in journals belonging to the Clarivate Journal of Citation Reports from 2012 to 2022. I have identified 46 empirical papers. Table 1 provides information for each of these papers on (1) country or countries analyzed; (2) estimation method; (3) type of frontier estimated; (4) type of institution evaluated; (5) time period covered by the sample; (6) number of units per year analyzed; and (7) journal where it was published. Table 2 summarizes these empirical studies in terms of the frontier estimated, utilized methodology, analyzed segment, output used, country or countries analyzed, and year of publication.

The definition of outputs, inputs, and their prices is the main step to use frontier methods. Cummins and Weiss (2000, 2013) provide the concept of insurance output from a theoretical perspective and define insurance outputs, output prices, inputs and input prices in practice. They claim that a modified version of the value-added approach is the most appropriate method for analyzing insurer efficiency. Risk-pooling and risk-bearing services, real financial services related to insured losses, and intermediation services are the three principal services in creating value under this approach. Cummins and Weiss (2013) highlight that the real value of incurred benefits plus addition to reserves for the life segment and the real value of losses incurred for the nonlife segment are satisfactory proxies for the amount of risk pooling/bearing and real insurance services. As stated before, the utilization of different output variables for the major lines of business offered by insurers in both segments (life and nonlife) is advisable. They also consider that the real value of invested assets is the most suitable output variable to proxy for the intermediation function. The prices of the insurance output variables are defined as premiums minus output divided by output, while the expected rate of return on the insurer's assets is the definition of the price for the intermediation function.

Appendix 1 presents the outputs, output prices, inputs and input prices used in the surveyed studies. A summary of the outputs is also presented in Table 2. The majority of studies have used the definitions of outputs and output prices discussed in Cummins and Weiss (2000, 2013). Most (28) used real incurred losses and/or real incurred benefits as the insurance output measure as well as real invested assets (23) as the intermediation output measure, indicating that when information is available, most studies have used the definitions of outputs and output prices recommended in the literature (see, e.g., Cummins et al. 1999a, b; Cummins and Rubio-Misas 2006; Cummins et al. 2009; Leverty and Grace 2010; Eling and Jia 2019). Nevertheless, it is remarkable here that eight papers used a two-stage network process to estimate operating efficiency where outputs from the first stage (intermediate measures) become the inputs for the second stage.

Within the context of the value-added approach, insurers utilize three main inputs: labor, material and business services, and capital (see Cummins and Weiss 2013). As labor and equity capital are quantitatively quite important for insurance firms, we observe that they are the two most frequently utilized inputs in the surveyed studies. We also notice that sometimes the labor input and the material and business services input are combined due to data unavailability (e.g., Huang and Eling 2013; Cummins and Rubio-Misas 2022). Furthermore, sometimes two types of capital are taken into

[5] Cummins and Weiss (2013) provide an ample discussion of the Malmquist index approach to analyze the TFP change of firms and its components over time.

Table 1 Empirical studies on efficiency and productivity in the insurance industry using frontier methods (2012–2023)

Author	Country	Method	Frontier type	Type of institution	Sample period	No. of units/year	Publication
Leverty and Grace (2012)	USA	DEA	Production/cost/revenue	P-L	1989–2000	1409	JRI
Biener and Eling (2012)	21 countries	DEA	Production/cost	Life and nonlife	2002–2006	4761/3208/1552	EJOR
Cummins and Xie (2013)	USA	DEA	Production/cost/revenue	P-L	1993–2009	781	JPA
Huang and Eling (2013)	4 countries	DEA	Production	Nonlife	2000–2008	819	EJOR
Gaganis et al. (2013)	52 countries	SFA	Cost/profit	Life and nonlife	2003–2008	295	JPA
Abdul Kader et al. (2014)	17 countries	DEA	Production/cost	Takaful	2004–2007	45	IRFA
Kao and Hwang (2014)	Taiwan	DEA	Production	Nonlife	2000–2002	21	EJOR
Lu et al. (2014)	China	DEA	Production	Life	2006–2010	34	Omega
Chen et al. (2014)	Malaysia	DEA	Production	P-L	2008–2011	16	EM
Ma et al. (2014)	USA	DEA	Cost/revenue	P-L	1993–2008	229/119	GP
Yaisawarng, Avsavadachanukorn and Yaisawarng et al. (2014)	Thailand	SFA	Cost	Nonlife	2000–2007	46	JPA
Grace et al. (2015)	USA	DEA	Cost/revenue	Life and nonlife	2004 and 2006	366/963	JRI
Alhassan and Biekpe (2015)	South Africa	DEA	Production	Nonlife	2007–2012	76	GP
Dreassi and Schneider (2015)	Italy	SFA	Cost	Life	1998–2012	86	GP
Biener et al. (2016)	Switzerland	DEA	Production/cost/revenue	Life and nonlife	1997–2013	21/51/37	EJOR
Lim and Zhu (2016)	Taiwan	DEA	Production	Nonlife	Av. 2001–2002	24	EJOR
Wu et al. (2016)	Malaysia	DEA	Production	Life and nonlife	2008–2013	34	JORS
Wanke and Barros (2016)	Brazil	DEA	Production	Life and nonlife	1995–2013	97	EM
Rubio-Misas and Fernández-Moreno (2017)	Spain	DEA	Cost	Life and nonlife	2005–2012	203	SJFA
Bikker (2017)	The Netherlands	SFA	Cost	Health	1995–2012	60/53	GP
Eling and Schaper (2017)	14 countries	DEA	Production/cost	Life	2002–2013	970	EJOR
Biener et al. (2017)	Global reinsurance market	DEA	Production/cost/revenue	Life and nonlife	2002–2012	76	JBF
Karbhari et al. (2018)	21 countries	DEA/FDH	Production	Takaful	2002–2013	134	JIFMIM
Danquah et al. (2017)	Ghana	SFA	Cost	Life and nonlife	2005–2014	30	MDE
Ferro and León (2018)	Argentina	SFA	Production	Nonlife	2009–2014	83	GP

(continued)

Table 1 (continued)

Author	Country	Method	Frontier type	Type of institution	Sample period	No. of units/year	Publication
Eling and Jia (2018)	16 countries	DEA	Production	Life and nonlife	2006–2013	1346	IRFA
Nourani et al. (2018)	Malaysia	DEA	Production	Life and nonlife	2005–2012	27	SIR
Reyna and Fuentes (2018)	Mexico	DEA	Production/cost	Life and nonlife	2001–2013	57/81	JPA
Alshammari et al. (2019)	6 countries	SFA	Cost	Life and nonlife	2009–2016	86	RIBR
Eling and Jia (2019)	91 countries	DEA	Production	Life and nonlife	2003–2013	1616	PBFJ
Shavitha et al. (2019)	India	DEA	Production	Life	2012–2017	17	GP
Tone et al. (2019)	Malaysia	DEA	Production	Life and nonlife	2008–2016	30	Omega
Lee et al. (2019)	Malaysia	DEA	Production/cost	Takaful	2011–2015	19	IJIMEFM
Lim and Zhu (2019)	Taiwan	DEA	Production	Nonlife	Av. 2001–2002	24	Omega
Ohene-Asare et al. (2019)	Ghana	DEA	Cost	Life and nonlife	2005–2014	14/15	GP
Al Amri et al. (2021)	19 Islamic countries	DEA	Production/cost	Takaful	2004–2009	44	JIFMIM
Peng and Lian (2021)	2 countries	DEA	Production	Life	2008–2016	80	GP
Altuntas et al. (2021)	Germany	DEA	Production/cost/revenue/profit	P-L	1999–2009	84	AOR
Lim et al. (2021)	Malaysia	DEA	Production	Life and nonlife	2000–2017	56	GP
Cummins and Rubio-Misas (2021)	10 European countries	DEA	Cost/revenue	Life	1998–2014	505	EAP
Ho and Hsu (2021)	Taiwan	DEA	Production	Life	2012–2017	23	GP
Reyna et al. (2022)	Mexico	DEA	Production	Life and nonlife	1997–2018	46/71	GP
Cummins and Rubio-Misas (2022)	10 European countries	DEA	Cost/revenue	Life	1998–2014	505	AOR
Shaddady (2022)	16 countries	DEA/SFA	Production	Life and nonlife	2007–2017	125	GP
Rubio-Misas (2022)	Spain	DEA	Cost/profit	Life and nonlife	2002–2009	264/331	IJBM
Nourani et al. (2022)	China	DEA	Production	Life and nonlife	2014–2018	32	GP

Abbreviations: *DEA* data envelopment analysis, *SFA* stochastic frontier analysis, *FDH* free disposal hull, *P-L* property-liability, *JRI* Journal of Risk and Insurance, *EJOR* European Journal of Operational Research, *JPA* Journal of Productivity Analysis, *IRFA* International Review of Financial Analysis, *GP* The Geneva Papers, *JIC* Journal of Intellectual Capital, *JORS* Journal of Operational Research Society, *EM* Economic Modelling, *SJFA* Spanish Journal of Financial Analysis, *JBF* Journal of Banking and Finance, *JIFIM* Journal of International Financial Markets Institutions and Money, *MDE* Management Decisions Economics, *SIR* Social Indicators Research, *RIBR* Research in International Business and Finance, *PBFJ* Pacific-Basin Finance Journal, *IJIMEFM* International Journal of Islamic and Middle Eastern Finance and Management, *EAP* Economic Analysis and Policy, *IJBM* International Journal of Bank Marketing, *AOR* Annals of Operations Research

Table 2 Summary review of efficiency and productivity studies in the insurance industry from 2012 to 2022

Studies by type of frontier		**Studies by methodology**	
Frontier type	**No. of papers**	**Method**	**No. of papers**
Cost	24	DEA	37
Revenue	9	SFA	7
Production	33	FDH	1
Profit	3	DEA/SFA	1
Studies by insurance segment		**Studies by insurance output**	
Segment	**No. of papers**	**Output**	**No. of papers**
Life and nonlife	21	Losses or benefits	28
Nonlife (P-C)	12	Invested assets	23
Life	8	Premiums	17
Takaful	4	Investment income	9
Health	1		
Studies by country		**Studies by year**	
Country	**No. of papers**	**Year**	**No. of papers**
Multicountry	15	2012	2
Malaysia	6	2013	3
Taiwan	4	2014	6
USA	4	2015	3
China	2	2016	4
Ghana	2	2017	4
Mexico	2	2018	6
Spain	2	2019	7
Argentina	1	2021	6
Brazil	1	2022	5
Germany	1		
India	1		
Italy	1		
The Netherlands	1		
South Africa	1		
Switzerland	1		
Thailandia	1		

account: equity capital and debt capital (e.g., Grace et al. 2015; Altuntas et al. 2021; Rubio-Misas 2022b). As input prices, it is common to use indices such as the wage rate for labor, the business services deflator for material and business services, and the expected market return on equity capital for equity capital or a Treasury bill rate for debt capital. In addition, because physical measures of input quantity are usually not publicly available, the most frequent way to approach the quantity of physical inputs is dividing the expense item by a corresponding price index. Regarding inputs, I also notice that five papers in the context of a dynamic DEA model used carry-over variables, which are those continuously held from one term to another, reflecting dynamic components.[6] The most used

[6] The dynamic DEA model (see Tone and Tsutsui 2010) deals with the long-term dynamic process of a company by incorporating carry-over inputs, enabling the measurement of period-specific efficiency based on long-time optimization.

carry-over variables are equity capital and debt capital. That is, for instance, equity capital of an insurer is carried forward from one period to another continuously.

Table 2 shows the type of estimated frontier in the surveyed papers. The frontier is formed by the most efficient units of the reference set. The production frontier (with 33 studies including the 8 papers that use a two-stage network process) is the most estimated, followed by the cost frontier (with 24 studies), the revenue frontier (with 9 studies), and the profit frontier (with 3 studies). Table 2 also presents the methodology utilized to measure efficiency and productivity change. The DEA approach was the most frequently used method in the surveyed papers (37 studies), followed by SFA (7 studies).

Tables 1 and 2 present additional information. I notice that most reviewed papers focused on both insurance segments: life and nonlife. That is, out of the 46 surveyed papers, 21 analyzed both segments, 12 focused on the nonlife segment, 8 on the life segment, 4 on takaful (Islamic insurance) and 1 on health insurers. Regarding the analyzed countries, the majority of the reviewed papers are multicountry studies (15). Within the papers that analyzed one single country, 6, 4, and 4 are devoted to studying Malaysia, Taiwan, and the USA, respectively. In terms of publications, *The Geneva Papers* is the outlet where many of the reviewed articles have been published. Out of 46 papers, 13 were published in *The Geneva Papers*, 6 in the *European Journal of Operational Research*, 4 in the *Journal of Productivity Analysis*, and 3 in *Omega*.

4 Main Analyzed Topics, Hypotheses, and Findings

This section discusses the main analyzed topics, hypotheses tested, and principal findings of the surveyed papers on insurers' performance using frontier efficiency and productivity methods. I have distinguished 12 different application areas. Although many studies have made contributions to more than one topic, I tried to focus on the main fields of application. Table 3 provides information on the articles that conform to the different topics by specifying the main issue analyzed and/or hypotheses tested as well as the principal findings.

4.1 General Level of Efficiency and Relevance of Efficiency Measures

A first use of efficiency frontier methods was to evaluate the general level of efficiency in the insurance industry. Although most recent empirical studies in this field contribute to other topics in addition to providing information on the level of efficiency, papers continue to be published in which the latter is their main contribution. This is the case for the Danquah et al. (2017) and Ferro and León (2018) papers. Danquah et al. (2017) analyzed the cost efficiency of insurance firms in Ghana from 2005 to 2014 using SFA and found that the overall average cost inefficiency for the period was 53.8%, indicating a challenge of cost inefficiency in the insurance industry. Ferro and León (2018) estimated the technical efficiency of Argentina's nonlife insurance industry over the period 2009–2014 using SFA and found that the average technical efficiency yielded a value of 42%.

Additionally, in this area, I include the papers by Gaganis et al. (2013) and Eling and Jia (2019), which focused on the relevance of frontier efficiency measures. More precisely, Gaganis et al. (2013) analyzed whether the capital market valued firms' efficiency (both cost and profit) in a sample of listed insurers from 52 countries during the period 2002–2008 using the econometric approach of Battese and Coelli (1995). They showed a positive and statistically significant relationship between profit efficiency change and stock returns but no robust evidence that cost efficiency change is associated with stock returns. These results are particularly of interest for security analysts, firm managers and

Table 3 Topics analyzed, main hypotheses, and selected findings

Author	Issue/hypothesis	Selected findings
General level of efficiency and relevance of efficiency measures		
Danquah et al. (2017)	Cost efficiency of insurers from the perspectives of developing economies	Insurers operate with about 52.8% average cost inefficiency. Cost inefficiencies are explained by firm size, market share, capitalization, reinsurance, regulation, and business type
Ferro and León (2018)	Technical efficiency in the Argentina nonlife insurance industry	The average technical efficiency yielded a value of 42 percent
Gaganis et al. (2013)	The link between efficiency and stock returns	A positive and statistically significant relationship between profit efficiency change and stock returns
Eling and Jia (2019)	The link between efficiency and profitability	A significantly positive correlation between efficiency and profitability. This relationship was nonlinear and industry dependent
Solvency		
Leverty and Grace (2012)	Whether managers impact firm performance	A correlation exists between inefficient management and the likelihood of failure
		Superior managers are able to remove their firms from regulatory scrutiny sooner than inferior managers
		Managerial quality decreases the ultimate cost of insolvency
Rubio-Misas and Fernández-Moreno (2017)	The link between cost efficiency and insurer regulatory solvency ratio	Cost efficiency negatively affects the level of the regulatory solvency ratio
Eling and Jia (2018)	The impact of technical efficiency on failure probability	Technical efficiency negatively impacts the probability of failure
Al Amri et al. (2021)	The link between efficiency and solvency risk	More efficient insurers have larger distance to default
Scale and scope economies		
Cummins and Xie (2013)	Scale economies	Majority of firms below median size are operating with IRS. Majority of firms above median size are operating with DRS. A significant number of firms in each size decile achieved CRS. More diversified firms achieved efficiency and productivity gains
Yaisawarng et al. (2014)	Scale economies	Most insurers experienced economies of scale as they expanded their capital funds to meet new regulations
Alhassan and Biekpe (2015)	Scale economies	About 20% of the nonlife insurers operate at their optimum capacity. The majority of large insurers operate under DRS, but small, medium, and big insurers largely operate under IRS

(continued)

Table 3 (continued)

Author	Issue/hypothesis	Selected findings
Dreassi and Schneider (2015)	Scale economies	Bancassurers do not outperform independent insurers in terms of scale economies. Before 2007, bancassurers and joint ventures showed modest scale diseconomies while independent insurers appeared in equilibrium
Biener et al. (2016)	Scale economies	Firms from the "small" category mostly operate under IRS and CRS while those from the "large" group mostly operate under DRS
Bikker (2017)	Scale economies	After the health care reform of 2006, unused scale economies at around 20 percent were much higher than before the reform (4%), pointing to a relative increase in fixed costs
Biener et al. (2017)	Scale and scope economies	Scale-efficient reinsurers exhibit an optimal size that is larger than those for the primary insurance market. A focus strategy is appropriate for small reinsurers, whereas a product diversification strategy may suit some large reinsurers
Shavitha et al. (2019)	Scope economies	Evidence on the strategic focus hypothesis
Ohene-Asare et al. (2019)	Scale economies	Not many insurers have been operating at the optimal production scale over the period
Al Amri et al. (2021)	Conglomeration vs. strategic focus	Strategic focus is superior to diversification for takaful in terms of efficiency
Peng and Lian (2021)	If there is an optimal business model for three business dimensions: assets, funding, and income	The strategic focus is superior in terms of assets and income to diversification
Ho and Hsu (2021)	The impact of the introduction of microinsurance on efficiency	The introduction of microinsurance reduced pure technical efficiency of individual insurers
Organizational form, corporate governance		
Biener and Eling (2012)	Expense preference vs. efficiency structure	Evidence for the efficiency structure hypothesis in selected market segments
Abdul Kader et al. (2014)	The link between cost efficiency and board composition	The relative influence of board composition (e.g., the proportion of nonexecutive directors on the board) on the cost efficiency of takaful insurers depends on its interaction with other firm-specific characteristics such as board size
Karbhari et al. (2018)	The relationship between corporate governance attributes and technical and scale efficiencies on the global Takaful insurance operators	CEO/chair duality and board size increase technical efficiency. Nonexecutive directors, Shari'ah board, and institutional ownership improve scale efficiency
Alshammari et al. (2019)	Differences in efficiency between conventional insurers and takaful operators	Takaful operators are less efficient than conventional insurers
Lee et al. (2019)	The impact of corporate governance factors on cost efficiency	Board size and the proportion of nonexecutive directors impact cost efficiency negatively
Al Amri et al. (2021)	Tests differences between the mudharaba (profit-sharing) and the wakala (fee-based) model	Profit-sharing model performs better than the fee-based model

Risk management		
Grace et al. (2015)	Which aspects of enterprise risk management improve efficiency	The use of economic capital models, dedicated risk managers, and requiring the risk manager to report to the board and CEO improve operating performance
Wu et al. (2016)	The impacts of risk management committee characteristics and prestige on efficiency	A positive relationship among the risk management committee characteristics (number of directors, the proportion of independent directors, the number of titles of honor carried by directors) and efficiency. The prestige of the risk management committee could minimize the positive effect
Altuntas et al. (2021)	Tests if enterprise risk management (ERM) facilitates economies of scale and scope	ERM facilitates economies of scale and economies of scope with respect to revenue complementarities
Other management issues		
Lu et al. (2014)	The link between intellectual capital and efficiency	Intellectual capital affected operating efficiency positively in a truncated regression model
Biener et al. (2016)	Tests whether international diversification impacts efficiency	The internationalization of the Swiss insurance business has a positive impact on life insurer efficiency, but not for p/c. Entering the non-European market for Swiss life and p/c insurers is not as beneficial as entering European markets
Nourani et al. (2018)	To estimate the intellectual capital efficiency at three levels in the insurance industry	Deficiencies occurred in the human and structural capital stages as opposed to physical capital stage
Distribution systems		
Ma et al. (2014)	The impact of abandonment on the use of contingent commissions on insurer performance	The abandonment of the use of contingent commissions before (after) the Spitzer investigation was (was not) rewarded with higher performance. These results were attributed to the cost of reputation preservation
Rubio-Misas (2022a, b)	To investigate why bancassurance coexists with alternative insurance distribution channels	Results support the product-quality arguments. The use of the banking channel was found to increase cost efficiency, but the distribution channel(s) utilized did not affect profit inefficiency
Market structure		
Biener et al. (2017)	Tests the efficient structure hypothesis	Results support the efficient structure hypothesis: cost-efficient reinsurers can charge lower prices without sacrificing profitability
Alshammari et al. (2019)	The impact of market concentration variables on cost efficiency	Conventional insurers and takaful undertakings operating in more concentrated markets tended to be less efficient in supporting the Quiet Life hypothesis

(continued)

Table 3 (continued)

Author	Issue/hypothesis	Selected findings
Regulatory change		
Alhassan and Biekpe (2015)	Whether regulatory reforms improved efficiency	A slower catch-up effect, implying that less efficient insurers were able to catch up with efficient insurers
Reyna and Fuentes (2018)	Whether deregulation improved performance	An increase in costs arising from changes in the prices and quantities of inputs
Ohene-Asare et al. (2019)	Whether the introduction of the Insurance Act of 2006 improved cost productivity growth	After the introduction of the Act, there was cost productivity growth between 2007 and 2009 and from 2010 to 2011
Lim et al. (2021)	The impact of the implementation of risk-based capital requirements on the performance of the Malaysian insurance sector	A different effect of the implementation of risk-based capital on the performance of conventional insurers compared to takaful firms
Cummins and Rubio-Misas (2022)	Integration and convergence in efficiency and technology gap	Convergence in cost and revenue efficiency. Convergence in cost/revenue technology gap. Financial crisis affected convergence in efficiency and technology gap negatively in terms of costs but not in terms of revenues
Environmental conditions		
Huang and Eling (2013)	Whether environment affects the efficiency of nonlife insurers	The environment variables growth of GDP, consumer price index, and deposit interest rate are adverse environmental factors with respect to efficiency
Eling and Schaper (2017)	The impact of changing environment on efficiency and productivity	General economic, capital market, and market conditions were important drivers of efficiency
Cummins and Rubio-Misas (2021)	Tests the role of financial market development and institutional quality on the integration of EU life insurance markets	National stock market development and institutional quality enhance cost performance and integration of EU life insurance markets
		Better outcomes in national institutional quality decrease the metatechnology revenue efficiency ratio
Reyna et al. (2022)	The impact of low interest rate on efficiency	There is an inverse relationship between interest rate and efficiency
Shaddady (2022)	The influence of business environment, political risk, governance, and Shari'ah compliance on efficiency in the insurance market across MENA countries	Interest rate, inflation, oil price, and financial freedom show a negative effect on efficiency, while unemployment, economic growth, effective exchange rate, political risk, and the quality of regulation show a positive impact
Productivity analysis		
Cummins and Xie (2013)	Productivity change and components	Significant gains in TFP. Higher technology investment is positively related to efficiency and productivity improvements

Huang and Eling (2013)	Productivity change and components	An increase in technical change and a decrease in technical efficiency change
Kao and Hwang (2014)	To calculate a common-weight global Malmquist productivity index (MPI) for measuring performance changes between two periods	The MPI of a DMU of two periods is the product of those of the two processes (premium acquisition and profit generation) between the same periods
		The performance of the analyzed insurers improved during the analyzed period
Chen et al. (2014)	The impact of intellectual capital on productivity	Intellectual capital is positively related to changes in productivity
Ma et al. (2014)	The effect of abandonment on the use of contingent commissions on insurer total factor productivity (TFP) change	Firms that abandoned contingent commissions sometime during 1994–2003 experienced TFP gains. Abandoners experienced similar amounts of reduction in TFP as nonabandoners during the post-Spitzer period
Yaisawarng et al. (2014)	Total factor productivity change and its components	A regress in technology from 2000 to 2004 was the dominant contributor to a negative TFP growth of 1.15%. Beginning in 2005, the industry experienced technological progress at a rate of 2%, contributing to TFP growth
Alhassan and Biekpe (2015)	Total factor productivity change and its components	Improvements in productivity of about 10%, mostly attributable to innovations in production of insurance outputs resulting in outward shifts of the production frontier
Biener et al. (2016)	Total factor productivity change and its components	Yearly average productivity increased in the p/c and reinsurance markets due to similar degrees of improvement in technology and efficiency. A decline in technology for life insurance firms gave rise to a decline in productivity
Eling and Schaper (2017)	Productivity change and components	An efficiency increase that leads to an increase in total factor productivity
Ferro and León (2018)	Technological change	Evidence of a negative technological change during the sample period
Reyna and Fuentes (2018)	Productivity effect and its components	Even though some companies achieved cost reductions from technological progress or improvements in efficiency, the market did not show signs of significant productivity gains
Shavitha et al. (2019)	Productivity change and components	Total factor productivity of the life microinsurance (LMI) portfolio insurance companies grew due to the contribution of technical efficiency rather than technical progress, while pure efficiency change was negative
Ohene-Asare et al. (2019)	Cost productivity growth and its components	Cost productivity growth of 3.2% over the sample period, being different for life insurers (2.3%) compared to nonlife insurers (4%). The cost productivity growth was basically driven by the progress in overall efficiency
Lim et al. (2021)	Total factor productivity change and its components	Total factor productivity change and technical efficiency change decreased by 1.1% on average over the period
Methodology issues		
Huang and Eling (2013)	Incorporation of uncontrollable variables in the efficiency analysis using multistage DEA	The model allows distinguishing between managerial inefficiency and inefficiency due to environmental conditions

(continued)

Table 3 (continued)

Author	Issue/hypothesis	Selected findings
Kao and Hwang (2014)	Proposing a multiperiod model which is able to measure the overall and period efficiencies at the same time and derive a mathematical relationship between them	The efficiency scores measured from the multiperiod model could be used to calculate a common weight global Malmquist productivity index (MPI) between two periods
Lu et al. (2014)	Applying the dynamic slack-based model (DSBM) to evaluate insurers´ operating efficiency to take into account carry-over activities	The importance of taking into account carry-over activities related to debt capital and equity capital
Lim and Zhu (2016)	Demonstrating that the duality in the standard DEA naturally migrates to the two-stage network DEA	Developing formulas to obtain frontier projections and divisional efficiency scores using a set of dual variables
Wanke and Barros (2016)	Proposing an integrative, two-stage approach involving the determination of a DEA meta-frontier in the first stage and the use of a number of data-mining techniques in the second stage	There is a heterogeneous impact on the efficiency levels of the geographical regions served and the type of insurance offered by insurance companies
Nourani et al. (2018)	Applying a network DEA slack-based model to address the connectivity between stages and time effect	Deficiencies occurred in the human and structural capital stages as opposed to the physical capital stage
Tone et al. (2019)	Modelling investment assets as a carry-over variable in a dynamic two-stage network DEA model	This model leads to improved discriminatory power of overall performance, particularly investment efficiency
Lim and Zhu (2019)	Tests whether and how the primal-dual correspondence between the multiplier model and the envelopment model can be obtained for two-stage network DEA models	Developing an axiomatic derivation of some two-stage network DEA models focusing on the basic two-stage serial process structure
Cummins and Rubio-Misas (2022)	Tests the suitability of using the meta-frontier framework	Presence of heterogeneity in production possibility sets among countries
Nourani et al. (2022)	Using different DEA frameworks to evaluate insurers' efficiency	Network models generally perform better in discriminating inefficient DMUs

investors since the efficiency performance measures appear to offer more information—as compared to traditional indicators—with regard to firms' ability to maintain and improve their profits relative to their peers. Eling and Jia (2019) evaluated the relationship between efficiency and profitability in a global dataset of life and nonlife insurers from 91 markets over the period 2003–2013. They estimated technical efficiency using the two-stage DEA model of Kao and Hwang (2008) and documented a significantly positive correlation between firm efficiency and profitability. They found that this relationship was nonlinear and industry dependent. The positive impact of efficiency became smaller as the firm approached to the best practice, and efficiency was more critical to the profitability of life insurers than to that of nonlife insurers. One of the implications of these results for managers of life insurers and of less efficient firms is that they should pay attention to their operational management because the increase in very low efficiency is easily translated into profit.

4.2 Solvency

One application of frontier efficiency analysis in insurance is to investigate if firm performance is likely to be related to solvency. The findings of this topic are particularly relevant as far as firm failure avoidance is concerned. Although this application area has not been distinguished in the Cummins and Weiss (2000, 2013) surveys, I have found 4 papers in the analyzed period in this chapter. Related to solvency, Leverty and Grace (2012), on a sample of property liability (P-L) insurers from the USA over the period 1989–2000, estimated production, cost, and revenue frontiers using DEA and tested if managerial ability was related to firm distress. First, they examined whether the CEO was able to minimize firm costs, maximize revenues, operate at an efficient scale, adopt the best practice technology and use inputs according to their marginal productivities as a definition of managerial ability. They showed that manager fixed effects play an economically important role in influencing firm efficiency. Then, they evaluated if CEOs influenced firm performance when their firm was in financial distress and found that managerial ability was inversely related to the amount of time a firm spent in distress, the likelihood of a firm's failure and the cost of failure, indicating that the managers of failed firms were less skilled than their counterparts.

Rubio-Misas and Fernández-Moreno (2017) estimated the cost efficiency of Spanish insurers over the period 2005–2012 using DEA and evaluated its impact on a firm's regulatory solvency ratio. They found that cost inefficient insurers maintained higher regulatory solvency ratios, probably due to the overutilization of equity capital. Eling and Jia (2018), using a sample of European insurers from 16 countries over the period 2006–2013, found that DEA technical efficiency was negatively correlated with the probability of failure. Al Amri et al. (2021) analyzed takaful firms in 19 countries and estimated efficiency using DEA. They found an inverse relationship between cost efficiency and insolvency risk. The results of the last two papers suggest that managers in the insurance industry should focus on enhancing the efficiency of their firms to avoid insolvency.

4.3 Scale and Scope Economies

Economies of scale are a principal topic in the area of frontier efficiency analysis, especially in the justification of mergers. It is important for a firm to achieve the optimal scale to realize the goals of minimizing costs and maximizing revenues. Scale economies exist if the average cost per unit of output decreases as the volume of output rises. Expanding firms' fixed costs over a larger volume of output is the most common source of scale economies. Insurers' fixed costs are related not only

to the costs of production such as computer systems but also to the costs of financial capital and managerial expertise. If income volatility is inversely related to size, working at a large scale may decrease insurers' cost of capital due to the risk-reducing impact of the law of large numbers in the insurance risk pools. Economies of scale may additionally arise if working at a large scale allows managers to become more specialized and consequently more proficient in developing different tasks (Cummins and Weiss 2013). Nevertheless, firm expansion may produce inefficiencies. Additional costs may arise, for instance, because the internal communication and control of large organizations need expensive systems in the management structure. In addition, larger organizations are more prone to building managerial conflicts and agency costs.

Cummins and Xie (2013) provided an extensive analysis of scale economies in their analysis of the USA property-liability insurance industry over the period 1993–2009 using DEA. They found that most insurers in the six smallest size deciles operated with IRS while most insurers in the four largest deciles operated with DRS. The results also show that in every size decile, at least 6% of insurers operated with CRS, indicating that it is possible to realize CRS even for large insurers. They concluded that further consolidation of the insurance industry cannot be justified on the grounds of scale economies in the five largest size deciles. However, the substantial number of firms in the five smallest size deciles operating with IRS suggested that such firms may be able to benefit from consolidation or organic growth. Yaisawarng et al. (2014) estimated a stochastic cost frontier for nonlife insurers from Thailand from 2000 to 2007 and examined whether scale economies existed. They found that these firms were, on average, 74–79% cost efficient and that most of them experienced economies of scale as they expanded their capital funds to meet new regulations. Alhassan and Biekpe (2015) analyzed scale economies using the DEA approach in the South African nonlife insurance industry from 2007 to 2012. They found that about 20% of nonlife insurers operated at their optimum capacity and that while most large insurers operated under DRS, small, medium and big insurers largely operated under IRS. The existence of scale inefficiencies made the authors suggest regulatory action in the form of mergers and consolidation policies. Dreassi and Schneider (2015), using a Translog cost function, investigated differences in scale economies of different ownership models (independent insurers, insurers totally owned by banks (*bancassurers*) and joint ventures) in the Italian life insurance industry over the period 1998–2012. They found that *bancassurers* did not outperform independent insurers in terms of scale economies. Before 2007, *bancassurers* and joint ventures showed modest scale diseconomies while independent insurers appeared in equilibrium. They observed that changes in demand and shocks in the financial sector during the postcrisis period generated scale economies for all groups, particularly for independent insurers, suggesting that the level of scale economies may change significantly in a mature market when external shocks reshape market conditions.

Biener et al. (2016) analyzed scale economies in the Swiss insurance industry over the period 1997–2013 using DEA. They found that firms from the "small" category mostly operate under IRS and CRS, while those from the large group mostly operate under DRS, suggesting that mergers among small- and medium-sized insurers could be justified on efficiency grounds. Bikker (2017) investigated scale economies of the Dutch health-care insurance market over the period 1995–2012 using a Translog cost function and focusing on the impact of the 2006 health-care reform, which replaced the dual system of public and private insurance with a single compulsory health insurance scheme. He found that after the reform, unused scale economies at around 20% were much higher than before the reform (4%), pointing to a relative increase in fixed effects. Biener et al. (2017) evaluated economies of scale of the global reinsurance market over the period 2002–2012 using DEA. They showed that scale-efficient reinsurers of an optimal size were larger than those in the primary insurance market or other financial service industries. They also found that cost-efficient

reinsurers can charge lower prices without sacrificing profitability. These findings suggest that further consolidation is expected and that policymakers should be cautious when adopting anti-concentration measures in the global reinsurance market. Ohene-Asare et al. (2019) evaluated economies of scale in the Ghana insurance industry from 2005 to 2014 using DEA; the sample size is a probable limitation of this study. They found that not many insurers (both in the life and nonlife segment) were operating under CRS. Nevertheless, at the end of the period, 71% (62%) of life (nonlife) insurers operated under IRS.

Scope economies are another important issue in industrial economics, particularly relevant for insurers because of the existence of two different segments, each involving different lines of business. Two are the main hypotheses regarding economies of scope: the *strategic focus hypothesis* and the *conglomeration hypothesis*. The first maintains that firms can best add value by focusing on the core business, i.e., separate productions are more costly and/or revenue efficient. The second holds that operating diverse businesses can add value because of exploitation costs and revenue scope economies. Scope cost economies arise because of production complementarities, that is, the joint use of inputs. For instance, a single prospect list may be used by an insurer that writes both life insurance and nonlife insurance. The same could happen with executive talent and brand names, which are inputs that may give rise to production complementarities. However, scope revenue economies appear due to consumption complementarities because customers are willing to pay more to a joint producer to reward the amount of savings that they obtain in reducing searching costs or enhancing product quality (Cummins and Weiss 2013).

Biener et al. (2017) also studied economies of scope in their analysis of the global reinsurance industry. They concluded that a focus strategy is appropriate for small reinsurers, whereas a product diversification strategy may suit some larger reinsurers due to significant revenue scope economies. Shavitha et al. (2019) used the DEA approach to evaluate scope economies in the Indian insurance industry from 2012 to 2017 on a sample without the availability of appropriate input measures. They tested if insurers offering life microinsurance (LMI) along with life and health policies were more technically efficient than their counterparts who kept out of the LMI market. The results showed that product diversification had adverse effects on the technical efficiency of insurers, supporting the *strategic focus hypothesis*. In the same line, Ho and Hsu (2021), in their analysis of life insurers from Taiwan over the period 2012—2017, found that the introduction of microinsurance reduced the pure technical efficiency of individual insurers. Al Amri et al. (2021), in their analysis of takaful firms, tested the two main hypotheses regarding scope economies. They estimated efficiency using DEA and found that the *strategic focus* was superior to *conglomeration* for takaful in terms of efficiency, suggesting that regulators should be skeptical of scope economies as justification for mergers and acquisitions in the takaful industry. Peng and Lian (2021) evaluated life insurers in China and India from 2008 to 2016 using a game cross-DEA model, which comparatively measured insurers' performance using Nash equilibrium weights. They investigated whether there was an optimal business model for three business dimensions: assets, funding, and income. They found that specialization in terms of assets and income was more efficient than diversification.

4.4 Organizational Form and Corporate Governance

The effect of organizational form on performance is a well-developed field of research within frontier efficiency analysis in the insurance industry since two types of organizational forms coexist in many countries: stocks owned by stockholders and mutuals owned by customers, i.e., policyholders. The two most prominent hypotheses developed in this area are the *efficient structure hypothesis* and the

expense preference hypothesis. The *efficient structure hypothesis* claims that the two organizational forms are sorted into market segments where they have comparative advantages in minimizing costs and maximizing revenues due to differences in managerial discretion, maturity, and access to capital (see Mayer and Smith 1988). This argument predicts that stocks will be more successful in lines that require more managerial discretion (such as complex lines), while mutuals are expected to be more successful in lines that require less managerial discretion (such as standardized lines). However, the *expense preference hypothesis* states that mutuals will be less successful than stocks in minimizing costs and maximizing revenues due to unresolved agency conflicts since the available mechanisms for controlling owner-manager conflicts are relatively weak in mutuals (see Mester 1989). These hypotheses are not mutually exclusive, and mutuals could be more successful in low managerial discretion lines even though mutual managers exhibit expense preference behavior (Cummins et al. 1999b). Biener and Eling (2012) employed DEA cross-frontier analysis (a method that had previously been utilized to test these hypotheses in other studies, such as Cummins et al. (1999b); Cummins et al. (2004), or Erhemjamts and Leverty (2010)) to test the *efficient structure* and the *expense preference hypotheses* on a sample of insurers from 21 countries over the period 2002–2006. They estimated technical and cost efficiency using DEA and provided evidence for the *efficiency structure hypothesis* in selected market segments and no indication of the *expense preference hypothesis*. The results provided insight into the competitiveness of stock and mutual insurers from different countries.

I also include two papers in this section where the takaful organizational form was analyzed. Alshammari et al. (2019), using a stochastic cost frontier function, compared the efficiency of takaful firms and conventional insurers of Gulf Cooperation Council (GCC) countries for the period 2009–2016. They found that takaful operators were less efficient than conventional insurers. Al Amri et al. (2021) tested the differences in performance between the two major takaful organizational forms—the *mudharaba* (profit-sharing) and the *wakala* (fee-based) models—for compensating managers in a sample of takaful firms belonging to 19 different countries. The profit-sharing model was hypothesized to be more effective in aligning the incentives of managers and policyholders. The results showed that profit-sharing takaful firms were significantly more efficient (cost, technical, and scale) than fee-based takaful firms and, therefore, that agency costs were associated with fee-based (wakala) takafuls.

Furthermore, three articles—Abdul Kader et al. (2014), Karbhari et al. (2018), and Lee et al. (2019)—investigate corporate governance issues in the market with inconclusive results. Abdul Kader et al. (2014) investigated the influence of board composition on cost efficiency in nonlife takaful firms operating in 17 Islamic countries over the period 2004–2007. They found that the relative influence of board composition, such as the proportion of nonexecutive directors on the board, on cost efficiency depended on its interactions with other firm-specific characteristics, such as board size. Karbhari et al. (2018) using alternative efficiency measurement techniques (DEA, FDH, Order M DEA, and DEA bootstrap) on a sample of takaful undertakings from 21 countries over the period 2002 to 2013 found that CEO/chair duality and board size had a positive link with technical efficiency, while nonexecutive directors, Shari'ah board, and institutional ownership improved scale efficiency. Lee et al. (2019) evaluated the technical and cost efficiency of the Malaysian takaful industry from 2011 to 2015 using DEA. They examined corporate governance factors that influence efficiency and found a negative impact of board size and the proportion of nonexecutive directors on cost efficiency.

4.5 Risk Management

Three articles apply frontier efficiency analysis to evaluate issues related to risk management. Grace et al. (2015) investigated which aspects of enterprise risk management (ERM) improved efficiency

in the P/L and life insurance industries in the USA in 2004 and 2006. They used the DEA approach to estimate cost and revenue efficiency. They found that ERM practices resulted in economically and statistically significant increases in cost and revenue efficiency. They particularly showed that a single economic capital model increased revenue efficiency, dedicated risk managers improved cost efficiency, and that requiring the risk manager to report to the board or CEO increased revenue efficiency. These findings give information to managers about where they can allocate capital to the appropriate methods for their firms. Wu et al. (2016) evaluated the effects of risk-management committee characteristics and prestige on the efficiency of Malaysian insurance companies from 2008 to 2013 using a dynamic network DEA approach to estimate efficiency. This article represents a relevant application of the dynamic network DEA approach in the insurance industry due to the meticulous selection of inputs, intermediates, carry-overs, and outputs. They showed a positive relationship between risk management committee characteristics and efficiency. These characteristics refer to the number of directors on the RM committee, the proportion of independent directors, and the number of titles of honor carried by directors. They also showed that the prestige of the risk management committee could minimize the positive effect of these characteristics on efficiency. Altuntas et al. (2021) analyzed whether the use of the enterprise risk management (ERM) approach helped firms achieve economies of scale and scope in the German property-liability insurance industry over the period 1999–2009. They estimated scale and scope economies using DEA and used detailed survey data of German property-liability insurers to construct continuous measures of ERM quality. They found that ERM quality positively moderated both the size-scale efficiency relationship and the diversification-revenue scope efficiency relationship. These findings indicate that ERM facilitates economies of scale and economies of scope with respect to revenue complementarities, suggesting that ERM creates value through its impact on economies of scale and scope.

4.6 Other Management Issues

In this section, I include papers that addressed other management issues not included in other application areas, such as internationalization diversification or the role of intellectual capital. Biener et al. (2016), in their analysis of the Swiss insurance industry from 1997 to 2013, tested whether internationalization diversification impacted the efficiency of the Swiss insurance industry from 1997 to 2013. They showed that the internationalization of the Swiss insurance business had a positive impact on life insurer efficiency (technical, cost, and revenue) but not on p/c. They also found that entering the non-European market for Swiss life and p/c insurers was not as beneficial as entering European markets.

Two studies utilize frontier efficiency analysis to evaluate the role of intellectual capital in the performance of the insurance industry. Lu et al. (2014) applied a DEA dynamic slack-based measurement model to evaluate the relationship between the intellectual capital and efficiency of Chinese life insurance companies for the period 2006–2010. In a truncated regression model, they showed that intellectual capital positively affected operating efficiency. Nourani et al. (2018) used a dynamic network DEA model to estimate intellectual capital (IC) efficiency at three levels in the Malaysian insurance industry over the period 2005–2012. They found a lack of IC efficiency, which was mainly due to poor performance in terms of human capital and structural capital, indicating that both should be strengthened.

4.7 Distribution Systems

Two papers analyzed issues related to the link of distribution systems on insurer performance using frontier efficiency analysis. Ma et al. (2014) utilized the DEA approach to estimate cost and revenue efficiency in the P-L insurance industry of the USA over the period 1993–2008 and evaluated the impact of abandoning the use of contingent commissions on insurer performance. They showed that the abandonment of the use of contingent commissions before (after) the Spitzer investigation was (was not) rewarded with higher performance. The results corresponding to the period after the Spitzer investigation were attributed to the cost of reputation preservation.

Rubio-Misas (2022b) estimated the cost efficiency and profit inefficiency of the Spanish insurance industry over an 8-year sample period using DEA. She investigated why bancassurance coexists with alternative distribution systems in the long run, considering that the bank channel is known to involve lower costs than traditional distribution systems. She tested the *product-quality hypothesis*, which maintains that the higher costs of some distribution systems represent expenses associated with producing higher product quality, greater service intensity, and/or skills to solve principal-agent conflicts. The results supported these arguments since the use of the banking channel was found to increase cost efficiency, but the distribution channel(s) utilized did not affect profit inefficiency. These findings indicate that the establishment of a regulatory environment that allows the sale of insurance products through banking networks would enable the banking channel and alternative distribution systems to be sorted into market niches, where each channel has comparative advantages in production and/or agency costs.

4.8 Market Structure

I include the papers by Biener et al. (2017) and Alshammari et al. (2019) in this section. Biener et al. (2017), in their analysis of the global reinsurance market, tested the *efficient structure* hypothesis, which predicts that efficient firms are more competitive and gain market share due to their ability to charge lower prices without sacrificing profitability. Their results provided support to this hypothesis since cost-efficient reinsurers can charge lower prices while preserving profitability. Alshammari et al. (2019) examined the impact of market concentration variables (the Herfindahl-Hirschman index and insurers' market share) on the cost efficiency of conventional insurers and takaful undertakings in GCC countries between 2009 and 2016 using a stochastic frontier cost function. They found that conventional insurers and takaful undertakings operating in more concentrated markets tended to be less efficient. They explain this finding in support of the *Quiet Life hypothesis*, which assumes that higher market power may lead managers to reduce their efforts, which reflects in decreasing operating efficiency.

4.9 Regulatory Change

Five articles investigated the impact of regulatory change on insurers' performance using frontier efficiency analysis. Alhassan and Biekpe (2015) tested whether regulatory reforms toward the adoption of solvency assessment management and the adoption of the Solvency II improved efficiency. Their results from the convergence analysis showed a slower catch-up effect, implying that less efficient insurers were able to catch up with efficient insurers. Reyna and Fuentes (2018) analyzed whether the NAFTA deregulation of 2000, which allowed the setting up of companies in

Mexico with 100% foreign capital, improved the performance of Mexican insurance firms. In doing so, they conducted a DEA cost efficiency analysis of the insurance industry in Mexico over the period 2001–2013. Cost frontier estimation was grounded in a joint production technology of desirable and undesirable outputs modeled in an input-oriented fashion to analyze the sources that explain the cost variations of insurance companies in the analyzed period in an intertemporal manner. They found that improvements in productivity did not occur and observed an increase in costs arising from changes in the prices and quantities of inputs, concluding that changes in regulation were not an incentive to reduce costs by improving productivity.

Ohene-Asare et al. (2019) evaluated the impact of the introduction of the Insurance Act of 2006 (Act 724), which required the separation of life and nonlife businesses, on the performance of the Ghanaian insurance industry. They analyzed performance by estimating a cost Malmquist productivity index (see Maniadakis and Thanassoulis 2014) of insurers in Ghana over the period 2005–2014. In general, an increase in the productivity capabilities of the industry appeared. Nevertheless, although they found cost productivity growth between 2007 and 2009 and from 2010 to 2011, since 2012, the industry has faced a cost productivity decline. Lim et al. (2021) analyzed the efficiency and productivity of insurance operators in Malaysia before and after risk-based capital (RBC) and risk-based capital for takaful (RBCT) implementation. They estimated the efficiency (technical, pure technical, and scale) and productivity of the Malaysian insurance industry from 2000 to 2017 using DEA. The results showed a different effect of the implementation of risk-based capital on the performance of conventional insurers compared with takaful firms. While conventional insurers were more efficient during the pre-RBC period, takaful firms showed increased efficiency following RBCT implementation. Furthermore, while conventional insurers showed a decrease in productivity in the post-RBC implementation period, takaful undertakings did not show changes in productivity after the introduction of RBCT.

Cummins and Rubio-Misas (2022) analyzed integration and convergence in the efficiency and technology gap of 10 EU life insurance markets over the period 1998–2014. They evaluated whether the regulatory steps taken to promote integration among EU insurance markets led to a convergence in efficiency and technology gaps in those markets. They applied the meta-frontier DEA approach to estimate both the efficiency and technology gaps and the main concepts of convergence from the economic growth literature, β-convergence and σ-convergence, in the analysis. The results showed convergence in cost/revenue efficiency among major EU life insurance markets during the sample period. They also found convergence in the cost/revenue technology gap among these markets, indicating that they became more technologically homogeneous during the sample period. However, the results indicated that the global financial crisis led to a slowdown in the progress of integration and convergence in the efficiency and technology gap of EU life insurance markets in terms of cost efficiency but not in terms of revenue efficiency, suggesting that the crisis harmed competition in input prices but not competition in output prices.

4.10 Environmental Conditions

I have selected five papers in which the role that environmental conditions play in the performance of insurance firms using frontier efficiency analysis is a main issue. Nevertheless, environmental conditions at the country level used to be taken into account in many frontier efficiency studies, particularly in multicountry studies. Huang and Eling (2013) analyzed the technical efficiency of nonlife insurers in BRIC (Brazil, Russia, India, China) countries over the period 2000–2008 by incorporating uncontrollable variables in the efficiency analysis using multistage DEA. This approach

captures cross-country differences, such as the political and economic environment, distinguishing between managerial inefficiency and inefficiency due to environmental conditions. They found that the growth of GDP, consumer price index, and deposit interest rate were adverse environmental factors with respect to efficiency. Eling and Schaper (2017) applied a multistage DEA approach to analyze the impact of the changing environment on the efficiency and productivity of European life insurers from 14 countries over the period 2002–2013. They found that general economic, capital market, and insurance market conditions were important drivers of efficiency. These findings revealed that the differences in business conditions across countries harmed the efficiency of life insurers that were exposed to relatively unfavorable conditions, leading to competitive disadvantages for those companies.

Cummins and Rubio-Misas (2021) evaluated the role that national financial market development and institutional quality played in the integration of EU life insurance markets. They analyzed 10 EU life insurance markets over a 17-year sample period and used the meta-technology cost/revenue efficiency ratios, estimated under the meta-frontier DEA framework, as measures of integration. The results showed that national stock market development and institutional quality enhanced cost performance and the integration of EU life insurance markets. They also found that in countries where *bancassurance* was the main life insurance distribution channel, banking sector development contributed to integration in terms of revenue efficiency. However, the results also showed that better outcomes in national institutional quality decreased the metatechnology revenue efficiency ratio, suggesting that life insurance prices were lower in countries with better institutions. Some implications of this research are that regulators and policymakers should be concerned with designing programs involving the development of the stock market as well as the convergence of financial services via *bancassurance* in order to improve performance and achieve a more integrated EU life insurance market.

Reyna et al. (2022) examined the Mexican insurance industry over a 22-year period to evaluate whether insurers have improved their economic performance through efficiency as a reaction to the prolonged period of low interest rates. They estimated technical efficiency using DEA and found an inverse relationship between efficiency and interest rates in support to the view of the positive effect of prolonged low interest rate periods on insurer performance. Shaddady (2022) investigated the influence of the business environment, political risk, governance, and Shari'ah compliance on efficiency in the insurance market from 2007 to 2017. He evaluated the efficiency of listed insurers from MENA countries using a multistage DEA-SFA model. The results indicated that while interest rates, inflation, oil prices, and financial freedom impacted efficiency negatively, unemployment, economic growth, effective exchange rates, political risks, and the quality of regulation impacted it positively.

4.11 Total Factor Productivity

Fourteen papers provide an analysis of total factor productivity (TFP). As stated before, productivity is often analyzed in frontier efficiency studies to give a dynamic analysis of performance. Nine out of fourteen papers have estimated TFP change using the DEA Malmquist approach, two the SFA, one (Chen et al. 2014) the DEA Malmquist in two-stage production systems, one (Reyna and Fuentes 2018) the DEA intertemporal decomposition of the input cost variation of Grifell-Tatjé and Lovell (2015) and one (Ohene-Asare et al. 2019) the cost-Malmquist productivity index of Maniadakis and Thanassoulis (2014). Single countries in which productivity in the insurance industry is analyzed are Argentina, Ghana, India, Malaysia (twice), México, South Africa, Switzerland, Taiwan, Thailand, and

the USA (twice). There are also two multicountry studies. Most studies have focused on estimating the evolution of TFP change and its components over the analyzed period.

4.12 Methodological Issues

Several papers have made a methodological contribution although some of them use data on the insurance industry to simply verify the proposed model(s). Huang and Eling (2013) used the multi-stage DEA approach to incorporate both controllable and uncontrollable variables in the efficiency analysis of the insurance industry in BRIC countries. The use of uncontrollable variables allowed capturing cross-country differences and distinguishing managerial inefficiency and inefficiency due to environmental conditions. Kao and Hwang (2014) developed a multiperiod, two-stage DEA model to measure the overall and period efficiencies at the same time, with the first expressed as the weighted average of the latter. The efficiencies measured from the model could be used to calculate a common weight global Malmquist productivity index (MPI) between two periods. Lu et al. (2014) applied the dynamic slack-based measure DEA model of Tone and Tsutsui (2010) to investigate the operating efficiency of life insurers using a dynamic production process. This model incorporates carry-over inputs (in this case, debt capital and equity capital) to measure period-specific efficiency based on long-time optimization. They verified that the efficiency scores of the DEA dynamic slack-based model differed significantly from those of the traditional DEA model.

Lim and Zhu (2016) demonstrated that the duality in standard DEA naturally migrates to two-stage network DEA. They developed formulas to obtain frontier projections and divisional efficiency scores using a set of dual variables. The case of Taiwanese nonlife insurance companies (see Kao and Hwang 2008) was revisited using this newly developed approach. Wanke and Barros (2016) used different data mining techniques in conjunction with the DEA meta-frontier in a two-stage approach to predict the performance of the Brazilian insurance industry. The inputs and outputs considered were restricted by the availability of the data to different measures of assets and premiums, respectively. They found a heterogeneous impact on the efficiency levels of the geographical regions served and the type of insurance offered by insurance companies. Nourani et al. (2018) applied a network DEA slack-based model to address the connectivity between the stages as well as incorporating the time effect.

Tone et al. (2019) modeled investment assets as a carry-over variable in a dynamic two-stage network DEA model, which was applied to the Malaysian insurance industry from 2008 to 2016. The results indicated that this model led to improved discriminatory power of overall performance, particularly investment efficiency, indicating that if insurers focused on the dynamics of their investment assets, they would be in a better position to build business competitiveness over long-term periods. Lim and Zhu (2019) tested whether and how the primal-dual correspondence between the multiplier model and the envelopment model could be obtained for two-stage network DEA models. They developed an axiomatic derivation of some two-stage network DEA models focusing on the basic two-stage serial process structure. The Taiwanese nonlife insurance companies studied by Kao and Hwang (2008) were used to apply the models. Cummins and Rubio-Misas (2022) tested the suitability of using the DEA meta-frontier framework for a sample of 10 European countries. The results confirmed the presence of heterogeneity in production possibility sets among countries and, thus, that the use of the DEA meta-frontier framework was appropriate. Nourani et al. (2022) used different DEA frameworks (a classical slack-based model (SBM), a network SBM, a dynamic SBM, and a dynamic network SBM) to evaluate efficiency in the Chinese insurance industry over the period 2014–2018. They found that the network models generally performed better in discriminating inefficient DMUs.

4.13 Trends and Gaps

Extended geographic coverage, particularly in terms of multicountry studies, was an important contribution to research on efficiency and productivity in the insurance industry conducted in the last decade. However, most of this research has focused on the production and cost frontier, with works that estimate the revenue or profit frontier scarce. This lack is sometimes justified by the unavailability of data to calculate output prices. Nevertheless, a comprehensive analysis of insurers' performance should complement cost efficiency analysis with revenue or profit efficiency analysis. Furthermore, recent works in this field have used refined methodologies, particularly those related to DEA, with an important number of articles applying network DEA and dynamic DEA models with carry-over variables. However, papers that use SFA are scarce. Applying recent developments in SFA (see, e.g., Kumbhakar et al. 2020) to evaluate particular questions of the insurance industry could be an avenue for future research.

There are two topics included in this review that were not considered in previous surveys as separate areas: solvency and environmental conditions. While the first basically includes papers investigating if firm efficiency is related to solvency, the second involves articles analyzing how environmental conditions influence efficiency. More research is needed on these two topics, particularly on whether insurer efficiency during normal times affects risk during subsequent financial crises or on how environmental conditions not yet studied, such as some cultural factors, may influence efficiency. Furthermore, unlike previous surveys, papers published on the topic of mergers and acquisitions have not been found. Therefore, there is room for investigating the efficiency and productivity effects of M&As as well as ex ante predictions of potential acquirers and targets according to firm characteristics in terms of efficiency. Additionally, other topics under research deserve additional attention. For instance, more research is needed in the distribution system area to test the *product quality* hypothesis vs. the *market imperfection* hypothesis. The applications of frontier efficiency to the market-power vs. efficient-structure debate about the determinants of profitability in multicountry analyses also offer an area for future research.

5 Summary and Conclusions

This chapter provides an in-depth study of research on efficiency and productivity in the insurance industry utilizing frontier methodologies that has been published in journals belonging to the Clarivate Journal of Citation Reports from 2012 to 2022. These methodologies constitute benchmarking ways for estimating efficiency and productivity and have become mainstream in measuring the performance of firms and other types of decision-making units using accounting data. They allow the estimation of technical, cost, revenue, and profit frontiers; the two main approaches for estimating them are econometric methods and nonparametric methods. There is no consensus on the best method for deciding the best-practice frontier against which the relative efficiencies of firms are calculated. Nevertheless, the DEA approach—the most outstanding nonparametric method—has been the most frequently used method in the surveyed papers, with 37 studies. This approach provides a particularly convenient way for decomposing cost or revenue efficiency into pure technical, scale, and allocative efficiency. The production frontier is the most estimated frontier (with thirty-three studies including eight papers that use a two-stage network process), followed by the cost frontier (with 24 papers). However, one can observe the scarce number of works that estimate revenue and profit frontiers, with nine and three studies, respectively. This lack was sometimes justified because data to calculate

output prices were not available. Nevertheless, a comprehensive analysis of insurers' performance should include cost efficiency and also revenue or profit efficiency since, according to traditional microeconomic theory, firms are profit maximizers by minimizing costs and maximizing revenues.

The information collected from utilizing frontier methods to evaluate the performance of insurers can be utilized in a variety of ways, the main ones being enhancing managerial performance, basically by informing management about the effects of new strategies and technologies; providing a guide to regulators and policymakers with regard to the appropriate response to problems; and testing economic hypotheses. The reviewed articles represent contributions to twelve application areas. There are two topics not considered in previous surveys as separate areas: the solvency topic, which basically includes papers investigating if firm performance is likely to be related to solvency, and the environmental conditions topic, with articles analyzing the role that environmental variables play in the performance of insurers. However, unlike previous surveys, papers published in the area of mergers and acquisitions have not been found, a topic in which there is room for investigating not only the efficiency and productivity effects of M&As but also ex ante predictions of potential acquirers and targets on the basis of the set of firm characteristics (particularly in terms of frontier efficiency characteristics). Furthermore, there are other underresearched insurance topics, providing avenues for future research. This is the case in the distribution system area, where more research is needed to test the *product quality* hypothesis vs. the *market imperfection* hypothesis or the market structure area in which hypotheses derived from the industrial organization literature may be investigated. Nevertheless, a fruitful topic of the surveyed papers refers to methodological issues. These contributions imply important refinements of the methods. Yet, some papers lack a detailed analysis of insurers' value creation process, which would improve the utility of these advancements for management.

Acknowledgments The author gratefully acknowledges financial support from the Spanish Ministry of Science and Innovation (Project PID2021-127736NB-I00).

Appendix

Appendix 1 Outputs, Output Prices, Inputs, and Input Prices

Author	Output volume	Output price	Input type	Input volume	Input price
Leverty and Grace (2012)	Present value of real losses incurred	(Real premiums earned—output volume)/output volume	Adminstrative labor	Current dollar expenditures associated with administrative labor/administrative labor price	Average weekly wage rate, in the state in which the insurer's home office is located, for the property-liability insurer for SCI 6331
	Average real invested assets	Expected rate of return on the insurer's assets	Agent labor	Current dollar expenditures associated with agent labor/agent labor price	Average weekly wage rate for insurance agents for SCI 6411
			Material and business services	Current dollar expenditures associated with material and business services/material and business services price	Average weekly wage rate for business services for SCI 7300
			Financial equity capital	The average of the beginning and end-of-year real equity capital	Average 90-day Treasury bill rate in year t, plus the long-term average market risk premium on large company stocks
			Debt capital	Real loss reserves plus real unearned premium reserves	Total investment income minus expected investment income attributed to equity capital divided by average policyholder-supplied debt capital
Biener and Eling (2012)	Net incurred claims (benefits) plus addition to reserves		Labor and business service	Operating expenses/input price	ILO October Inquiry wage per year
	Total investments		Debt capital	Total liabilities	Long-term government bond rates
			Equity capital	Capital and surplus	20-year-average MSCI stock market return indices
Cummins and Xie (2013)	Real present value of losses incurred (smoothed)	(Premiums earned—output volume)/output volume	Administrative labor	Total expenditures on administrative labor/price of labor	National average weekly wage rate for SIC 6331

	Average invested assets	Weighted average of expected investment return on stocks plus the expected return on interest-bearing assets	Agent labor	Total expenditures on agent labor/price of labor	National average weekly wage rate for SIC 6411
			Material and business services	Total material and business services expenses/input price	Weighted avg. of prices indices for business services for expenses components from Best's Aggregates & Averages
			Financial equity capital	Real average value of equity capital	Fama-French three-factor cost of capital estimated for traded insurers, assigned to nontraded based on Best's ratings
Huang and Eling (2013)	Net premium written		Labor and business service	Management expenses/price of labor	
	Total Invested Assets		Equity capital	Paid-in capital plus capital surplus	
			Debt capital	Net technical reserves	
Gaganis et al. (2013)	Insurance premiums		Management and commissions (variable input)		Commissions and management expenses/total assets
	Value of invested assets		Claims or benefits (variable input)		Net claims (benefits)/total assets
			Equity (quasi-fixed input)	Equity capital	
			Technical reserves (quasi-fixed input)	Net technical reserves	
Abdul Kader et al. (2014)	Gross premiums contributions		Labor	Number of employees	The estimated total wage bill divided by the number of employees
			Physical capital	Insurer's total assets	Total capital expenses/total assets

(continued)

Appendix 1 (continued)

Author	Output volume	Output price	Input type	Input volume	Input price
Kao and Hwang (2014)	Underwriting profit		Operating expenses	Salaries and various types of costs incurred in daily operations	
	Investment profit		Insurance expenses	Expenses paid to agencies, brokers, and solicitors and other expenses associated with marketing the service of insurance	
	Direct written premiums (intermediate)		Direct written premiums (intermediate)	Premiums received from insured clients	
	Reinsurance premiums (intermediate)		Reinsurance premiums (intermediate)	Premium received from ceding companies	
Lu et al. (2014)	Incurred benefits plus addition to reserves		Labor and business service	Various types of expenses incurred in daily operations	
	Investment profits		Debt capital (carry-over)	Total amount of liabilities from the balance sheet	
			Equity capital (carry-over)	Total amount of owner's equity from the balance sheet	
Chen et al. (2014)	Incurred claims plus addition to reserves		Operating expenses	Operating expenses	
	Investment assets		Total liabilities	Total liabilities	
			Equity	Equity	
Ma et al. (2014)	Present value of losses incurred	(Premiums earned—output volume)/output volume	Administrative labor	Administrative labor expenses/input price	Average weekly wage for employees in insurance companies
	Real invested assets	Weighted average of the expected return on stocks and the realized return on interest-bearing assets	Agent labor	Agent labor expenses/input price	Average weekly wage for insurance agents

			Material and business services	Material and business services expenses/input price	Weighted average of price indices for business services from the expense page of Best's Aggregates and Averages
			Equity capital	Equity capital	Fama-French three-factor cost of capital estimated for traded insurers, assigned to nontraded based on Best's ratings
Yaisawarng et al. (2014)	Number of policies issued		Labor		Total salaries and welfare/total number of administrative and staff employees
	Total invested assets		Business services		(Net commissions and brokerages plus net loss adjustment expenses plus building and equipment expenses plus other operating expenses plus other underwriting expenses)/thousand baht of sum insured
			Capital		Profits from operations/capital input
Grace et al. (2015)	Present value of losses incurred (benefits) plus addition to reserves	(Net premiums earned—output volume)/output volume	Administrative labor	Total expenditure on administrative labor/administrative labor price	Average weekly wage for SCI6331
	Real Invested assets	Expected rate of return on the insurer's assets	Agent labor	Total expenditure on agent labor/agent labor price	Average weekly wage rate for insurance agents for SCI 6411
			Business services and materials	Total expenditure on business services and materials/business services and material price	Average weekly wage rate for business services for SCI 7300

(continued)

Appendix 1 (continued)

Author	Output volume	Output price	Input type	Input volume	Input price
			Financial equity capital	Financial equity capital	Based on the insurer's A.M. Best Company financial rating, the cost of capital is 12% for firms rated in the A range, 15% for firms in the B range, and 18% for firms below B
			Policyholder-supplied debt capital	Policyholder-supplied debt capital	Average corporate credit spread of similarly rated firms
Alhassan and Biekpe (2015)	Premiums earned or incurred losses		Management expenses	Management expenses	
	Investment income		Total equity	Book value of equity	
			Total liability	Borrowed funds	
Dreassi and Schneider (2015)	Gross written premiums		Distribution Channels		Distribution costs/gross premiums
			Human resources, marketing, and IT activities		Administrative expenses/gross written premiums
Biener et al. (2016)	Real benefits (losses) incurred	(Net premiums earned—output volume)/output volume	Labor and business services	Operating expenses/price of labor	Annual wage variable for the Swiss insurance sector
	Real value of total investment	Expected rate of return on the insurer's assets	Debt capital	Debt capital	10-year average returns of Swiss confederations bonds
			Equity Capital	Equity capital	7-year averages of yearly rates of return of the Swiss Market Index
Lim and Zhu (2016)	Underwriting profit		Operational expenses	Operational expenses	
	Investment profit		Insurance expenses	Insurance expenses	
	Direct writing premiums (intermediate)		Direct writing premiums (intermediate)	Direct writing premiums (intermediate)	
	Reinsurance premiums (intermediate)		Reinsurance premiums (intermediate)	Reinsurance premiums (intermediate)	

Wu et al. (2016)	Operating income		Labor and business service expenses	Operating expenses	
	Operating cash flow		Debt capital (carry-over)	Total amount of liabilities (carry-over)	
	Incurred claims plus additions to reserves (intermediate)		Equity capital (carry-over)	Total amount of shareholders' equities (carry-over)	
	Investment Assets (intermediate)		Property, plant, and equipment (carry-over)	Property, plant, and equipment (carry-over)	
			Incurred claims plus additions to reserves (intermediate)	Incurred claims plus additions to reserves (intermediate)	
			Investment assets (intermediate)	Investment assets (intermediate)	
Wanke and Barros (2016)	Direct Premiums		Current assets	Current assets	
	Insurance premiums		Real assets	Real assets	
	Retained Premiums		Long-term fixed assets	Long-term fixed assets	
	Earned Premiums		Other long-term assets	Other long-term assets	
Rubio-Misas and Fernández-Moreno (2017)	Value of real losses incurred		Business expenses	Business expenses/input price	Average monthly wage for employees in the Spanish insurance sector
			Debt capital	Debt capital	1-year Spanish Treasury bill rate
			Equity capital	Equity capital	Average rate of total return on the Madrid Stock Exchange Index for the 10-year period preceding each of the sample period
Bikker (2017)	Gross Premiums written or claims incurred		Labor		The logarithm of the hourly cost of labor

(continued)

Appendix 1 (continued)

Author	Output volume	Output price	Input type	Input volume	Input price
	Total assets				
Eling and Schaper (2017)	Net benefits plus changes in reserves		Labor	Total operating expenses/country-specific prices of labor	Country-specific prices of labor obtained from the International Labor Organization
	Total invested assets		Debt capital	Debt capital	Long-term interest rates obtained from the OECD
			Equity capital	Equity capital	15-year rolling returns on MSCI country-specific stock market indices
Biener et al. (2017)	Losses	(Premium-output volume)/output volume	Labor	Number of employees	Annual average wages for the insurance sector
	Total invested assets	Annual realized industry-wide investment returns	Material and business services	Operating expenses minus employee costs	Consumer price indices of respective country-years
			Equity capital	Capital and surplus	Overall ROE of the reinsurance industry each year
Karbhari et al. (2018)	Total amount contribution received		Administrative expenses	Administrative expenses incurred	
			Labor	Number of labor utilized	
			Labor expenses	Total amount of salaries paid to staff	
			Capital assets	Pieces of properties owned and investments held	
Danquah et al. (2017)	Net incurred benefits (claims)		Operating expenses		Operating expenses/total assets
	Investment income		Equity capital		Equity capital/total assets
			Debt capital		Debt capital/total assets

Ferro and León (2018)	Total losses		Direct labor costs	Exploitation expenses—marketing expenses	
			Commissions to brokers	Total production expenses	
			Financial capital	Equity capital plus liabilities	
Eling and Jia (2018)	Insurance benefits (losses)		Labor	Operating expenses/annual wage for the insurance sector in respective market-years	
	Total invested assets		Equity capital	Equity capital	
			Debt capital	Debt capital	
Nourani et al. (2018)	Investment income		Management expenses	Management expenses	
	Net profit		Staff expenses	Staff expenses	
	Premiums earned (intermediate)		Property, plant, and equipment (carry-over)	Property, plant, and equipment (carry-over)	
	Operating cash flow (intermediate)		Equity (carry-over)	Equity (carry-over)	
	Total Investment (intermediate)		Premiums earned (intermediate)	Premiums earned (intermediate)	
	Net claims plus addition to reserves (intermediate)		Operating cash flow (intermediate)	Operating cash flow (intermediate)	
			Total investment (intermediate)	Total investment (intermediate)	
			Net claims plus addition to reserves (intermediate)	Net claims plus addition to reserves (intermediate)	
Reyna and Fuentes (2018)	Insurance policies issued and paid (desirable output)	Average insurance premium per line of product	Labor costs	Annual labor cost/labor price	Index of national minimum wage
	Reported claims (undesirable output)	Average value of each event per line of business net after adjustment expenses	Other administrative expenses	Value of fixed assets at the end of the year	Share of other operating expenses with respect to the value of fixed assets

(continued)

Appendix 1 (continued)

Author	Output volume	Output price	Input type	Input volume	Input price
			Acquisition expenses	Annual direct premiums	Share of annual acquisition expenses with respect to the direct premiums of the year
			Adjustment expenses	Value of direct claims	Share of the amount of adjustment expenses with respect to the settlement of claims or direct claims during year
Alshammari et al. (2019)	Net claims incurred		Labor		Salary of all staff
	Invested assets		Debt capital		Product of rate of return on debt and technical provision
			Equity capital		Product of average stock market return indices and total capital
Eling and Jia (2019)	Profit before tax		Labor	Operating expenses/annual wage for the insurance sector in respective market-years	
	Net premiums written (intermediate)		Equity capital	Capital and surplus	
	Total invested assets (intermediate)		Debt capital	Total liabilities	
			Net premiums written (intermediate)	Net premiums written (intermediate)	
			Total invested assets (intermediate)	Total invested assets (intermediate)	
Shavitha et al. (2019)	Incurred benefits		Number of branches	Number of branches	
	Number of policies sold		Commissions charged	Commissions charged	
Tone et al. (2019)	Premiums earned		Management expenses	Management expenses	
	Investment income		Fixed assets (carry-over)	Fixed assets (carry-over)	
	Incurred claims plus addition to reserves (intermediate)		Investment assets (carry-over)	Investment assets (carry-over)	

			Incurred claims plus addition to reserves (intermediate)	Incurred claims plus addition to reserves (intermediate)	
Lee et al. (2019)	Net contribution income		Labor	Management expenses plus operating expenses	Average annual wage in Malaysia
	Net investment income		Debt capital	Amount invested in Islamic bonds issued by the Malaysian government	Annual return of government investment issues
	Provision outstanding claims		Equity capital	Amount invested in Islamic-related securities in the market	Annual return of the Shari'ah EMAS Index
Lim and Zhu (2019)	Underwriting profit		Operational expenses	Operational expenses	
	Investment profit		Insurance expenses	Insurance expenses	
	Direct written premiums (intermediate)		Direct written premiums (intermediate)	Direct written premiums (intermediate)	
	Reinsurance premiums (intermediate)		Reinsurance premiums (intermediate)	Reinsurance premiums (intermediate)	
Ohene-Asare et al. (2019)	Net premiums		Operating expenses	Operating expenses	Operating expenses deflated by the GDP deflator as a proportion of the total assets of an insurer
	Investment income		Equity capital	Equity capital	Shareholder equity ratio
			Debt capital	Debt capital	Debt to total assets ratio
Al Amri et al. (2021)	Premiums		Labor		Operating expenses/number of employees
	Total investment		Equity capital		Expected rate of return on the stock market of each country by year
			Technical provisions		Each country's annual rate of interest on lending
Peng and Lian (2021)	Benefits		Operating expenses	Operating expenses	

(continued)

Appendix 1 (continued)

Author	Output volume	Output price	Input type	Input volume	Input price
	Investment income		Equity capital	Equity capital	
			Debt capital	Debt capital	
Altuntas et al. (2021)	Present value of losses incurred net of reinsurance	Real premiums earned—output volume/output volume	Labor and business services	Operating expenses/input price	German Census Bureau's index for insurance business services (BI index)
	Total invested assets	Realized investment income return for the year	Equity capital	Book value of surplus	Insurer's expected return on equity
			Debt capital	Technical provisions net of reinsurance	Two-to-three-year German Treasury Bill rates
Lim et al. (2021)	Total investment		Operating expenses	Net commissions paid and management expenses	
	Net claims (benefits) paid		Equity capital	Capital and surplus	
	Net premiums (contributions) for conventional (takaful)		Debt capital	Total liabilities	
Cummins and Rubio-Misas (2021)	Value of real incurred benefits plus addition to reserves	(Premiums—output volume)/output volume	Operating expenses	Operating expenses/input price	Eurostat Index based on wages and salaries
	Real value of invested assets	Net investment income to invested assets	Debt capital	Debt capital	OECD 10-year-Threasury-Bill rates of the specific country
			Equity capital	Policyholders' surplus	20-year rolling average of the yearly rates of return of the country-specific MSCI stock market indices
Ho and Hsu (2021)	Premium income		Labor	Number of employees	
	Service charge		Business services	Indemnity payment	
	Net investment gains		Equity capital	Capital share and additional paid-in capital	
Reyna et al. (2022)	Issued and paid policies (desired output)		Administrative expenses	Annual administrative expenses divided by the national index of prices for the consumer of that year	

	Reported claims (undesired output)		Acquisition costs	Annual acquisition expenses divided by the national index of prices for the consumer of that year	
Cummins and Rubio-Misas (2021)	Value of real incurred benefits plus addition to reserves	Premiums—output volume/output volume	Operating expenses	Operating expenses/input price	Eurostat Index based on wages and salaries
	Real value of invested assets	Net investment income to invested assets	Debt capital	Debt capital	OECD 10-year-Threasury-Bill rates of the specific country
			Equity capital	Policiholders' surplus	20-year rolling average of the yearly rates of return of the country-specific MSCI stock market indices
Shaddady (2022)	Net income		Underwriting expenses	Underwriting expenses	
	Net premium written		Equity capital	Equity capital	
			Net claims incurred	Net claims incurred	
Rubio-Misas (2022)	The value of real incurred losses (benefits)	(Premiums earned—output volume)/output volume	Equity capital	Equity capital	Average return of the total return of the Madrid Stock Exchange Index for the 10-year period preceding each year of the sample period
			Business expenses	Business expenses/input price	Average monthly wage paid to employees in the Spanish insurance sector
			Debt financial capital	Debt financial capital	One-year Spanish Treasury bill rate
Nourani et al. (2022)	Premiums earned (intermediate)		Operating expenses	Operating expenses	
	Investment income		Total debt (carry-over)	Total debt (carry-over)	
	Net profit		Total equity (carry-over)	Total equity (carry-over)	
			Total Investment (carry-over)	Total Investment (carry-over)	

References

Abdul Kader H, Adams M, Harwick P, Kwon WJ (2014) Cost efficiency and board composition under different takaful insurance business models. Int Rev Financ Anal 32:60–70

Aigner D, Lovell K, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econ 6:21–37

Al Amri K, Cummins JD, Weiss MA (2021) Economies of scope, organizational form, and insolvency risk: evidence from the takaful industry. J Int Financ Mark Inst Money 70:101259

Alhassan AL, Biekpe N (2015) Efficiency, productivity and returns to scale economies in the non-life insurance market in South Africa. Geneva Pap 40(3):493–515

Alshammari AA, Alhabshi SMSJ, Saiti B (2019) The impact of competition on cost efficiency of insurance and takaful sectors. Evidence from GCC markets based on the Stochastic Frontier Analysis. Res Int Bus Financ 47:410–427

Altuntas M, Berry-Stölzle TR, Cummins JD (2021) Enterprise risk management and economies of scale and scope: evidence from the German insurance industry. Ann Oper Res 299(1):811–845

Banker RD (1993) Maximum likelihood consistency and data envelopment analysis: a statistical foundation. Manag Sci 39:1265–1273

Banker RD, Natarajan R (2008) Evaluating contextual variables affecting productivity using data envelopment analysis. Oper Res 56(1):48–58

Banker RD, Natarajan R, Zhang D (2019) Two-stage estimation of the impact of contextual variables in stochastic frontier production model using data envelopment analysis: second stage OLS versus bootstrap approaches. Eur J Oper Res 278(2):368–384

Battese GM, Coelli TJ (1995) A model for technical inefficiency effects in stochastic frontier production function for panel data. Empir Econ 20:325–332

Battese GM, Corra GH (1977) Estimation of production frontier models with application to the pastoral zone of Eastern Australia. Aust J Agric Econ 21:167–179

Berger AN, Humphrey D (1992) Measurement and efficiency issues in commercial banking. In: Griliches Z (ed) Output measurement in the services sectors. University of Chicago Press, Chicago

Berger AN, Humphrey D (1997) Efficiency of financial institutions: international survey and direction for future research. Eur J Oper Res 98(2):175–212

Berger AN, Hanweck GA, Humphrey D (1987) Competitive viability in banking: scale, scope, and product mix economies. J Monet Econ 20:501–520

Biener C, Eling M (2012) Organization and efficiency in international insurance industry: a cross-frontier analysis. Eur J Oper Res 221:454–468

Biener C, Eling M, Wirfs JH (2016) The determinants of efficiency and productivity in the Swiss insurance industry. Eur J Oper Res 248:703–714

Biener C, Eling M, Jia R (2017) The structure of the global reinsurance market: an analysis of efficiency, scale and scope. J Bank Financ 77:213–229

Bikker JA (2017) Competition and scale economy effects of the Dutch 2006 health-care insurance reform. Geneva Pap 42:53–78

Charnes A, Cooper W, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Chen FC, Liu ZJ, Kweh WL (2014) Intellectual capital and productivity of Malaysian general insurers. Econ Model 36:413–420

Cooper WW, Seiford LM, Zhu J (2011) Handbook of data envelopment analysis. Springer, New York

Cummins JD, Rubio-Misas M (2006) Deregulation, consolidation and efficiency: evidence from the Spanish insurance industry. J Money Credit Bank 38(2):323–355

Cummins JD, Rubio-Misas M (2021) Country factor behavior for integration improvement of European life insurance markets. Econ Anal Policy 72:186–202

Cummins JD, Rubio-Misas M (2022) Integration and convergence in efficiency and technology Gap of European life insurance markets. Ann Oper Res 315:93–119

Cummins JD, Weiss MA (2000) Analyzing firm performance in the insurance industry using frontier efficiency and productivity methods. In: Dionne G (ed) Handbook of insurance. Kluwer Academic Publishers, Norwell, pp 767–829

Cummins JD, Weiss MA (2013) Analyzing firm performance in the insurance industry using frontier efficiency and productivity methods. In: Dionne G (ed) Handbook of insurance, 2nd edn. Springer, New York, pp 795–861

Cummins JD, Xie X (2013) Efficiency, productivity, and scale economies in the us property-liability insurance industry. J Prod Anal 39:141–164

Cummins JD, Zi H (1998) Measuring economic efficiency of the US life insurance industry: econometric and mathematical programming techniques. J Prod Anal 10:131–152

Cummins JD, Tennyson S, Weiss MA (1999a) Consolidation and efficiency in the US life insurance industry. J Bank Financ 23:325–357
Cummins JD, Weiss MA, Zi H (1999b) Organizational form and efficiency: an analysis of stock and mutual property-liability insurers. Manag Sci 45:1254–1269
Cummins JD, Rubio-Misas M, Zi H (2004) The effect of organizational structure on efficiency: evidence from the Spanish insurance industry. J Bank Financ 28(12):3113–3150
Cummins JD, Dionne G, Gagne R, Nouira A (2009) Efficiency of insurance firms with endogenous risk management and intermediation activities. J Prod Anal 32(2):145–159
Cummins JD, Weiss MA, Xie X, Zi H (2010) Economies of scope in financial services: a DEA efficiency analysis of the US insurance industry. J Bank Financ 34(7):525–1539
Danquah M, Otoo DM, Baah-Nuakoh A (2017) Cost efficiency of insurance firms in Ghana. Manag Dec Econ 39:213–225
Debreu G (1951) The coefficient of resource utilization. Econometrica 19:273–292
Dreassi A, Schneider MC (2015) Bancassurance and scale economies: evidence from Italy. Geneva Pap 40(1):89–107
Eling M, Jia R (2018) Business failure, efficiency, and volatility: evidence from the European insurance industry. Int Rev Financ Anal 59:58–76
Eling M, Jia R (2019) Efficiency and profitability in the global insurance industry. Pac Basin Financ J 5:101190
Eling M, Schaper P (2017) Under pressure: how the business environment affects productivity and efficiency of European life insurance companies. Eur J Oper Res 258:1082–1094
Erhemjamts O, Leverty JT (2010) The demise of mutual organizational form: an investigation of the life insurance industry. J Money Credit Bank 42(6):1011–1036
Fare R, Grosskopf S, Margaritis D (2008) Efficiency and productivity: malmquist and more. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productivity efficiency and productivity growth. Oxford University Press, New York
Farrel MJ (1957) The measurement of productive efficiency. J R Stat Soc A 120:253–281
Ferro G, León S (2018) A stochastic frontier analysis of efficiency in Argentina's non-life insurance market. Geneva Pap 43:158–174
Gaganis C, Hasan I, Pasiouras F (2013) Efficiency and stock returns: evidence from the insurance industry. J Prod Anal 40:429–442
Grace MF, Leverty JT, Phillips RD, Shimpi P (2015) The value of investing in enterprise risk management. J Risk Insur 82(2):289–316
Greene WH (2008) The econometric approach to efficiency analysis. In: Fried HO, Lovell K, Schmidt SS (eds) The measurement of productive efficiency and productivity growth. Oxford University Press, New York
Grifell-Tatjé E, Lovell CAK (2015) Productivity accounting: the economics of the business performance. Cambridge University Press, New York
Grosskopf S (1996) Statistical inference and nonparametric efficiency: a selective survey. J Prod Anal 7:161–176
Hancock D (1985) The financial firm: production with monetary and nonmonetary goods. J Polit Econ 93(5):859–880
Ho SJ, Hsu HH (2021) The effect of microinsurance on the insurance market: evidence from Taiwan. Geneva Pap 46:130–145
Huang W, Eling M (2013) An efficiency comparison of the non-life insurance industry in the BRIC countries. Eur J Oper Res 226:557–591
Kao C, Hwang SN (2008) Efficiency decomposition in two-stage data envelopment analysis: an application to the non-life insurance companies in Taiwan. Eur J Oper Res 185:418–429
Kao C, Hwang SN (2014) Multi-period efficiency and Malmquist productivity index in two-stage production systems. Eur J Oper Res 232:512–521
Karbhari Y, Muye I, Hassan AFS, Elnahass M (2018) Governance mechanisms and efficiency: evidence from an alternative insurance (Takaful) market. J Int Financ Mark Inst Money 56:71–92
Kumbhakar SC, Lovell CAK (2000) Stochastic frontier analysis. Cambridge University Press, New York
Kumbhakar SC, Parmeter CF, Zelenyuk V (2020) Stochastic frontier analysis: foundations and advances I. In: Ray SC, Chanibers R, Kumbhakar SC (eds) Handbook of production economics. Springer, Singapore, pp 1–40
Lee HS, Cheng FF, Har WM, Nassir AM, Ab Razak NH (2019) Efficiency, firm-specific and corporate governance factors of the Takaful insurance. Int J Islam Middle East Financ Manag 12(3):368–387
Leverty JT, Grace MF (2010) The robustness of output measures in property-liability insurance efficiency studies. J Bank Financ 34(7):1510–1524
Leverty JT, Grace MF (2012) Dupes or incompetents? An examination of management's impact on firm distress. J Risk Insur 79(3):751–783
Lim S, Zhu J (2016) A note on two-stage network DEA model: frontier projection and duality. Eur J Oper Res 248:342–346
Lim S, Zhu J (2019) Primal-dual correspondence and frontier projections in two-stage network DEA models. Omega 83:236–248

Lim QM, Lee HS, Har WM (2021) Efficiency, productivity and competitiveness of the Malaysian insurance sector: an analysis of risk-based capital regulation. Geneva Pap 46:146–172
Lu WM, Wang WK, Kweh QL (2014) Intellectual capital and performance in the Chinese life insurance industry. Omega 42:65–74
Ma YL, Pope N, Xie X (2014) Insurer performance and intermediary remuneration: the impact of abandonment of contingent commissions. Geneva Pap 39:373–388
Maniadakis N, Thanassoulis E (2014) A cost Malmquist productivity index. Eur J Oper Res 154(2):396–409
Mayer D, Smith CW (1988) Ownership structure across lines of property-casualty insurance. J Law Econ 31:351–378
Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18:435–444
Mester LJ (1989) Testing for expense preference behavior: mutual versus stock savings and loans. RAND J Econ 20:483–498
Nourani M, Chandran VGR, Kweh QL, Lu WM (2018) Measuring human, physical and structural capital efficiency performance of insurance companies. Soc Indic Res 137:281–315
Nourani M, Kweh QL, Ting IWK, Lu WM, Strutt A (2022) Evaluating traditional, dynamic and network business models: an efficiency-based of Chinese insurance companies. Geneva Pap 47:905–943
Ohene-Asare K, Asare JKA, Turkson C (2019) Dynamic cost productivity and economies of scale of Ghanaian insurers. Geneva Pap 44:148–177
Peng L, Lian Z (2021) Diversification and efficiency of life insurers in China and India. Geneva Pap 46:710–730
Ray SC (2004) Data envelopment analysis: theory and techniques for economics and operations research. Cambridge University Press, New York
Reyna AM, Fuentes H (2018) A cost efficiency analysis of the life insurance industry in Mexico. J Prod Anal 49:49–64
Reyna AM, Fuentes HJ, Núñez JA (2022) Response of Mexican life and non-life insurers to the low interest rate environment. Geneva Pap 47:409–433
Rubio-Misas M (2022a) Analysis of insurers' performance using frontier efficiency and productivity methods. The great contributions by David Cummins and Mary Weiss. Risk Manag Insur Rev 25(4):445–489
Rubio-Misas M (2022b) Bancassurance and the coexistence of multiple insurance distribution channels. Int J Bank Mark 40(4):724–745
Rubio-Misas M, Fernández-Moreno M (2017) Solvency surveillance and financial crisis: evidence from the Spanish insurance industry. Span J Financ Account
Shaddady A (2022) Business environment, political risk, governance, Shariah compliance and efficiency in insurance companies in the MENA region. Geneva Pap 47:861–904
Shavitha B, Banerjee S, Shetty A (2019) Product diversification versus technical efficiency of conglomerate life microinsurance companies: evidence from India. Geneva Pap 44:527–547
Tone K, Tsutsui M (2010) Dynamic DEA: a slack-based measure approach. Omega 38:145–156
Tone K, Kweh QL, Lu WM, Ting IWK (2019) Modelling investments in the dynamic network performance of insurance companies. Omega 88:237–247
Wanke P, Barros CP (2016) Efficiency drivers in Brazilian insurance: a two-stage DEA meta frontier-data mining approach. Econ Model 53:8–22
Wu YC, Kweh QL, Lu WM, Azizan NA (2016) The impacts of risk-management committee characteristics and prestige on efficiency. J Oper Res Soc 67:813–829
Yaisawarng S, Asavadachanukorn P, Yaisawarng S (2014) Efficiency and productivity in the Thai non-life insurance industry. J Prod Anal 41:291–306

Index

A

Accident, 18–22, 25, 29, 34, 89, 230, 235, 339, 353, 354, 363, 364, 386–392, 402–404, 406, 407, 412, 417, 422, 432, 447–450
Accident experience, 22
Accidents distribution, 364, 386, 387, 392
Adverse selection
  competitive markets, 21
  empirical tests, 21
  monopoly, 20, 22, 24
  moral hazard, 6, 17, 19, 21, 24–25, 201, 258, 338, 340, 343, 347–349, 355, 361–364, 386, 388–391, 395
  risk classification, 7, 21, 234, 239
  Rothschild-Stiglitz, 20–23, 339, 362
  screening, 23, 211
  signaling, 23, 24
Alternative risk transfer, 170–171, 200, 213–214, 217, 274, 277
Asymmetric information
  adverse selection, 17, 21, 24, 338, 339, 361, 362, 386–388
  econometric models, 380–381
  empirical analysis, 361–395
  experience rating, 388
  insurance fraud, 20, 417
  liability insurance, 18
  life insurance, 21
  moral hazard, 7, 17, 338, 361, 386, 388
Asymmetric learning, 19, 20, 25, 389, 390, 395
Auditing, 18, 40, 362, 385, 417, 424
Automobile insurance, 19–21, 25, 57, 69, 70, 228, 229, 232, 363, 386, 389–392, 420, 421, 423, 432, 447–449, 459, 461

B

Bonus-malus, 19, 362, 390, 391, 402, 447–451
Build-up, 57, 140, 157, 212, 267, 271, 278, 296, 327, 344

C

Catastrophe bond pricing, 184
Catastrophe bonds, 72, 73, 103, 152–153, 155, 169–190, 274
Catastrophe insurance, 100, 102–105, 117, 153, 155, 169, 240
Catastrophe options, 37
Catastrophe risk, 37–40, 103, 104, 169–174, 176–178, 183–186, 189, 258, 275
Categorization, 22–24, 170–173, 317, 437
Causality, 19, 40, 87, 231, 361–395
Causality effects models
  difference-in-differences, 369–370
  generalized method of moments, 362, 392–394
  instrumental variables, 362, 367–369, 372–374, 376, 377, 379, 381, 383, 395
Causality in applied econometrics, 362, 367, 371–376
Causal tests, 389–392
Claim frequency, 391, 392, 404, 405, 409
Classification, 7, 21, 23, 202, 229, 234, 239, 252–253, 303, 361, 385–387, 414, 418–422, 427, 429, 433–435
Climate change, 30–32, 55–57, 63, 67, 71–73, 75, 76, 81–86, 89, 91–93, 99, 103, 118, 124, 128, 131, 138, 143, 145–150, 155, 156, 190, 206, 231
Climate events, 30, 99–118
Climate risk, 5, 14, 31, 32, 37, 70–72, 75, 81–94
Coinsurance, 6, 9–11, 18, 153, 260, 271, 349, 362
Contract choice, 19, 229, 364, 386–388, 390, 407
Correlated risks, 88
Costs of claims, 452, 458
Count-data models, 404, 411
COVID-19, 111, 136, 145, 190, 210, 231, 237, 251, 252, 254–270, 272, 274–278, 291, 315, 333, 432
Cyber insurance, 190, 199–220, 271
Cyber insurance market, 190, 199, 200, 205, 210–212, 214–217

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

G. Dionne (ed.), *Handbook of Insurance*,
https://doi.org/10.1007/978-3-031-69561-2

Cyber risk
asymmetric information, 199, 200, 215, 218
heavy tail distribution, 212, 215
model risk, 214
non-linear dependencies, 200, 203, 214, 215

D
Data, 19, 58, 82, 101, 132, 176, 201, 227, 260, 291, 312, 337, 361, 402, 418, 447, 463
Data envelopment analysis (DEA), 464–466, 468–472, 477–488
Data science, 403, 417–442
Demand for insurance, 8–11, 16, 22, 67, 70, 81, 83, 86–92, 291, 340
Dependent risks, 201
Digital transformation, 225–228, 231, 233, 236–238, 245, 246, 423
Disaster damages, 55, 63, 73, 82, 110, 116, 117
Disaster insurance, 55–76, 81, 82, 85–93, 101–105, 117, 118
Divestment market
demand, 314–319
supply, 314–319
factors determining, 347
improved division, 353–355

E
Econometric estimation
accidents distribution, 364, 386
adverse selection, 364
asymmetric information, 364, 380, 386
insurance fraud, 417, 419–425, 442
moral hazard, 364, 386
Econometric methodology, 378
Econometric models, 20, 371, 373, 380–381, 404, 407, 419, 442
Economic equilibrium, 8, 17, 19
Economics of risk and uncertainty, 17
Economics on causality, 366–371
Economies of scale and scope, 475, 483
Efficiency and division of labor, 338, 345
Empirical estimation, 75, 183, 184, 340, 367, 370, 391, 468
Empirical measure, 92, 101, 361, 364, 395
Empirical results, 8, 216, 383, 439
Empirical tests
adverse selection, 20–25, 340–342, 363, 390
adverse selection *versus* moral hazard, 21
dynamic contracts, 390
insurance fraud, 432
moral hazard, 21
Rothschild-Stiglitz model, 21
Equilibrium, 4, 6–8, 17–23, 33, 34, 39, 85, 88, 91, 230, 340, 342, 344, 347, 350, 356, 362, 383, 474, 480, 481
Expected utility
optimal insurance, 9, 16
transaction costs, 186
Experience rating
bonus-malus systems, 390, 402, 448–451
costs of claims, 388
count models, 411
empirical results, 448
longitudinal data, 390
Poisson models, 404, 411
Extreme weather events, 30, 31, 83–85, 93

F
Finance and insurance, 4, 5, 30, 33–39
Financial stability, 85, 311, 319, 325–333
Financial theory, 5, 33
Firm performance, 26, 463–468, 473, 479, 489
Firm risk, 365, 366, 377
Frequency of claims, 203, 229, 391, 392, 403–407, 409, 449
Frontier efficiency, 463–465, 467–472, 479, 482–486, 488, 489

G
Generalized method of moments (GMM), 36, 362, 392–395, 457–460
Geneva Association, 3, 4, 28, 36, 231

H
Household damages, 99, 101, 107–110, 116, 117
Household financial resilience, 99–118
Household recovery, 100–102, 106, 109–111, 113, 117, 118

I
Incentive for safe driving, 389–391, 448, 450
Information, 4, 60, 84, 101, 126, 176, 199, 228, 252, 299, 338, 361, 401, 417, 447, 464
Information problems, 19, 24, 361–365, 386, 388, 390, 395
Insurability, 32, 36, 37, 57, 124–128, 148–154, 200, 201, 205, 211, 213, 215, 217, 229, 231, 252, 257, 258, 260, 268, 278
Insurability of wildfire losses, 124–129, 148–151
Insurance benefits, 66, 354, 497
Insurance claims frequency, 391, 392, 404, 405, 409
Insurance classification, 421, 422
Insurance contracts, 5, 9, 10, 17–20, 23, 25, 30, 34, 37, 57, 61, 62, 65, 66, 70, 82, 89, 104, 124, 126, 127, 234, 239, 258, 263, 265, 290, 292–294, 297, 298, 304, 305, 326, 345, 362, 385–388, 417, 432, 447–449, 459
Insurance demand, 5, 6, 9–12, 14, 33–35, 37, 65, 67, 69, 74, 83, 86–92, 104, 107, 272, 361, 362, 392–395
Insurance economics, 3–40, 87, 211, 215, 463, 464
Insurance fraud, 20, 115, 242, 417–442
Insurance fraud analysis, 419–426
Insurer performance
econometric frontier method, 464, 466

economies of scale and scope, 473, 479–481
efficiency and productivity, 463–489
efficiency and productivity concepts, 463, 467, 485
efficiency score, 465, 487
insurance inputs, 322, 467, 490–501
insurance output, 465, 467, 490–501
mathematical programming method, 464, 465
total factor productivity, 322, 477, 486–487
Insuring wildfire losses, 124–129
InsurTech, 225–246

L
Labor productivity, 24
Life insurance
asymmetric information, 4, 6, 17–25, 199, 218, 339, 342, 361–395
contract, 9, 10, 18, 65, 66, 70, 239, 294, 345, 387, 417
distribution, 8, 25–26, 231, 242, 306, 423, 486
economic accounts, 21
equity-indexed contracts, 40, 300–302, 304, 305
financial products, 64, 289–306
interest guarantees, 297
life insurance products, 297
looked-back guarantees, 294, 295, 297
non-exclusivity, 36, 127, 240
possibilities, 10, 226, 228, 231, 233, 238, 294, 314, 414, 425
private equity, 311–333
unit-linked contracts, 40, 263, 289, 290, 292–294, 297, 298, 302–306
world insurance market, 29, 177, 344, 385, 417, 461
Liquidity creation, 35, 361, 362, 392–395

M
Machine learning algorithms, 40, 187, 188, 227, 229, 233–235, 401–414, 417–422, 425, 427, 430, 438, 439, 441
Managing wildfire losses, 62, 124–137, 143, 147–151, 155
Market capacity, 73, 227, 313
Market characteristics, 177–183
Market regulation, 71
Market structure, 4, 8, 25–33, 70, 75, 257, 284, 489
Memory in longitudinal data, 25, 390
Miyazaki–Wilson–Spence equilibrium (MWS), 230
Monopoly, 22, 24, 128, 153, 339

N
Natural disaster, 29, 55, 81, 99, 126, 169, 271
Natural disaster insurance
demand, 230, 234, 242
supply, 30, 33, 85, 100, 103, 104
No-claim discounts, 447, 449, 450, 452, 453, 456, 461

O
Optimal insurance
adverse selection, 20–24
linear transaction costs
moral hazard, 17–20
nonlinear transaction costs, 18
risk aversion of the insurer, 6, 11, 16
Optimal insurance contracts
adverse selection, 20
liability insurance, 14, 37
moral hazard, 19
Optimal risk sharing
adverse selection, 212
moral hazard, 212, 276
Option pricing, 34, 185, 294, 315

P
Pandemic risk, 190
Pandemic risk insurability, 251–278
Pandemic risk management, 251–278
Panel data, 364, 370, 390, 391, 393, 394, 406, 407, 448, 451, 457–459
Partial insurance coverage, 6, 11, 14, 20, 23, 362, 363
Performance of insurance firms, 464–468
Poisson model, 203, 404, 405, 411, 458
Policy holder protection, 311, 319, 325, 326
Predictive performance, 419, 421, 429, 432, 433, 439–441
Pricing, 4, 9, 15, 18, 23, 29, 34, 71, 72, 76, 84, 103, 114, 149, 152, 170, 183–190, 229, 236, 259, 271, 275, 295, 324, 389, 392, 406, 410, 451
Pricing methods, 185, 410–412
Private and social insurance
division of labor, 257, 338, 345
efficient asset allocation, 296, 305
improved division of labor, 353–355
Private equity investment, 311–333
Private equity ownership, 319–333
Private information, 22–24, 68, 229, 230, 364
Productivity analysis, 464–468, 472
Property-liability insurance, 4, 7, 29, 480, 483

Q
Q-forward, 186

R
Regulatory constraints, 57, 257
Reinsurance demand, 35, 37, 361, 362, 392–394
Reinsurance market, 4, 30, 37–39, 72, 73, 128, 148, 152, 187, 480, 481, 484
Risk attitude, 4, 14, 88, 365
Risk aversion, 4–6, 8, 11, 13, 14, 16, 17, 21, 22, 66, 185, 301, 344, 346, 348–350, 364, 386, 387
Risk management, 7, 12, 26, 31, 33, 35, 36, 40, 83, 100, 107, 113, 143, 155, 189, 199, 201, 205, 210, 211, 216, 217, 236, 238, 251–278, 303–306, 314, 361–395, 482
Risk scores, 403, 413
Risk transfer
affordability, 85, 93, 278
capacity, 277, 278

Rothschild-Stiglitz equilibrium
empirical tests, 21
existence, 23

S
Separating equilibrium, 7, 20, 21
Signaling model, 4, 18, 23, 24, 112, 211
Single crossing property, 13
Single-period contracts
adverse selection, 388
moral hazard, 388
Social insurance, 23, 24, 36, 37, 276, 337, 338, 342, 349, 353–355
Statistical models, 233, 407, 448–459, 461
Structural characteristics, 60, 173–178
Systemic risk, 28, 200, 202, 205, 217, 258, 330, 333

T
Technical efficiency, 27, 466, 467, 472, 479, 481, 482, 485, 486
Telematic data, 229, 402–414
Telematic information, 40, 401–414
Telematics, 228–230, 234, 401–414
Telematics in insurance
contract choice, 229
ethical issues, 413–414
machine learning algorithms, 410
regulatory constraints, 413–414
telematics data, 414
Total factor productivity, 322, 467, 477, 486–487
Traffic violations, 389, 448, 450
Transaction costs, 3, 6, 21, 22, 33, 35, 85, 213, 296, 306, 344, 387

U
Unemployment insurance, 74, 112, 370

V
Verification, 18, 176, 217, 339, 437

W
Wage rates, 345, 471, 490, 491, 493
Wildfire losses, 62, 124–137, 146–154
Wildfire risk, 60, 64, 123–125, 128, 137–147, 152, 155–158
Wilson equilibrium, 5, 7, 20, 342