Essays in Causal Machine Learning

DISSERTATION
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Daniel Goller

from

Germany

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Dr. Tim Pawlowski

Dissertation no. 5106

ADAG 4farbig, St.Gallen, 2021

Essays in Causal Machine Learning

DISSERTATION
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

Daniel Goller

from

Germany

Approved on the application of

Prof. Dr. Michael Lechner

and

Prof. Dr. Tim Pawlowski

Prof. Francesco Audrino, PhD

Dissertation no. 5106

ADAG 4farbig, St.Gallen, 2021

The University of St.Gallen, School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science, hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, May 28, 2021

The President:

Prof. Dr. Bernhard Ehrenzeller

Acknowledgements

To write a dissertation is a great privilege. I would like to thank my supervisor Michael Lechner for giving me this opportunity. His support and guidance were most important during the dissertation phase and his helpful and target-oriented input and feedback kept me on the right track. I appreciate the pleasant atmosphere we had at the chair with stimulating and entertaining conversations with colleagues and friends - Hugo Bodory, Daniel Boller, Petyo Bonev, Hannah Busshoff, Sandro Heiniger, Georgi Lautliev, Jana Mareckova, Carina Steckenleiter, Anthony Strittmatter. I have learned a lot, not only academically, but also personally. A special thank you goes to Michael Knaus, Gabriel Okasa, and Michael Zimmert, who had an open ear for all my questions and problems at all times, and Alex Krumer, who gave me a different perspective on what is important in academia.

To write a dissertation is an exciting journey in which not only my colleagues participated, but also my friends in St.Gallen and around the world. Tobias, Filippo, Max, Robert, Leonie, Vitali, Chris, Piotr, etc. – thank you for being in my life. Moreover, thanks to my father, Rainer, for his support and my brother, Nico, for inspiring conversations.

To write a dissertation implies to experience ups and downs. Being able to share both is essential for getting through this time well. I was and I am happy to share those moments with two special persons. First, my mother, Lore, who made me what I am today and who trusted me in every decision I made. Second, my partner, Gitti, who has experienced the ups and downs the most, while doubling the joy and halving the burden. I am deeply grateful to both of you for your unconditional support.

St.Gallen, June, 2021 Daniel Goller

Table of Contents

Table of Contents v

List of Tables viii

List of Figures x

Summary xi

Zusammenfassung xii

1 Introduction 1

2 Analysing a built-in advantage in asymmetric darts contests using causal machine learning 3

2.1 Introduction 4

2.2 Literature review 7

2.3 Setting and data 9

2.3.1 Setting 9

2.3.2 Database 10

2.3.3 Descriptive statistics 11

2.4 Methodology 12

2.4.1 Notation and framework 12

2.4.2 Identification 13

2.4.3 Methods 14

2.5 Results 17

2.5.1 Average treatment effects 17

2.5.2 Individualized treatment effects 18

2.5.3 Group average treatment effects 20

2.5.4 Sensitivity checks 23

2.5.5 Discussion 25

2.6 Conclusion 27

Appendices 29

3 Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed 37

3.1 Introduction 38

3.2 Institutional background 42

3.3 Data 43

3.3.1 Dataset 43

3.3.2 Treatment and sample selection 44

3.3.3 Descriptive statistics 44

3.4 Methodology 46

3.4.1 Target, notation and identification 46

3.4.2 Empirical Monte Carlo Simulation 48

3.4.3 Matching estimator 51

3.4.4 Propensity score estimation 51

3.5 Simulation 55

3.6 Empirical application 61

3.7 Conclusion 63

Appendices 65

4 Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues 78

4.1 Introduction 79

4.2 Description of the leagues' schedules 83

4.2.1 General structure of the leagues 83

4.2.2 The French Ligue 1 84

4.2.3 The Spanish La Liga 85

4.2.4 The German Bundesliga 1 85

4.2.5 The English Premier League 85

4.3 Data and descriptive results 86

4.3.1 Database 86

4.3.2 Definition of heterogeneity 87

4.3.3 Variables and descriptive statistics 88

4.4 Empirical strategy 94

4.4.1 Selection into treatment 94

4.4.2 Estimation 95

4.5 Results 97

4.5.1 Pooled data 97

4.5.2 Individual leagues 100
4.5.3 Robustness tests 103
4.5.4 Discussion 104
4.6 Conclusion 105
Appendices 107
References 117

List of Tables

Table 2.1: Descriptive statistics 11
Table 2.2: Base results - best linear predictors 18
Table 2.3: Difference in characteristics of 10% most and least affected 20
Table 2.4: Social pressure vs. home (dis-)advantage 22
Table 2.5: Average and group average effects, MCF 24
Table 2.6: Base results - best linear predictors, sensitivity to clustering 25
Table 2.7: Descriptive statistics, all variables 29
Table 2.8: Different definitions of 'Home' 32
Table 2.9: List of tournaments 35
Table 3.1: Descriptive statistics of selected covariates 45
Table 3.2: Empirical Monte Carlo Study 49
Table 3.3: Summary of DGP's 50
Table 3.4: Summary of simulation results, propensity score estimation 56
Table 3.5: Matching-quality 58
Table 3.6: Summary of simulation results, matching 60
Table 3.7: Empirical treatment effect estimation, matching 61
Table 3.8: Covariate balancing in application 62
Table 3.9: Simulation results for N=4,000 and ~10% share of treated 65
Table 3.10: Propensity score results for N=4,000 and ~10% treated 66
Table 3.11: Simulation results for N=4,000 and ~25% share of treated 67
Table 3.12: Propensity score results for N=4000 and ~25% treated 68
Table 3.13: Simulation results for N=16,000 and ~5% share of treated 69
Table 3.14: Propensity score results for N=16,000 and ~5% treated 70
Table 3.15: Simulation results for N=16,000 and ~10% share of treated 71
Table 3.16: Propensity score results for N=16,000 and ~10% treated 71
Table 3.17: Simulation results for N=16,000 and ~25% share of treated 72
Table 3.18: Propensity score results for N=16,000 and ~25% treated 72
Table 3.19: Radius matching protocol 77
Table 4.1: Frequency of matches by days 84
Table 4.2: Descriptive statistics of the main variables 89
Table 4.3: Descriptive statistics, Ligue 1 & La Liga 91
Table 4.4: Descriptive statistics, Premier League & Bundesliga 92
Table 4.5: Levels and effects of playing on four non-frequent days, all data 98
Table 4.6: Effects of different definitions of non-frequent days, all data 99

Table 4.7: Effects of playing on four non-frequent days, leagues; Panel A ... 101
Table 4.8: Effects of playing on four non-frequent days, leagues; Panel B ... 102
Table 4.9: Descriptive statistics for all available variables 107
Table 4.10: Estimation of propensity score 112
Table 4.11: Effects of playing on five non-frequent days for each league 114
Table 4.12: Effects excluding the third most frequent day for each league ... 115

List of Figures

Figure 2.1: Sorted Effects 19
Figure 2.2: Built-in advantage of starting contestant by ability 21
Figure 2.3: Differential built-in advantage by home and ability 23
Figure 2.4: BIA by maximum number of legs 26
Figure 2.5: Propensity scores 31
Figure 2.6: Differential BIA by ability, alternative specification 32
Figure 2.7: Built-in advantage of starting contestant by ability; MCF 33
Figure 3.1: Propensity scores by treatment status, N=4,000, 10% treated 57
Figure 3.2: Propensity scores by treatment status, N=16,000, 10% treated 58
Figure 3.3: Propensity scores by treatment status, N=4,000, 25% treated 73
Figure 3.4: Propensity scores by treatment status, N=16,000, 5% treated 74
Figure 3.5: Propensity scores by treatment status, N=16,000, 10% treated 74
Figure 3.6: Propensity scores by treatment status, N=16,000, 25% treated 75
Figure 3.7: Propensity scores by treatment status, Probit 75
Figure 3.8: Propensity scores by treatment status, Random Forest 76
Figure 3.9: Propensity scores by treatment status, LASSO 76

Summary

This doctoral thesis contains three studies in different fields of economics. The first and third chapter are in the field of operational research. The second chapter investigates a methodological question in the field of labour economics. Common to all chapters is that they deal with causal questions and that tools from the machine learning literature are used to answer these questions.

The first chapter investigates a sequential contest with two players in darts, in which one of the contestants enjoys an advantage. The first-mover in those contests has potentially more, but never less moves in the match, which results in higher winning probability for the starting player. The application of causal machine learning methods shows that contestants with low performance measures and little experience benefit the most from moving first. Furthermore, contestants performing in a venue close to their hometown do not experience this advantage, unlike contestants competing in venues away from their hometown.

The second chapter investigates whether methods from the machine learning toolbox for estimating the propensity score in matching estimators lead to more credible estimations. A main finding from the chapter is that the choice of the method is highly relevant. Especially for empirical applications in active labour market policy estimating the propensity score by LASSO based logit models delivers most credible results, while the usage of Random Forests may lead to a deterioration of the performance in specific situations.

In the third chapter, this well performing combination of LASSO-based estimation of the propensity score in a radius-matching estimator found in the second chapter is used as estimation strategy. This work investigates schedule-related issues in the European football leagues and establishes differential winning probabilities in home matches for underdog teams depending on the schedule. The findings suggest that the current schedule favours underdog teams with fewer home matches on days that do not correspond to the usual match days.

Zusammenfassung

Diese Doktorarbeit enthält drei Studien in verschiedenen Bereichen der Volkswirtschaftslehre. Das erste und das dritte Kapitel sind im Bereich der operationellen Forschung angesiedelt. Das zweite Kapitel untersucht eine methodische Frage auf dem Gebiet der Arbeitsökonomie. Allen Kapiteln ist gemeinsam, dass sie sich mit kausalen Fragen befassen und dass zur Beantwortung dieser Fragen Methoden aus der Literatur des Machine Learnings verwendet werden.

Im ersten Kapitel wird ein sequentieller Wettkampf mit zwei Dartspielern untersucht, bei dem einer der beiden Teilnehmer einen Vorteil genießt. Der Spieler, welcher das Spiel anfangen darf, hat potenziell mehr, aber nie weniger Züge im Spiel, was zu einer höheren Gewinnwahrscheinlichkeit für den startenden Spieler führt. Bei der Anwendung sogenannter Causal Machine Learning Methoden zeigt sich, dass Wettkämpfer mit niedrigen Fähigkeiten und wenig Erfahrung am meisten davon profitieren starten zu dürfen. Weiterhin haben Teilnehmer, die an einem Ort in der Nähe ihrer Heimatstadt Wettkämpfe bestreiten, diesen Vorteil nicht, im Gegensatz zu Teilnehmern, die in Städten entfernt von ihrer Heimatstadt antreten.

Im zweiten Kapitel wird untersucht, ob Methoden des Machine Learnings zur Schätzung der Propensity Scores in Matching Schätzern zu glaubwürdigeren Ergebnissen führen im Vergleich zu klassischen Methoden. Als eine Schlussfolgerung dieses Kapitels kann festgestellt werden, dass die Wahl der Methode von hoher Relevanz ist. Insbesondere für empirische Programmevaluation in der aktiven Arbeitsmarktpolitik liefert die Schätzung des Propensity Scores durch LASSO-basierte Logit-Modelle die glaubwürdigsten Ergebnisse, während die Verwendung von Random Forests in bestimmten Situationen zu weniger zuverlässigen Ergebnissen führen kann.

Im letzten Kapitel wird diese gut funktionierende Kombination einer LASSO-basierten Schätzung des Propensity Scores in einem Radius-Matching Schätzer, wie in Kapitel 2 gefunden, verwendet. In dieser Arbeit werden Spielplan bezogene Fragen in den europäischen Fussball-Ligen untersucht und unterschiedliche Gewinnwahrscheinlichkeiten für Aussenseiter-Teams im Zusammenhang mit dem Spielplan gefunden. Die Ergebnisse deuten darauf hin, dass der aktuelle Spielplan Aussenseiter-Teams mit weniger Heimspielen an Tagen, die nicht den üblichen Spieltagen entsprechen, begünstigt.

1 Introduction

The use of methods originating in the machine learning literature to answer causal questions has developed in recent years. The rise of so-called causal machine learning methods promises to become at least an extension to the classical methods, if not the new standard in the evaluation of economic problems. At the same time, research in this area is still in its development phase. The three parts of this thesis contribute to a better understanding of the newly developed possibilities and emerging challenges in using machine learning techniques for analysing causal questions.

Each of the chapters is an independent study and can be read without knowledge of the others. Nevertheless, the chapters, apart from the overall theme, share some features. One chapter deals with improving credibility of a classical method by implementing machine learning. The knowledge gained in this work is used in the two chapters of this thesis answering empirical questions in operational research.

The first, single authored, chapter is investigating fairness in a specific form of sequential contests in the sports of darts. In those contests there is a built-in advantage for one of the contestants, which is found to substantially increase the winning probability for the specific contestant. By using causal machine learning methods, this chapter thoroughly investigate this built-in advantage. The results show that individuals benefit from this advantage in different degrees. Especially, the most experienced and high ability contestants profit least. Furthermore, contestants do not benefit from this advantage at all if they compete in a contest close to their home town. This phenomenon is found to occur due to social pressure from the audience.

The second chapter is joint work with Michael Lechner, Andreas Moczall, and Joachim Wolff. It is a simulation study in the context of programme evaluation for active labour market policy for long-term unemployed. It investigates whether machine learning method for estimating the propensity score in a matching-type estimator lead to more credible estimates. The study concludes that the choice of the method is highly relevant. Using a Random Forest to estimate the propensity score leads to a deterioration of the performance in specific situations, which are common in the evaluation of active labour market policy. The most credible results are found for estimating the propensity score using a LASSO-based Logit. Finally, an evaluation of training programmes for long-term unemployed with a rich data set shows that the estimates vary considerably depending on the method chosen to predict the propensity score.

This LASSO-based Logit to estimate the propensity score in a matching-type estimator is used in the third chapter, which is joint work with Alex Krumer. This study investigates the well-known phenomenon of a home advantage in soccer. More specifically, the home advantage for teams depending on the schedule in the English, French, Italian and German football leagues are investigated. Differential home advantages are found for underdog teams in case they play on unusual compared to usual match days. This finding highlights the importance of ensuring a fair schedule. Otherwise, underdog teams that play fewer home matches on days that are not the usual match days will be favoured. The study also finds that fewer visitors come to the stadiums on these unusual match days. This implies that underdog teams with more home games on these days are disadvantaged in two ways: Financially and in terms of fairness.

2 Analysing a built-in advantage in asymmetric darts contests using causal machine learning

Daniel Goller

Abstract: We analyse a sequential contest with two players in darts where one of the contestants enjoys a technical advantage. Using methods from the causal machine learning literature, we analyse the built-in advantage, which is the first-mover having potentially more but never less moves. Our empirical findings suggest that the first-mover has an 8.6 percentage points higher probability to win the match induced by the technical advantage. Contestants with low performance measures and little experience have the highest built-in advantage. With regard to the fairness principle that contestants with equal abilities should have equal winning probabilities, this contest is ex-ante fair in the case of equal built-in advantages for both competitors and a randomized starting right. Nevertheless, the contest design produces unequal probabilities of winning for equally skilled contestants because of asymmetries in the built-in advantage associated with social pressure for contestants competing at home and away.

Keywords: Causal machine learning, heterogeneity, contest design, social pressure, built-in advantage, incentives, performance, darts

JEL classification: C14, D02, D20, Z20

2.1 Introduction

Contest designers put a lot of effort in designing contests in various fields. For sports contests, fairness for the contestants and attractiveness for the spectators are the main restrictions (Szymanski, 2003; Wright, 2014; Arlegi and Dimitrov, 2020). The main fairness criterion for contests is fulfilled if equally skilled contestants have equal winning probabilities (e.g. Arlegi and Dimitrov, 2020). To ensure this, most contest designs balance potential advantages, e.g. home and away matches within one contest to balance the home advantage. Hence, many contest studies analyse symmetric contests that are designed to have either a balanced or no technical advantage but empirically uncover side effects, such as the order of actions (Ginsberg and Van Ours, 2003) or scheduling effects (Goller and Krumer, 2020).

In contrast, we investigate the fairness of an unbalanced, one match sequential contest with two competitors in the sport of darts. This contest has a built-in advantage (BIA) for the first moving contestant, who has an advantage in potentially more, but never less situations within a match.[1] In this study, we analyse a simple question: Is the two player, sequential contest with BIA a fair contest, or are there specific groups of individuals who are systematically disadvantaged by the contest design?

The outcome of a contest is about human behaviour, and even though behavioural responses are potentially among the most individual, there is hardly any personalized evidence. The importance of individual behavioural responses calling for more individualized effects estimation is demonstrated in the work of González-Díaz, Gossner, and Rogers (2012). They found large differences of individual 'critical abilities' of tennis players affecting the performance and success. While it is impossible to observe or estimate a true individual effect owing to the unobservable outcomes in situations of non-realized interventions, there is a huge potential to find or describe those individuals who benefit most or least from some given treatment, which is the BIA in the contest design. For designing rules for competitions, identifying the most benefiting group of individuals might help improve contest design in terms of fairness, attractiveness, or competitive balance. This is especially true if specific groups of

[1] Sequential contests with different kinds of advantages are widely present. Examples are legislators giving an advantage, in form of patents, to a first moving contestant in a market, job posting committees giving an advantage to specific groups of the population in a contest for a job, incumbency advantage in political elections, or in sports contests such as the snooker championship or NBA and NHL playoffs.

individuals systematically realize differential effects and are therefore (dis-) advantaged by the contest design.

The relation of incentives and performance is widely discussed in the literature and is among the fundamentals in contests (Baumeister, 1984; Rosen; 1986; Ehrenberg and Bognanno, 1990; Ariely, Gneezy, Loewenstein, and Mazar, 2009). Humans behave differently in situations with increased level of incentives, which are present in many fields of daily life; such as job talks, presentations or speeches in the public, competitions and decisions at the workplace or in sports. Often those situations are subject to higher rewards or reputation that might increase or decrease the performance of individuals. In darts, social incentives might play a crucial role, owing to euphoric spectators and the otherwise very standardized environment. For athletes competing near their hometown the stakes are higher. Having a supporting crowd around might be motivating, which we call social support from now on. Social pressure is the counterpart, as being watched can also be a burden inducing pressure.[2] The increased focus magnifies the reputation gained if performing well, as well as the loss of reputation, otherwise. This is a known phenomenon in the social facilitation literature (e.g. Zajonc, 1965; Butler and Baumeister, 1998), which is found to have an influence even in cases in which not the athlete, but only the audience has high expectations (Baumeister, Hamilton, and Tice, 1985; Strauss, 1997).

This motivated us to analyse differential effects in the BIA on an individualized level, as well as of the respective groups associated with social incentives and ability in a flexible way. To the best of our knowledge we are the first to empirically analyse a BIA in an asymmetric two players sequential contest. This study especially contributes in investigating the fairness of such a contest with a thorough analysis of the BIA using state-of-the-art methodology.

This work relies on recent advances in econometric methods made in the emerging *Causal Machine Learning* literature. These new methods are more flexible compared to parametric approaches. Moreover, they can deal flexibly with rich data, which makes some identifying assumptions more credible, and they are useful for the analysis of heterogeneous treatment effects.

[2] While social support is supposed to improve performance, social pressure results in worse performance (Zajonc, 1965; Baumeister, Hamilton, and Tice, 1985).

For this, we use a dataset from the increasingly popular sports of darts, which offers several benefits when investigating individual responses. Particular for darts is a friendly but euphoric atmosphere created by the spectators. Moreover, conditions at different venues are highly standardized. A noteworthy advantage in comparison to various other sports is that there is no direct interaction between the contestants, which could lead to problems in assessing individual performance, otherwise. Furthermore, the outcome is precisely measurable, the rewards are high enough for competitors to take that task seriously, humans are observed while performing their usual job (Levitt and List, 2008), clear rules avoid subjective decisions by referees, and almost no external influences.

As expected, we found the average effect for the technical advantage of moving first to deliver an about 8.6 percentage points higher probability to win a match. Despite this advantage, the contest is ex-ante fair for players with symmetric BIAs and a randomized allocation to be the first-mover. The contest cannot be regarded as fair in case of systematically differential BIAs, which is systematic effect heterogeneity associated to specific groups. The empirical analysis shows equal BIAs for equally skilled contestants, but the personalized effects span from about -5 to +15 percentage points for individuals with different characteristics showing that there is a substantial heterogeneity in the effect of the BIA. Those contestants with lower performance measures and less experience profit most from the BIA. Differential effects are found in line with the social pressure hypothesis, as contestants playing in a neutral environment benefit from an 8.8 percentage points higher treatment effect compared to those playing in a supportive environment. Interestingly, this differential effect in the BIA is found only for the first-mover player, while the second moving player is not affected by this social pressure. Therefore, this is not a general home disadvantage but is related to social pressure only for the first moving contestant. This results in an unfair contest design, since equally skilled contestants do not necessarily benefit from equal winning probabilities.

In the following section the related literature is discussed. Section 2.3 provides an insight into the setting of darts and details the data used. Section 2.4 presents the methodological challenges and the estimation procedures used to obtain the results in Section 2.5, followed by a conclusion in Section 2.6.

2.2 Literature review

There are few studies on the sport of darts, starting with Tibshirani, Price, and Taylor (2011) analysing the statistics behind darts and the ways to get the highest expected payoffs. Liebscher and Kirschstein (2017) predicted winning probabilities for individual players in the world darts championship. While they implicitly incorporated the BIA in their prediction, they did not intend to estimate causal effects. Ötting, Deutscher, Langrock, Gehrmann, Schneemann, and Scholten (2020), and Klein Teeselink, van Loon, van den Assem, and van Dolder (2020) discovered no or low performance decrements under pressure for professional darts players by looking at more or less pressure owing to varying importance of situations within a match. Substantial decrements among youth and amateur players are found by Klein Teeselink et al. (2020), which might suggest a selection of more choking resistant individuals among professionals.

One of the most fundamental relations in contests is the role of incentives on performance, discussed in the behavioural and psychological (e.g. Baumeister, 1984; Masters, 1992; Butler and Baumeister, 1998; Ariely et al., 2009) and the economics literature (e.g. Stiglitz, 1976; Shapiro and Stiglitz, 1984; Rosen, 1986; Prendergast, 1999; Lazear, 2000).[3] Several works have investigated the role of social incentives, for example Cao, Price, and Stone (2011) and Toma (2017), which used data on basketball and found no effect of pressure on performance when a team plays at home compared to road games. Other works find this to be a source of pressure that impacts performance negatively. Ariely et al. (2009) found high social rewards to act counterproductive for performance in a set of experiments. Harb-Wu and Krumer (2019) investigated shooting performance in biathlon and found home athletes to miss more shots compared to athletes from other countries. Furthermore, the authors suggested that performance decrements are present only for the best ranked quarter of athletes. Despite the direct psychological effect, there is an indirect effect from public expectations (Baumeister, Hamilton, and Tice, 1985, Butler and Baumeister, 1998, Strauss, 1997). Furthermore, Baumeister et al. (1985) and Strauss (1997) found

[3] While there is a general agreement, that a small increase in incentives has a positive effect on performance, economic theory, such as the efficient wage hypothesis, suggests that stronger incentives lead to more effort and higher output. Behavioural research finds performance decrements under circumstances of increased importance mainly for skill-based tasks. As it can be assumed that darts is a skill-based task, we focus on research from this literature.

performance decrements among individuals in cases in which the audience, but not the individual expects success.

Symmetric contests with either a balanced or no advantage are analysed with regard to differential effects for specific groups in several works. Ginsburg and Van Ours (2003) found different winning probabilities in a musical contest depending on the order of actions. Dohmen (2008) and Harb-Wu and Krumer (2019) reported performance decrements among home contestants compared to non-home contestants. In other works, the groups of first-movers (e.g. Apesteguia and Palacios-Huerta, 2010) or the second-movers (e.g. Page and Page, 2007) are found to have higher winning probabilities than the respective other. Recently, Goller and Krumer (2020) established differential winning probabilities in home matches for underdog teams depending on scheduling of the matches. Asymmetric contests, in which certain contestants are favoured by the contest design, are examined in various theoretical works. Examples are the role of the incumbency advantage for autocrats' investment decisions (Konrad, 2002) or in political campaigns (Meirowitz, 2008), and more generally the influences of head-starts or handicaps in contests (Kirkegaard, 2012; Segev and Sela, 2014). In a more general scheme, a BIA is found for relative age effects in youth sports, in which children born at a certain time of year are favoured by the calendar year system, which groups children for competitive purposes (for a review, see Musch and Grondin, 2001; for a solution on how operational research can improve fairness, see Hurley, 2009).

Using sports data, machine learning methods are up to now almost exclusively used for prediction tasks, rarely in causal studies for average effects but, to the best of our knowledge, not yet for the systematic estimation of heterogeneous effects.[4] Especially in recent years, a number of methods have been proposed in the *Causal Machine Learning* literature for estimating effect heterogeneity (Wager and Athey, 2018; Lechner, 2018; Athey, Tibshirani, and Wager, 2019; Zimmert and Lechner, 2019; among others). Machine learning methods (or statistical learning methods, see e.g. Hastie, Tibshirani, and Friedman, 2009) were implemented or modified to suit the classical causal framework and developed to become useful in analysing causal questions (Athey, 2017; Athey and Imbens, 2019). While there is some work

[4] Examples of prediction tasks are forecasting the winner of the darts world championship (Liebscher and Kirschstein, 2017) or predicting the final table in the German soccer Bundesliga (Goller et al., 2018). Average causal effects estimations can be found for schedule-related issues in soccer, as in Krumer and Lechner (2018) and Goller and Krumer (2020). Further, Faltings, Krumer, and Lechner (2019) used a method from the causal machine learning literature in their work on favouritism in the Swiss soccer league but did not find any systematic heterogeneity in their effects.

establishing theoretical guarantees (e.g. Chernozhukov et al., 2018; Wager and Athey, 2018; Zimmert and Lechner, 2019) and simulating the performance of the newly developed estimators (notably Knaus, Lechner, and Strittmatter, 2020a), empirical applications of those new estimators and the systematic estimation of heterogeneous effects become increasingly popular (examples are Davis and Heller, 2017; Cockx, Lechner, and Bollens, 2019; Athey and Wager, 2019; Knaus, Lechner, and Strittmatter, 2020b).

2.3 Setting and data

2.3.1 Setting

Darts players commit themselves to play in one of the two major federations, the British Darts Organization (BDO) or the Professional Darts Corporation (PDC). While both federations hold their own tournaments, the PDC receives more media attention and distributes higher prize money than the BDO.[5] The most important tournaments are the (PDC/BDO) World Darts Championship held in December and January each year. Additionally, during the year, several other major tournaments are held in different places. The majors are joined by different minor tournaments throughout the year, relevant for accumulating ranking relevant prize money and qualification for the major tournaments. For a full list of tournaments considered in this work, see Appendix 2.E.

All BDO and PDC darts tournaments operate under the rules of the Darts Regulation Authority. Most matches are played in the (best-of-K) legs format, where K is an odd number. This implies that a player who wins (K+1)/2 legs wins the match. To win a leg, the contestants must score exactly 501 points and complete with a 'double' – a special region on the border of the darts board in which the score achieved is doubled. The two contestants perform their moves sequentially; for each move, three darts are to be thrown. Once a player reaches 501 points, the opponent is not allowed to 'catch-up'. Therefore, the first-mover in a given leg potentially has up to three darts more compared to the non-starting player. Having an odd number of maximum legs played, the first-mover has a technical advantage for the whole match.

[5] There are also combined tournaments for which the players of BDO and PDC can participate. Further, the players may change their affiliation any time.

Before each match the starter is determined in a shootout, which is one dart each player, with the player closest to the centre of the darts board starting the match. How this challenge of non-random determination of moving first is solved is discussed in Section 2.4.2. For a more formal model the interested reader can refer to Appendix 2.C.

2.3.2 Database

To investigate the BIA, we use data from the sports of darts. Starting from the year 2009 until 2019, matches from the most important BDO and PDC tournaments were extracted from the software, *dartsforwindows*, containing information on match statistics and outcomes. This resulted in a total of 11,604 matches played by 394 different players.[6] Importantly, the variable of interest, i.e. *starting the first leg*, as well as the outcome variable, *winning the match*, is generated. This is complemented by venue and tournament characteristics, such as prize money, which is standardized to make it comparable over the studied time period.[7] Personal characteristics of the players, such as nationality, hometown, and date of birth, among others are collected and utilized to create additional variables. Those contain the *age* or the number of years the player has played darts at the time of the respective match, or the distance between the hometown and the venue. *Home* and *Venue in country of birth* variables are created if the player lives within 100 kilometres of the venue, and is born in that country, respectively.[8] Since matches vary in their potential length, the data base contains the maximum number of legs to play (*bestoflegs*) in the respective match. Finally, (pre-match) performance measures and players statistics, such as the *3-darts average* (cumulated 3 darts score, averaged over all the matches in the past 2 years), *rankings*, etc., are added. For the full list of covariates and sources, the reader can refer to Appendices 2.A and 2.F, respectively.

[6] After excluding all matches that included players with less than 5 matches, and few other players whose characteristics were missing.

[7] Since prize money in tournament *m* raises substantially over the years *t* this is standardized as: $Prizemoney\ standardized = \frac{Prizemoney_{mt} - min_t(Prizemoney_t)}{\max_t(Prizemoney_t) - \min_t(Prizemoney_t)}$.

[8] The main definition of the variable *Home* of a radius of 100 kilometres around the venue is chosen to specify the hometown as close to the venue as possible, i.e. at a distance that family and supporters can easily accommodate, and to observe enough individuals as 'home contestants' to have sufficient statistical power. Further, to investigate the robustness regarding this decision the radius is varied to 50, 150, and 200 km around the venue.

2.3.3 Descriptive statistics

Table 2.1: Descriptive statistics

Variable	Mean	Starting Player	Non-starting Player	Diff / SD
	(1)	(2)	(3)	(4)
Game outcomes				
Wins match	0.50	0.55	0.45	0.10
Wins first leg	0.50	0.63	0.37	0.27
Game characteristics				
Starts first leg	0.50	1.00	0.00	
Televised	0.32			
Ranking tournament	0.90			
Prize money	51,145 (79,918)			
Prize money standardized	0.13 (0.25)			
Player characteristics				
3 darts average	91.82 (4.53)	91.99 (4.49)	91.65 (4.56)	7.54
Ranking	121.74 (292.77)	113.01 (254.99)	130.47 (326.21)	5.97
Accumulated Matches	231.15 (270.17)	238.13 (274.46)	224.17 (265.81)	5.17
Left handed	0.07	0.07	0.07	0.80
Years playing	19.97 (9.82)	20.07 (9.79)	19.87 (9.85)	2.05
Years playing professional	11.91 (6.21)	11.90 (6.17)	11.91 (6.25)	0.23
Age	36.54 (9.73)	36.46 (9.63)	36.61 (9.83)	1.57
Ven. in country of birth	0.38	0.38	0.38	0.21
Home	0.08	0.08	0.08	0.36
Distance to venue	1,410 (3,272)	1,423 (3,307)	1,398 (3,242)	0.77
Observations	11,604	11,604	11,604	

Notes: This table presents average values and standard deviations (in parentheses for non-binary variables) of selected variables. In the last column the difference for the outcome variables from starting (2) and non-starting (3) player, respectively the standardized difference for the player characteristics is reported. Each observation represents one match. Values that are equal for all columns are reported only once.

The outcome variable under consideration is *winning the match* and is depicted in the descriptive statistics presented in Table 2.1. Further, information is provided by categorising the players according to the variable of interest, i.e. *starting the first leg*, in columns (2) and (3). The difference in the outcome variables for starting and non-starting players can be observed in column (4). If starting the first leg is determined randomly (instead of performing the shootout) these numbers would be the average

treatment effects. Descriptive evidence that moving first is subject to selection effects can be observed in the standardized differences (SD) reported in column (4) for the players characteristics.[9] Starting players have on an average a higher *3-darts average*, a better *ranking* position, and more *accumulated matches*. For the other characteristics the standardized difference is low. The full list of variables can be found in Table 2.7 in Appendix 2.A.

2.4 Methodology

2.4.1 Notation and framework

The typical notation for binary treatment effects estimation, following Rubin (1974) is used. Suppose the outcome obeys the observational rule: $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$, where D_i denotes the treatment status $d \in (0,1)$, $Y_i(d)$ the potential outcome under treatment status d. Furthermore, we define X_i to contain the covariates necessary to account for confounding, and Z_i represents those variables investigated in the heterogeneity analysis.[10]

The first estimand of interest is the average treatment effect (ATE), $\theta = E(Y_i(1) - Y_i(0))$. This represents the average effect for all units on the highest level of aggregation. Contrarily, the estimand on the lowest aggregation level is the individualized average treatment effects (IATE), $\theta(x) = E(Y_i(1) - Y_i(0)|X_i = x)$. The group average treatment effect (GATE) represents an intermediate aggregation level according to heterogeneity variables Z_i, $\theta(z) = E(Y_i(1) - Y_i(0)|Z_i = z)$. Both conditional average treatment effects (CATEs), the GATEs and IATEs, are useful for detecting heterogeneity among the observed units, which are 'hidden' in ATE estimates.[11]

To understand the relationship of those three estimands of interest, integrating the IATEs over the characteristics of the groups $Z_i = z$ leads to the GATEs, while integrating over the characteristics of the entire population results in the ATE. Finally,

[9] It is unclear in the literature how large the standardized difference can be to still claim the average of the value of a respective variable to be balanced among treated and non-treated. Moreover, this is only an indication for average balancing. In this work we were very cautious even with seemingly low SD and took non-random treatment allocation seriously, as described in Section 2.4.2.

[10] In principle, X_i and Z_i might contain distinct variables or overlap, partly or completely. In this work Z_i is an ad-hoc selected subset of X_i, used for the heterogeneity analyses and investigating the hypotheses. Throughout the work, random variables are indicated by capital letters and realizations of these random variables by lowercase letters.

[11] GATE and IATE are special cases of the CATE. The more precise terms are used in this section; throughout this work, the terms are used synonymously.

only one of the potential outcomes is observable, as a unit can either be treated ($D_i = 1$) or non-treated ($D_i = 0$); the other remains counterfactual. In the following section, we shall discuss how to 'solve' this fundamental problem of causal inference (Holland, 1986).

2.4.2 Identification

Most crucial for estimating causal effects is a credible identification strategy. In the case of randomized treatment assignment, non-treated units can directly be used to construct the counterfactual outcome. In the underlying case, with non-random treatment assignment, we impose the following assumptions to identify the estimands of interest based on selection-on-observables. First, the conditional independence assumption (CIA): $Y_i(1), Y_i(0) \perp D_i | X_i = x$, and second, common support (CS): $0 < P[D_i = 1 | X_i = x] < 1$.

The first assumption, the CIA, states that the potential outcomes are independent of the treatment assignment conditional on the confounders. This implies that all variables affecting both, treatment assignment and outcome, are observed and contained in X_i. The second assumption, CS, requires that treatment possibilities are bounded away from 0 and 1. This assumption is testable and of no issue in this work. (see Figure 2.5 in Appendix 2.B.1)

At this point, we would like to highlight the two different roles of covariates. First, covariates (or confounders; X_i) are required to make the CIA credible. In other words, the characteristics that are responsible for selection into the treatment status have to be accounted for to obtain an 'as-good-as-randomized' situation. This is necessary to obtain causal treatment effects, which are free from selection bias. Second, covariates (or heterogeneity variables; Z_i) are used to form groups of observations for which heterogeneous effects are to be estimated.

In this application, the non-random determination of the treatment, i.e. starting the first leg, is a result of the shootout to select the starter as already discussed in Section 2.3.1. The consequence of this shootout is arguably mainly influenced by the ability of the players and their experience with this situation. As indicated in Section 2.3.2, there are two measures of ability, the previous ranking and the performance in form of the previous average scores achieved. Experience is measured by the cumulated number of matches played, age, number of years of playing darts and the number of years played at a professional level. Furthermore, we control for other

variables potentially influencing the selection into treatment as described in Section 2.3. Since it is unclear in which (functional) form the potential confounders are most relevant to account for the selection into treatment, a flexible estimation technique, namely, a double machine learning approach with non-parametrically estimated nuisance functions, described in the following section, will be used. Having this set of observed potential confounders, it is credible that the CIA is satisfied.

2.4.3 Methods

If one considers using a classical linear regression approach to estimate a treatment effect, there are some implicit assumptions to think about. As already discussed, the conditional independence assumption is crucial. Further, there are the assumptions of a constant treatment effect and a linear effect of the confounders X_i on Y_i. In most studies, there are neither scientific nor methodological reasons to motivate these latter assumptions. Therefore, adopting a methodology not imposing those assumptions is desirable. The upcoming *Causal Machine Learning* literature is agnostic about these assumptions.

The growing literature offers several solutions to estimate treatment effects in a flexible way. For a summary and simulation study covering many of the methods, refer to Knaus, Lechner, and Strittmatter (2020a). They found four causal machine learning methods with good performance in all their simulated settings, while other methods were unstable or did not perform well for estimating the CATE and ATE. Specifically, those are LASSO with covariate modification and efficiency augmentation (Tian, Alizadeh, Gentles, and Tibshirani, 2014), LASSO with R-Learning (Nie and Wager, 2017), Causal Forest with local centring (Athey, Tibshirani, and Wager, 2019; Lechner, 2018), and Random Forest with Double Machine Learning (Chernozhukov et al., 2018). For this work we chose to use Double Machine Learning, for which we have theoretical guarantees required for the strategy of our application, e.g. for the ATE (Chernozhukov et al., 2018) or the CATE (Semenova and Chernozhukov, 2017; Zimmert and Lechner, 2019). To check the sensitivity of the results being method dependent we chose to additionally perform the analysis using a Modified Causal Forest (Lechner, 2018).[12]

[12] An advantage is the easy implementation of cluster-based sampling, which is helpful in case observations are clustered. The potential drawback of the Modified Causal Forest method is efficiency loss owing to the requirement of sample splitting for weight-based inference, which can be circumvented by cross-fitting in the Double Machine Learning approach. Moreover, it works

In the following sections, we build on the observation of the two different roles of covariates. First, the procedure to account for selection into treatment by controlling for (potential) confounding factors is introduced in Section 2.4.3.1. Second, two ways how heterogeneity variables are used to investigate granular treatment effects are described in Section 2.4.3.2.

2.4.3.1 Double machine learning

The first stage of the estimation procedure uses the so-called *Double Machine Learning* (DML), introduced by Chernozhukov et al. (2018). This approach to overcome the selection problem builds on the augmented inverse probability weighting procedure, going back to Robins, Rotnitzky, and Zhao (1994, 1995):

$$Y_i^* = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i\big(Y_i - \mu_1(X_i)\big)}{p(X_i)} - \frac{(1 - D_i)\big(Y_i - \mu_0(X_i)\big)}{1 - p(X_i)},$$

and involves three nuisance parameters. $\mu_1(X_i) = E(Y_i|D_i = 1, X_i)$, modelling the conditional outcome mean if treated, $\mu_0(X_i) = E(Y_i|D_i = 0, X_i)$, the conditional outcome mean if not treated and $p(X_i) = E(D_i|X_i)$, the conditional probability to be treated.[13] This results in the orthogonal score (Y_i^*), which has various applications. The expected value of the orthogonal scores, i.e. $E(Y_i^*)$, is the average treatment effect. The lowest level of aggregation is obtained as $E(Y_i^*|X_i = x)$, while group average treatment effects can be obtained by the expected value conditional on the groups defined in Z as $E(Y_i^*|Z_i = z)$, which are discussed in detail in Section 2.4.3.2. The three nuisance parameters in general can be estimated by any well-suited estimation technique.[14]

To overcome the linearity assumption of effects by the confounders on the outcome, the non-linear and non-parametric Random Forest is used to estimate the nuisance parameters. The Random Forest algorithm, developed in Breiman (2001), is

---

with only discrete or discretized heterogeneity variables. For a more detailed description the interested reader is referred to Lechner (2018) in general, as well as the appendix for a short introduction and the process of its implementation in this work.

[13] Following Chernozhukov et al. (2018), the nuisance parameters are estimated based on a 2-fold cross-fitting, in which first the sample is randomly divided into 2 equally sized folds. In each of the folds, the prediction model is estimated, while in the respective hold-out fold, the model is used to predict the nuisance parameters. While no observation is used to predict its own nuisance, we can use all observations without running into problems of overfitting.

[14] There are requirements on the rate of convergence of the estimators (depending on level of aggregation, cross-fitting, etc.). Those requirements are met for standard regression models, such as OLS and Probit, as well as for specific methods coming from the machine learning literature, such as Random Forest or LASSO and others. In this work, the Random Forest is used, for which the required rates are given (compare Wager and Walther, 2015).

built as an ensemble of single Regression Trees, which are to some extent randomly constructed.[15] Each Regression Tree recursively splits the space of covariates into non-overlapping areas to minimize the MSE of the outcome prediction until it reaches some stopping criteria. The resulting structure is reminiscent of a rotated tree; one can observe the trunk gradually splitting up into finer branches. The averages of the outcomes of those observations falling into the same end-nodes (so called leaves) provide the predictions of the tree. Combining several of those tree predictions results in the final predictions of the Random Forest.[16] This DML step is to remove any (potential) confounding issues. The resulting orthogonal score is free from selection effects, and treatment effects can be constructed at various levels of aggregation.[17]

2.4.3.2 Conditional average treatment effects

As already mentioned, the scores can be used to directly obtain the average treatment effect by $\theta = E(Y_i^*)$. To estimate effects beneath the average level, the obtained orthogonal scores from the DML procedure can be used in different ways. Consider first using the Best Linear Predictor (BLP) framework proposed in Semenova and Chernozhukov (2017). Here, the orthogonal scores are used as pseudo-outcome in an ordinary least squares regression on covariates to solve the minimization problem: $\hat{\beta} = argmin \sum_{i=1}^{N} (\hat{Y}_i^* - \tilde{x}_i\beta)^2$, with $\tilde{x}_i$ containing a constant and x_i. The fitted values, $\hat{\theta}(\tilde{x}_i) = \tilde{x}_i\hat{\beta}$ are the best linear predictors of the IATEs. In fact, $\tilde{x}_i$ can be replaced by any subset of the covariate space. For example, replacing it by only a constant leads to the ATE; replacing $\tilde{x}_i$ by $\tilde{z}_i = (constant, z_i)$ leads to the GATEs. Standard errors can be computed as heteroscedasticity robust standard errors, which are valid, as shown in Semenova and Chernozhukov (2017).[18] Once the IATEs are estimated, they can be analysed in various ways. One possibility is to evaluate the distribution of the effects, to observe how different the effects are for the range of observations. Chernozhukov, Fernández-Val, and Luo (2018) suggested the Sorted Effects method, which sorts the individualized effects according to the effect size, coming with valid confidence intervals. Furthermore, this can be used to analyse whether the most and least affected individuals differ according to their characteristics.

---

[15] For each tree 50% of the data are subsampled. To further de-correlate the trees at each split, only a subset of the available covariates is randomly chosen.

[16] All the necessary tuning parameters are determined in a data-driven way using cross-validation. For more details on Random Forests, the interested reader can refer to the initial study by Breiman (2001) or the book of Hastie, Tibshirani, and Friedman (2009).

[17] For a more detailed discussion, the interested reader can refer to Knaus (2020).

[18] In a sensitivity check, we account for potential clustering on the individual contestant level.

The discussed method has its strength in analysing the total range of individualized effects and compactly summarising heterogeneity. The drawback is that an assumption of linearity of effects is imposed, as one runs quickly into the 'curse of dimensionality' with more than few heterogeneity variables when using non-parametric methods. To investigate specific hypotheses, involving few dimensions of the covariate space, Zimmert and Lechner (2019), as well as Fan, Hsu, Lieli, and Zhang (2019), proposed a non-parametric approach to not be dependent on the linearity assumption of the best linear predictor. The already estimated orthogonal scores can be used with (few) selected heterogeneity variables in a classical non-parametric kernel regression as follows:

$$\hat{\theta}(z) = \sum_{i=1}^{N} \frac{\mathcal{K}_h(z_i - z)\hat{Y}_i^*}{\sum_{i=1}^{N} \mathcal{K}_h(z_i - z)},$$

with $\mathcal{K}_h$ being a kernel function with bandwidth h, determined by cross validation and 90% undersmoothing, as suggested in Zimmert and Lechner (2019). The drawback of this procedure is that the dimension of Z_i is limited to obtain the required asymptotic guarantees. Z_i, in this case, includes one or two heterogeneity variables, which is unproblematic from a theoretical perspective and sufficient to analyse the hypotheses of interest in this work.

2.5 Results

2.5.1 Average treatment effects

The first result of the analysis is the average effect of the BIA for the starting player on winning the match. Column (1) in Table 2.2 shows the average treatment effect. The effect of the technical advantage with 0.0865 is large and precise enough to be statistically different from zero. The average effect for the starting contestants, therefore, amounts to 8.65 percentage points higher winning probability for the match. This effect is in line with what we would expect owing to the technical advantage implicit in the contest design.

Table 2.2: Base results - best linear predictors

	ATE	Home	Country of birth	All
	(1)	(2)	(3)	(4)
Intercept	0.0865***	0.0939***	0.0853***	0.0143
	(0.0089)	(0.0093)	(0.0112)	(0.2406)
Home		-0.0885***		-0.0987***
		(0.0323)		(0.0347)
Venue in country of birth			0.0032	0.0248
			(0.0184)	(0.0199)
Prize money standardized				-0.0736*
				(0.0488)
Ranking				0.0000
				(0.0000)
3 darts average				0.0009
				(0.0026)
Acc. Matches				-0.0001
				(0.0000)
Years playing				-0.0016
				(0.0016)
Years playing professional				-0.0026
				(0.0024)
Age				0.0020
				(0.0016)
Best of legs				0.0011
				(0.0020)
Observations	11,604	11,604	11,604	11,604

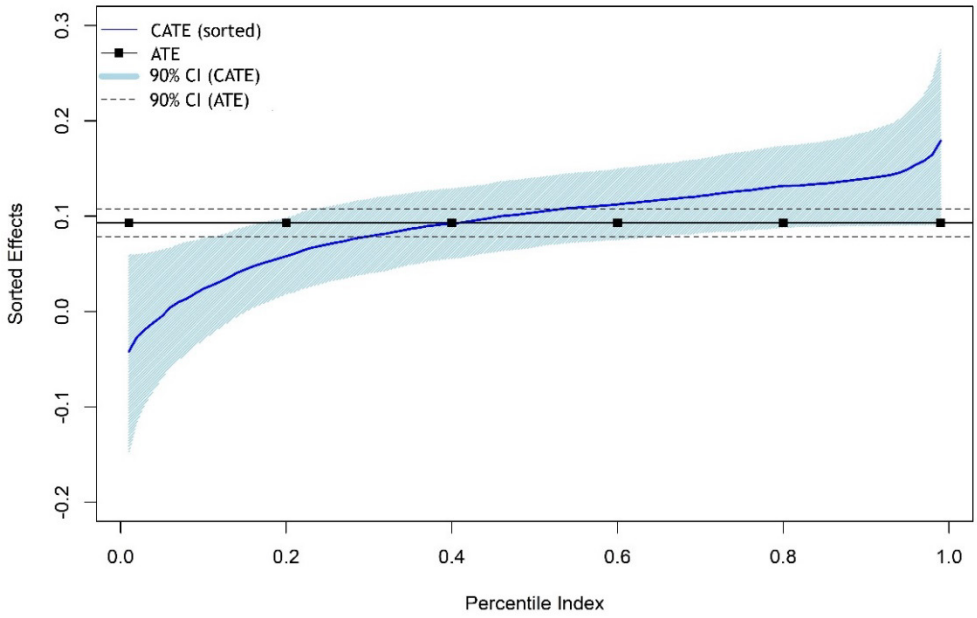
Notes: Best linear prediction as described in Semenova and Chernozhukov (2017); heteroscedasticity robust standard error in parentheses. *Home* is defined as the hometown of the contestant being in the radius of 100km to the venue. *, **, and *** represent statistical significance at the 10, 5, and 1% level, respectively.

While this is a sizeable effect, it would be interesting to note the kind of players that may profit more or less from this advantage. To analyse this, we look into more granular effects in the following sections.

2.5.2 Individualized treatment effects

Starting on the most granular level, we see in Figure 2.1 the individualized average treatment effects for the BIA sorted in size. The solid black line represents the average treatment effect with the dotted black lines being its confidence intervals. The solid blue line represents the sorted individualized effects, accompanied by the shaded blue confidence intervals.

Figure 2.1: Sorted Effects



Notes: Sorted Effects. 999 (weighted) bootstrap replications. Bias corrected. The blue line represents the sorted CATEs accompanied by the shaded 90% confidence interval. The black line with squares represents the ATE, accompanied by the dashed line representing the 90% confidence interval.

While most of the players have positive effects, there are also negative point estimates for specific players. For those players, the technical advantage might not be an actual advantage, even though this is not backed by statistical confidence. In total the individualized effects range from around –5 to +15 percentage points, suggesting that there is some heterogeneity in the effect.

A comparison of the 10% most affected (highest treatment effect) to the 10% least affected (lowest treatment effect) individuals in selected characteristics can be found in Table 2.3. This helps to evaluate the differences in specific characteristics for those with the highest and lowest treatment effects. The estimate represents the difference in the characteristics of the most and least affected. Joint p-values in contrast to the 'usual' p-values account for testing the estimates of, in this case, 10 tests (for more details the interested reader is referred to Chernozhukov, Fernández-Val, and Luo, 2018).

Table 2.3: Difference in characteristics of 10% most and least affected

Characteristics	Estimate	Joint p-value	p-value
Ranking	360.89 (109.66)	0.05	0.00
3 darts average	-5.54 (1.79)	0.07	0.00
Acc. matches	-488.31 (96.48)	0.00	0.00
Years playing	-13.13 (3.78)	0.03	0.00
Years professional	-11.34 (2.38)	0.00	0.00
Age	-6.87 (4.05)	0.66	0.05
Venue in country of birth	0.05 (0.21)	1.00	0.40
Home	-0.49 (0.13)	0.02	0.00
Prize money (standardized)	-0.28 (0.09)	0.09	0.00
Best of legs	-2.54 (2.62)	0.98	0.17

Notes: Bias corrected, 999 (weighted) bootstrap replications. Joint p-values account for the simultaneous inference conducted for all (ten) differences of variables. The estimate is the difference in the characteristics for the most affected (highest) minus the least affected (lowest treatment effect).

In general, the most affected, i.e. those with the highest realization of the technical advantage, have lower performance measures, less experience, competing away from home, and in tournaments with lower prize money, compared to the least affected. A clearly insignificant estimate can be found for *age*, which is arguably the worst proxy available for experience. Furthermore, only *home* as defined as the hometown being within a radius of 100km to the tournament venue seems to be differently allocated, while being born in the country (*Venue in country of birth*) of the venue is almost equally frequent among most and least affected.

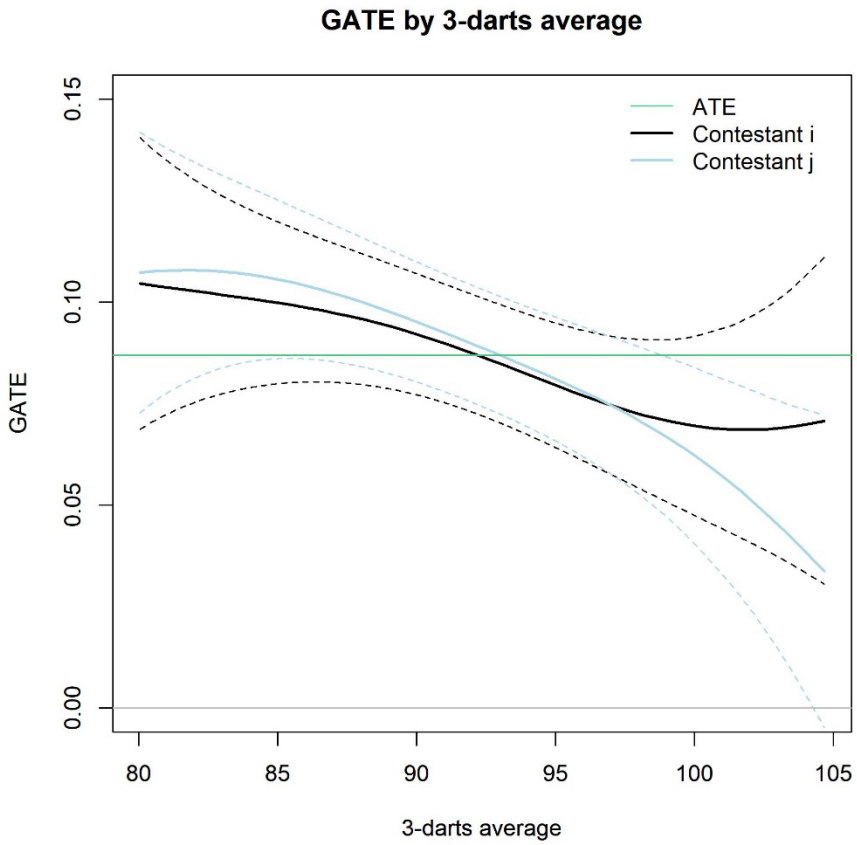
2.5.3 Group average treatment effects

Now we turn to investigating more specific effect heterogeneities. Two group effects are investigated: First, in Section 2.5.3.1, we analyse if there are differential effects in the BIA associated with the ability of the contestants. Second, we evaluate in Section 2.5.3.2 whether, motivated by the social pressure hypothesis, there are differential effects associated to competitors performing near their hometown.

2.5.3.1 Ability

A contest is fair if equally skilled contestants have equal winning probabilities. To discover if the BIA differs for equally skilled contestants on any level of ability, we therefore analysed the BIA of the first moving player (contestant *i*) associated with both competitors' ability (of *i* and *j*). As already discussed, our preferred measure of ability is the performance measure 3-darts average.

Figure 2.2: Built-in advantage of starting contestant by ability



Notes: The broken lines represent the 90% confidence intervals. Contestant *i* is the starting contestant, *j* the opponent.

In general, Figure 2.2 shows the GATEs to be smaller for higher ability athletes, in line with the results in Table 2.3. However, both contestants' GATEs are close to each other for every value of the 3-darts average. In other words, there is no evidence that the BIAs are different for equally skilled contestants. Note that, especially for low and high 3-darts average, there are fewer observations, resulting in larger confidence intervals.

2.5.3.2 Social incentives

We observe in Table 2.3 that those with the highest treatment effects are rather competing away from home, with an estimated difference of -0.49, compared to those with the lowest treatment effects. Investigating this more specifically, the group average effects are shown in Table 2.2. Column (2) presents the effect associated with competing at home (-0.0885) relative to those performing not at home. The technical advantage therefore, exists only for the group of individuals not performing at home (0.0939), and close to zero for those performing at home. The diminished winning probability of about 8.8 percentage points for the group of contestants performing in a friendly environment with support, i.e. at home, is in line with the social pressure and

contrary to the social support hypothesis, as found in Dohmen (2008) and Harb-Wu and Krumer (2019).[19] No substantial difference is found for the group of competitors born in the same country as the venues' country (0.0032, column (3) in Table 2.2). This indicates that the differential effect associated with competing at home can be attributed to social pressure from the audience, while there is no difference in the effect associated with the audience cheering for their compatriots.

Table 2.4: Social pressure vs. home (dis-)advantage

	ATE	Home (*i*)	Home (*j*)	Home (*i* & *j*)
	(1)	(2)	(3)	(4)
Intercept	0.0865***	0.0939***	0.0852***	0.0926***
	(0.0089)	(0.0093)	(0.0093)	(0.0097)
Home (contestant *i*)		-0.0885***		-0.0887***
		(0.0323)		(0.0323)
Home (contestant *j*)			0.0144	0.0168
			(0.0323)	(0.0323)
Observations	11,604	11,604	11,604	11,604

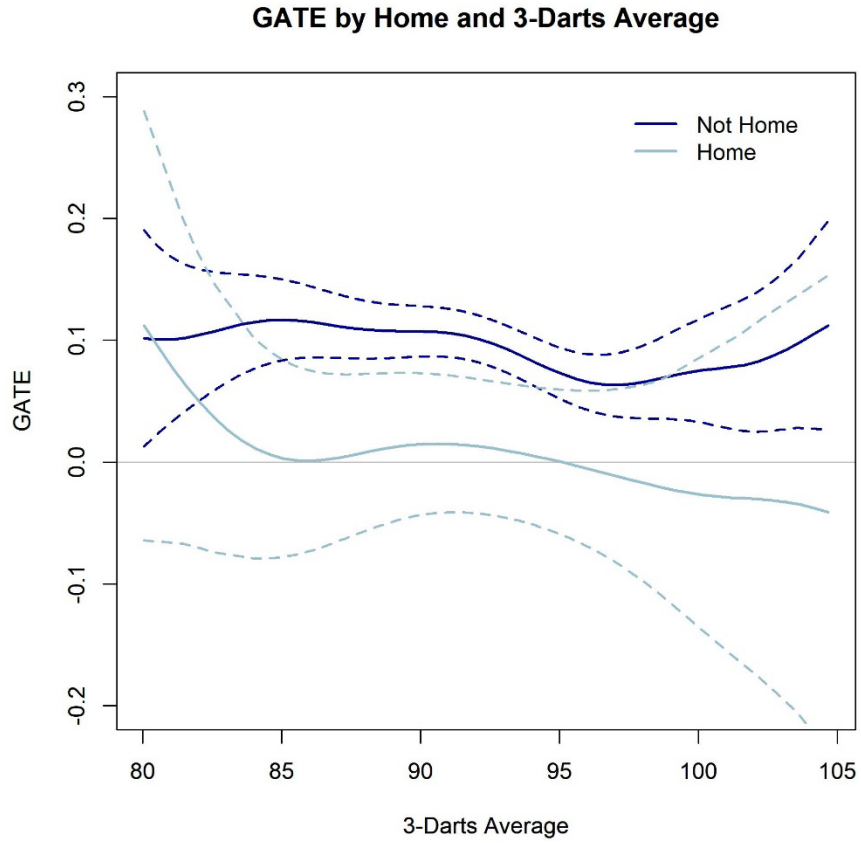
Notes: Best linear prediction as described in Semenova and Chernozhukov (2017). Heteroscedasticity robust standard error in parentheses. *Home* is defined as the hometown of the contestant being in the radius of 100km to the venue. Contestant *i* is the starting contestant, *j* is the opponent. *, **, and *** represent statistical significance at the 10, 5, and 1% level, respectively.

Interestingly, there is no differential effect for the second moving contestant (*j*) to be affected by social pressure or any type of home (dis-)advantage, as seen in Table 2.4. Columns (1) and (2) are equivalent to those presented in Table 2.2; columns (3) and (4) show the differential effect for non-starting contestants associated to competing in a venue near their hometown. In other words, competing in a friendly environment has no (negative) effect on the BIA for non-starting contestants. Therefore, there is no evidence for a general home (dis-)advantage, but evidence in line with social pressure for those starting the match.

On the contrary, an analogy with the finding of Harb-Wu and Krumer (2019), that the effect is driven by those athletes in the top quartile of the ability distribution, cannot be confirmed.

[19] Changing the definition of home in terms of a different radius around the venue can be found in Table 2.A.2 in Appendix 2.B.3. For a narrower definition of *home* contestants, the associated effect is larger, while the associated effect decreases if defining contestants' farther away from the venue as *home* contestants. This supports the general conclusion outlined previously.

Figure 2.3: Differential built-in advantage by home and ability



Notes: The dark blue line represents the GATEs for contestants not home, the light blue line for contestants performing at home. The broken lines represent the 90% confidence intervals.

Figure 2.3 displays the treatment effect for the starter associated with performing at home or not at home with respect to the 3-darts average, the most accurate measure of individual ability in darts. We discovered that the effect associated with playing at home fluctuates around the zero effect, while the effect for those playing away from home is always above. Especially for the boundaries of the figure, i.e. low and high 3-darts average, there are less observations; therefore, the estimates become imprecise. Replacing the 3-darts average by the previous position of the player in the ranking leads to similar conclusions; the results can be found in Figure 2.6 in Appendix 2.B.2.

2.5.4 Sensitivity checks

To assess the sensitivity of the results being method dependent we chose to additionally perform the analysis using a Modified Causal Forest (Lechner, 2018). For more details on the method and implementation the interested reader can refer to Appendix 2.D.

Table 2.5: Average and group average effects, MCF

Effects	Estimate	Standard error	t-value	p-value
ATE	0.0838	0.0147	5.70	0.00
GATE (home=0)	0.0911	0.0151	6.03	0.00
GATE (home=1)	0.0038	0.0357	0.11	0.91
Difference	0.0873	0.0353	2.47	0.01
With sample-based clustering				
ATE	0.0872	0.0152	5.74	0.00
GATE (home=0)	0.0931	0.0155	6.01	0.00
GATE (home=1)	0.0225	0.0365	0.62	0.54
Difference	0.0706	0.0363	1.94	0.05

Notes: Average Treatment Effect (ATE) and Group Average Treatment Effects (GATE), associated with performing in a venue near the hometown (home=1) and not near the hometown (home=0). *Home* is defined as performing at a tournament venue that is within a radius of 100km to the hometown of the athlete. Estimated using a Modified Causal Forest. Inference is conducted as weight-based inference. In the lower part of the table, effects are estimated using sample-based clustering on the individual player level; 1,000 Trees with 50% subsampling rate and the minimum leave size is each tree equals 2.

All estimated effects are in a comparable range to the estimates in the previous sections. In addition, the conclusions drawn hold in general. Table 2.5 shows the ATE estimate, as well as the GATE associated with performing at home resulting from the Modified Causal Forest estimation. Further, in the lower part of the table, we repeated the estimation but accounted for contestant specific clusters. For both estimations, we found a large differential effect associated to performing at home.

For the ability measure, 3-darts average, the pattern is in line with the findings from Section 2.5.3.1. In Figure 2.7 in Appendix 2.B.4 we observe lower GATEs for higher 3-darts averages of the starting contestant (i) and the opponent (j). Especially, for equal levels of ability the GATEs are similar. We can therefore conclude that the BIA is symmetric for equally skilled contestants. Note that, since the 3-darts average is discretized to form groups of roughly equal numbers of observation in each group, we can see that especially for low and high 3-darts averages there are less observations. In conclusion, this conceptually different method produces similar results, providing some evidence that the results presented in the previous sections are robust.

Table 2.6: Base results - best linear predictors, sensitivity to clustering

	ATE	Home	Country of birth	All
	(1)	(2)	(3)	(4)
Intercept	0.0875***	0.0948***	0.0863***	-0.0104
	(0.0095)	(0.0100)	(0.0120)	(0.2316)
Home		-0.0878***		-0.0968***
		(0.0338)		(0.0359)
Venue in country of birth			0.0030	0.0238
			(0.0214)	(0.0225)
Prize money standardized				-0.0936*
				(0.0488)
Ranking				0.0000
				(0.0001)
3 darts average				0.0011
				(0.0025)
Acc. Matches				-0.0001
				(0.0001)
Years playing				-0.0016
				(0.0016)
Years playing professional				-0.0024
				(0.0019)
Age				0.0019
				(0.0016)
Best of legs				0.0011
				(0.0018)
Observations	11,604	11,604	11,604	11,604

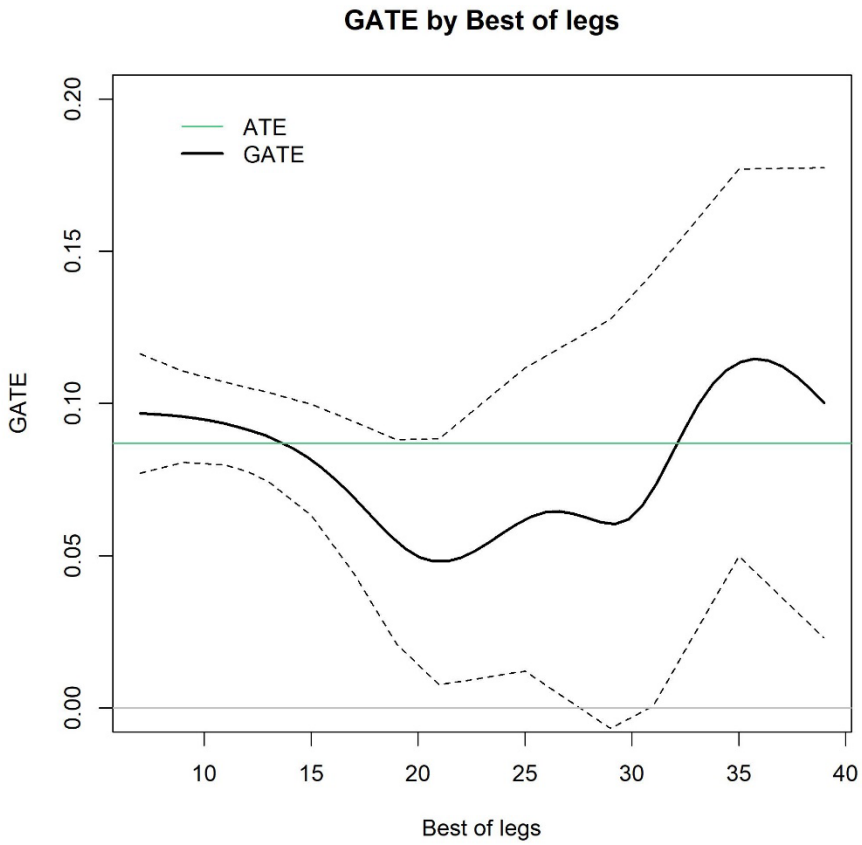
Notes: Best linear prediction as described in Semenova and Chernozhukov (2017); cluster-based sampling and clustered standard errors on the individual level in parentheses. *Home* is defined as the hometown of the contestant being in the radius of 100km to the venue. *, **, and *** represent statistical significance at the 10, 5, and 1% level, respectively.

Accounting for potential contestant specific clusters, the best linear prediction results from Table 2.2 are replicated with cluster-based sampling for the nuisance estimation, as well as clustered standard errors on the individual level in Table 2.6. While standard errors are slightly higher, we find the conclusions drawn previously are valid.

2.5.5 Discussion

A contest design in which one group systematically realizes differential BIAs is unfair. Especially when home contestants are affected, the competition design must be reconsidered and changed to have more fair and interesting contests.

Figure 2.4: BIA by maximum number of legs



Notes: The GATE is shown as solid black line; the broken lines represent the 90% confidence intervals. The green line represents the ATE.

A potential idea is to increase the length of the contest, such that the BIA decreases. While that looks like a neat idea in theory, we cannot provide clear evidence for it. Figure 2.4 provides the BIA associated to the number of maximum legs to play in a match. The GATE fluctuates around the average effect, with no clear tendency towards lower or no BIA for longer matches.

A second option is to remove the technical advantage from the contest design in form of even numbers of maximum legs to play, as for example, in tennis tournaments.[20] Further, Cohen-Zada, Krumer, and Shapir (2018) found that the order of serves in tennis tiebreak, which is the ABBA sequence, does not provide any advantage to any of the players. Adopting a similar structure might help to improve the fairness in darts contests.

[20] Another potential idea to resolve the BIA is to give the respective second-mover the chance to catch-up in case the starter of the leg finishes the leg.

2.6 Conclusion

The study discovered an average effect of about 8.6 percentage points higher winning probability induced by the BIA in the sequential darts contest with two players. Having a randomized allocation of being the first-mover for equally skilled contestants, the contest is fair for contestants with symmetric potential BIAs. While we found non-differential BIAs for equally skilled contestants, the personalized point estimates suggested substantial heterogeneity in the effect, ranging from about -5 to +15 percentage points with athletes having more experience and better performance measures to benefit least. In addition, estimated differential effects were in line with the social pressure hypothesis as found in previous works, especially in the behavioural economics literature. For the group of individuals competing in a friendly environment, the BIA was lower compared to those competing on neutral grounds, which was in line with the works of Dohmen (2008) and Harb-Wu and Krumer (2019) among others. This leads to not necessarily equal winning probabilities for contestants with equal abilities, and therefore, to an unfair contest.

For designing contests, the heterogeneous nature of the BIA in such types of contests should be considered to not unintentionally favour specific groups of contestants. Especially if fairness is a precondition for a contest, the presented findings should be taken into account. Despite fairness issues, local athletes might be important for the attractiveness of the contest. Introducing a fairer contest design may, therefore, improve the attractiveness as well. We did not find any evidence that an increase in the maximum length for a match leads to diminished BIAs, so the suggestion would be to change the design of the contest, such that there is no BIA for any contestant. Furthermore, the findings are of interest to individuals performing in situations of increased importance, either contestants in tournaments or in daily life situations. While preparing for those situations, one should put more emphasis on mental training.

Another insight from this study is that *Causal Machine Learning* is highly valuable for empirical studies. It enables us to provide a more complete picture of particular effects under investigation, as well as a way to investigate specific hypotheses in a flexible way. Furthermore, the two steps of removing potential selection bias and estimating treatment effects offers an intuitive, flexible, and robust way to improve credibility of empirical research.

Finally, we call for more, fine granular investigations in contest designs, particularly in situations in which there are potentially multiple incentives observable

(and distinguishable). Since we only have data on male professional players, we encourage further studies to analyse how the BIA affects women, as well as youth and amateur players in darts or other asymmetric contests. Furthermore, in future studies, the limitations of audience presence owing to the current corona pandemic could be used to analyse the BIA, without the direct influence coming from the audience. Methods from the *Causal Machine Learning* toolbox appear to be especially useful for this and should be considered for further empirical analyses.

Appendices

Appendix 2.A: Descriptive statistics

Table 2.7: Descriptive statistics, all variables

Variable	Mean	Starting Player	Non-Starting Player
Game outcomes			
Win match	0.50	0.55	0.45
Win first leg	0.50	0.63	0.37
Game characteristics			
Starts first leg	0.50	1.00	0.00
Best of leg	11.21 (3.43)		
Televised	0.32		
Ranking tournament	0.90		
BDO tournament	0.04		
Double in	0.01		
Prize money	51,145 (79,918)		
Prize money (standardized)	0.13 (0.25)		
Player characteristics			
3 darts average	91.82 (4.53)	91.99 (4.49)	91.65 (4.56)
Ranking	121.74 (292.77)	113.00 (254.99)	130.47 (326.20)
Accumulated matches	231.15 (270.17)	238.13 (274.46)	224.17 (265.81)
Left handed	0.07	0.07	0.07
Years playing	19.97 (9.82)	20.07 (9.79)	19.87 (9.85)
Years playing professional	11.91 (6.21)	11.90 (6.17)	11.91 (6.25)
Age	36.54 (9.73)	36.46 (9.63)	36.61 (9.83)
Venue in country in birth	0.38	0.38	0.38
Home (50km radius)	0.04	0.03	0.04
Home (100km radius)	0.08	0.08	0.08
Home (150km radius)	0.15	0.15	0.15
Home (200km radius)	0.21	0.22	0.21
Distance to venue	1,410.25 (3,271.76)	1,422.84 (3,307.26)	1,397.66 (3,241.93)

Table 2.7, continued

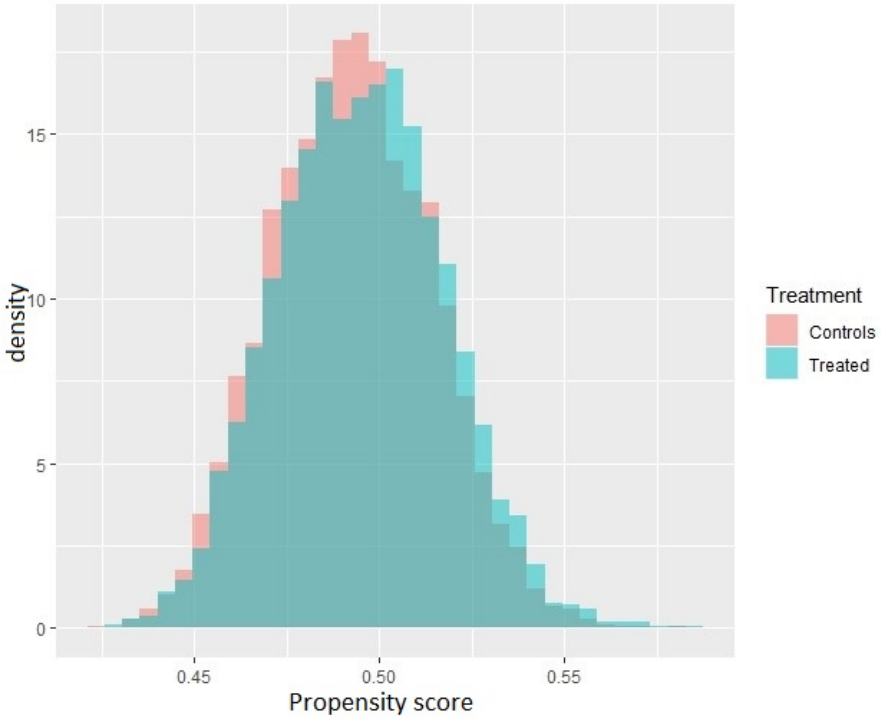
Variable	Mean	Starting Player	Non-Starting Player
Diff. 3 darts average		0.34 (5.76)	-0.34 (5.76)
Diff. ranking		-17.46 (406.21)	17.46 (406.21)
Diff. accumulated matches		13.96 (355.44)	-13.96 (355.44)
Diff. years playing		0.20 (13.89)	-0.20 (13.89)
Diff. years playing professional		-0.01 (8.76)	0.01 (8.76)
Diff. age		-0.15 (13.76)	0.15 (13.76)
Diff. distance to venue		25.18 (4,309.09)	-25.18 (4,309.09)
Year			
2009	0.01		
2010	0.02		
2011	0.02		
2012	0.02		
2013	0.02		
2014	0.02		
2015	0.03		
2016	0.07		
2017	0.07		
2018	0.34		
2019	0.37		
Observations	11,604	11,604	11,604

Notes: This table presents average values and standard deviations (in parentheses for non-binary variables). Values that are equal for all columns are reported only once.

Appendix 2.B: Additional results

Appendix 2.B.1: Common support assumption

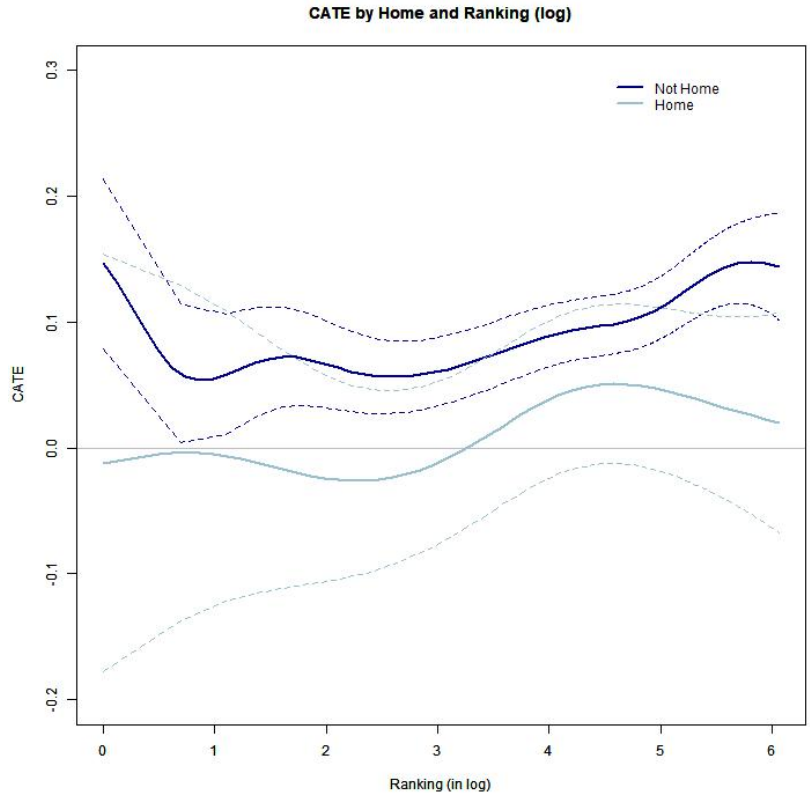
Figure 2.5: Propensity scores



Notes: Shown are the propensity scores for controls (in red) and treated (in blue). On the horizontal line you find the value of the propensity score.

Appendix 2.B.2: Alternative ability measure

Figure 2.6: Differential BIA by ability, alternative specification



Notes: The dark blue line represents the GATEs for contestants not home, the light blue line for contestants performing at home. The broken lines represent the 90% confidence intervals.

Appendix 2.B.3: Alternative definitions of home

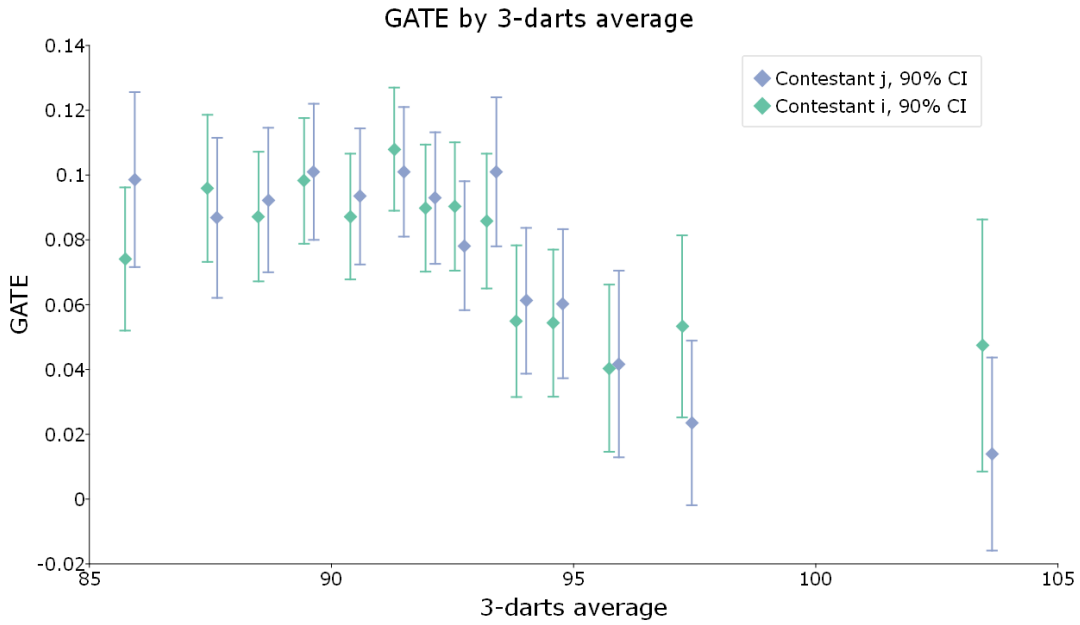
Table 2.8: Different definitions of 'Home'

	50 km	100 km	150 km	200 km	Distance (in km)
	(1)	(2)	(3)	(4)	(5)
Constant	0.0921***	0.0939***	0.0962***	0.0947***	0.0877***
	(0.0091)	(0.0093)	(0.0097)	(0.0100)	(0.0097)
Home	-0.1533***	-0.0885***	-0.0664***	-0.0393*	-0.0000
	(0.0471)	(0.0323)	(0.0249)	(0.0217)	(0.0000)

Notes: Best linear prediction as described in Semenova and Chernozhukov (2017). Heteroscedasticity robust standard error in parentheses. *, **, and *** represent statistical significance at the 10, 5, and 1% level, respectively. Different columns represent different definitions of the *Home* variable, for which the hometown of the contestant is within 50 km (column 1), 100 km (2), 150 km (3), 200km (4), and the continuous measure for the distance in km (5), to the venue.

Appendix 2.B.4: Sensitivity checks

Figure 2.7: Built-in advantage of starting contestant by ability; MCF



Notes: Contestant *i* is the starting player depicted in blue, accompanied by the 90% confidence intervals. Contestant *j* is the opponent depicted in green. Standard errors are calculated as weight-based standard errors.

Appendix 2.C: Contest model

We model a match in darts by adapting the lottery contest structure from Tullock (1980) contests, which is the major workhorse for modelling contests (for a survey on ways of modelling contests, see e.g., Nitzan, 1994). The probability to win for contestant i starting is: $Pr_{i,Starter}^{Win} = \frac{A_i}{A_i+\delta_i A_j}$ and for contestant j being the starter: $Pr_{j,Starter}^{Win} = \frac{A_j}{A_j+\delta_j A_i}$. We denote the ability of contestants i and j by A_i and A_j. A potential BIA for the contestants is denoted by $0 < \delta_i, \delta_j \leq 1$.[21] Symmetrically, following $Pr_{j,Non-Starter}^{Win} = 1 - Pr_{i,Starter}^{Win}$ and vice versa for contestant i, we get: $Pr_{i,Non-Starter}^{Win} = \frac{A_i}{A_i+{}^1/_{\delta_j} A_j}$ and $Pr_{j,Non-Starter}^{Win} = \frac{A_j}{A_j+{}^1/_{\delta_i} A_i}$. With the shootout before a match, there is a mechanism to allocate the right to start the contest, denoted by

[21] If $\delta_i = \delta_j = 1$, there is no BIA in the contest. In contrast to the 'classical' Tullock lottery contest the presented model is more flexible. We neither assume that there is no BIA nor that they are necessarily equal for both contestants.

$\pi(A_i, A_j)$ depending on the ability of contestants *i* and *j*. The ex-ante probability to win the contest for *i* is:

$$Pr_i^{Win} = \pi\big(A_i, A_j\big) \times \frac{A_i}{A_i + \delta_i A_j} + (1 - \pi\big(A_i, A_j\big)) \times \frac{A_i}{A_i + {}^{1}/_{\delta_j} A_j} \tag{2.1}$$

$$= \frac{A_i A_j\big(1+\delta_i\delta_j\big)+2\delta_j A_i^2}{2(A_i+\delta_i A_j)(\delta_j A_i + A_j)} + (1 - 2\pi\big(A_i, A_j\big)) \times \frac{A_i A_j(\delta_i\delta_j - 1)}{2(A_i+\delta_i A_j)(\delta_j A_i + A_j)} \tag{2.2}$$

$$= \frac{\big(1+\delta_i\delta_j\big)+2\delta_j}{2(1+\delta_i)(\delta_j+1)} + (1 - 2\pi) \times \frac{(\delta_i\delta_j - 1)}{2(1+\delta_i)(\delta_j+1)} \tag{2.3}$$

A contest is regarded as fair if contestants with equal abilities have equal winning probabilities. Therefore, from equation (2.2) to (2.3) contestants with equal ability, i.e. $A_i = A_j = 1$, are assumed.[22] From equation (2.3), it can be concluded that the contest provides equal winning probabilities for equally skilled contestants, i.e. 50 % each, in one of the two cases: 1.) there is no BIA ($\delta_i = \delta_j = 1$), or 2.) the treatment allocation is randomized ($\pi = 0.5$), and the BIA is equal for both contestants ($\delta_i = \delta_j$). The first case of no BIA can be tested empirically. For the second case, while treatment allocation can be assumed to be randomized for equally skilled contestants, the size and nature of the BIA are subject to an empirical analysis, which is the focus of Section 2.5.

Appendix 2.D: Modified causal forest

In addition to the DML procedure, we used a conceptually different method from the class of estimators, which also achieved good overall performance in the study of Knaus et al. (2020a), a variant of Causal Forests. More specifically, we employed the *Modified Causal Forest* (MCF) proposed in Lechner (2018).

The MCF offers a single procedure to estimate the various levels of aggregations, IATEs, GATEs, and ATE, coming with a unified inference approach. Further, it has an easy way of implementing sample-based clustering, which can be applied to account for contestant-specific clustering.

Generally, a causal forest estimator is built as an ensemble of many different *Causal Trees*, going back to Athey and Imbens (2016). Each *Causal Tree* splits the sample according to specific values of covariates sequentially into finer strata, so called leaves, until some stopping criterion is reached. The splits are conducted in a way to

[22] Note that, for ease of exposition, for contestants with equal abilities $\pi\big(A_i, A_j\big) = \pi$.

maximize the variance of treatment effects. This, in turn, mitigates selection effects and potentially uncovers effect heterogeneity in one step. Within final leaves, treatment effects are computed as the difference of the mean of the outcomes of treated and control units. Building many such trees on subsampled data and taking the average over the trees' effect estimates are the basis of causal forest estimators (Wager and Athey, 2018; Athey, Tibshirani, and Wager, 2019). Lechner (2018) refined this algorithm by implementing an improved splitting rule.

A potential drawback of this approach is a possible efficiency loss as the weight-based inference approach requires sample splitting. Moreover, in comparison to the Nonparametric CATE estimation discussed in the main part of this work, variables used in the analysis of effect heterogeneity needs to be discrete. For further details on the estimator, the reader can refer to Lechner (2018).

Appendix 2.E: Tournaments

Table 2.9: List of tournaments

Tournaments	Ranking Tournament*	Federation
BDO World Darts Championship	Yes	BDO
BDO World Master	Yes	BDO
BDO World Trophy	Yes	BDO
Champions League of Darts	No	PDC
European Championship	Yes	PDC
European Tour	Yes	PDC
Finder / Zuiderduin Masters	Yes	BDO
Grand Slam of Darts	No (2009-2014) Yes (2015-2019)	PDC / BDO
PDC World Darts Championship	Yes	PDC
Premier League Darts	No	PDC
Player Championship	Yes	PDC
The Masters	No	PDC
UK Open	Yes	PDC
World Cup of Darts	No	PDC
World Grand Prix	Yes	PDC
World Matchplay	Yes	PDC
World Series of Darts	No	PDC

Notes: *Ranking tournaments are open for qualification. Non-Ranking tournaments are invite-only tournaments. Money won in ranking tournaments count as *Order of Merit* ranking points. For more details, see: https://www.pdc.tv/pdc-order-merit-rules (PDC) and http://www.bdodarts.com/images/bdo-content/doc-lib/F/bdo-invitation-rules.pdf (BDO).

Appendix 2.F: List of sources

Dartsforwindows (software; from www.dartsforwin.com)

www.pdc.tv

www.bdodarts.com

www.dartsdatabase.co.uk

www.wikipedia.org

www.mastercaller.com/players

www.dartswdf.com/rules

3 Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed.

Daniel Goller, Michael Lechner, Andreas Moczall, and Joachim Wolff

Abstract: Matching-type estimators using the propensity score are the major workhorse in active labour market policy evaluation. This work investigates if machine learning algorithms for estimating the propensity score lead to more credible estimation of average treatment effects on the treated using a radius matching framework. Considering two popular methods, the results are ambiguous: We find that using LASSO based logit models to estimate the propensity score delivers more credible results than conventional methods in small and medium sized high dimensional datasets. However, the usage of Random Forests to estimate the propensity score may lead to a deterioration of the performance in situations with a low treatment share. The application reveals a positive effect of the training programme on days in employment for long-term unemployed. While the choice of the "first stage" is highly relevant for settings with low number of observations and few treated, machine learning and conventional estimation becomes more similar in larger samples and higher treatment shares.

Keywords: Programme evaluation, active labour market policy, causal machine learning, treatment effects, radius matching, propensity score

JEL classification: J68, C21

3.1 Introduction

A long and ongoing literature is concerned with the evaluation of active labour market programmes (ALMP) in a selection-on-observables setting. Propensity score (PS) based matching-type estimators are the established econometric workhorse in this literature (e.g., Imbens, 2004, 2015; Smith and Todd, 2005; Wunsch and Lechner, 2008; Lechner and Wunsch, 2009, 2013; Biewen, Fitzenberger, Osikominu, and Paul, 2014; Doerr, Fitzenberger, Kruppe, Paul, and Strittmatter, 2017; Caliendo, Mahlstedt, and Mitnik, 2017; Calónico and Smith, 2017; the meta study of Card, Kluve, and Weber, 2018, and references therein). A common issue in PS-based methods is the concrete specification of the PS. The past and current literature usually estimated the PS using a parametric model, i.e. Probit or Logit. Covariates and functional forms were commonly chosen in a fairly ad-hoc manner based on monitoring the balancing properties of the resulting estimated PS (compare Rosenbaum and Rubin, 1984; Dehejia and Wahba, 2002).

The emerging literature in statistical learning, might help to make this specification less ad-hoc.[23] In this paper, we investigate if machine learning methods can improve average treatment effect on the treated (ATET) estimation when used to predict the PS. Estimating the PS used in matching-type estimators with machine learning could help in three ways: 1) detecting variables of the selection process that might otherwise be omitted by the researchers, but are available in the data; 2) allowing for the appropriate degree of functional flexibility in the PS; 3) increasing the precision of the estimate by avoiding overfitting of the PS. These issues become more relevant with the increased availability of rich-covariate "big data" datasets, the handling of which requires suitable methods.

Although off-the-shelf machine learning methods have many well-documented advantages in prediction and classification, it is not obvious that using them for propensity score estimation in a matching framework will improve the estimation of causal effects. One potential reason is that they aim at a different target (compare Athey and Imbens, 2019). The goal of using a PS in matching estimation is to balance the covariate distribution of treated and non-treated units to obtain a quasi-experimental situation. Machine learning algorithms, if used for PS estimation, follow the goal to

[23] For an overview of statistical learning methods, see e.g. Hastie, Tibshirani, and Friedman (2009). Throughout the work, we use the terms statistical learning and machine learning interchangeably as if they were synonyms.

predict treatment participation given the covariates, as good as possible, by trading off bias and variance in out-of-sample comparisons. One example for why this may be a bad idea are covariates that are very good predictors of outcome but only weakly correlated with treatment assignment (compare e.g. Belloni, Chernozhukov, and Hansen, 2014). Since machine learning algorithms try to maximize predictive power (in a mean square error sense), they may omit such variables as they do not help much to predict the treatment, accepting a somewhat larger bias in the propensity score that is dominated by the resulting variance reduction. However, since these now-omitted variables are important predictors of the outcome, the small bias in the propensity score may translate to a large one in the ATET estimation.

While there are already some implementations of the idea of estimating the PS used in matching-type estimators with machine learning procedures (e.g. Krumer and Lechner, 2018; Goller and Krumer, 2020) there is little evidence on whether such an estimator actually has favourable finite sample properties. In early papers, Setoguchi, Schneeweiss, Brookhart, Glynn, and Cook (2008) and Lee, Lessler, and Stuart (2010) investigated the performance of machine learning methods for estimating the PS. Those papers based their simulations on a data generating process (DGP) which might be well suited for their targeted applications in the medical context. They found machine learning predictions to outperform the parametric baseline methods. Pirracchio, Petersen, and van der Laan (2015), and Cannas and Arpino (2019) used the same specifications as the two above-mentioned papers and found the Super Learner (van der Laan, Polley, and Hubbard, 2007) and the Random Forest, respectively, to perform best, while the other machine learning techniques did not work sufficiently well in terms of bias in PS matching. As these four studies used the same data generating process based on only ten covariates and a treatment share of 50 percent, they might be less informative for microeconometric applications in which the dimension of confounders is usually much higher and the treatment shares very likely to deviate from 50 percent.

In another recent work, Brown, Merrigan, and Royer (2018) evaluated machine learning PS estimation techniques in a simulation study. In particular, they found that Least Absolute Shrinkage and Selection Operator (LASSO), Boosting and Deep Learning outperformed the Random Forest and the baseline approach in terms of bias in their simulations. While they based the simulations on a high-dimensional empirical dataset with a low share of treated, this is only partially related to our question as they focus on using the PS as covariate in a Cox Proportional Hazard Model.

Hill, Weiss, and Zhai (2011) investigated a high-dimensional empirical problem and discussed strategies and challenges to understand which PS method to use. They illustrated the various potential strategies and the resulting wide range of different estimates, highly depending on the choice of the empirical researcher. As they did not observe the true effect, they were not able to point out which strategies worked best in their setup.

In conclusion, there is only limited practical advice from the existing literature on how to improve PS estimation with the goal of 'better' treatment effect estimation. Thus, our work contributes to the literature in evaluating the performance of classical and machine learning based PS estimators for matching-type estimators in a realistic labour market setting.

To be as close as possible to a real situation empirical researchers might face, we use a rich administrative dataset of German long-term unemployed persons in an Empirical Monte Carlo Simulation (EMCS), as suggested by Huber, Lechner, and Wunsch (2013) and Lechner and Wunsch (2013). Furthermore, we compare the different estimators in a real programme evaluation application.

Our database consists of a large sample of German unemployed means-tested benefit recipients at the end of 2009, most of them long-term unemployed, including all individuals participating in a specific training programme in the first quarter of 2010. There is a broad range of characteristics recorded for each individual, which includes all the quantifiable information relevant for the case-workers' decision to send the respective individual to a training programme or not.

We evaluate the effect of a training programme and simulate the performance of different PS estimators, using the radius matching on the propensity score with bias adjustment (RMBA) algorithm developed in Lechner, Miquel, and Wunsch (2011), which performed best in the simulation of Huber, Lechner, and Wunsch (2013). To be more precise, we use two different machine learning techniques, namely Random Forest and LASSO to estimate the PS. We choose these two as they use very different approaches in a non-parametric sense: Random Forest approximate the PS locally, similar to non-parametric regression, while LASSO with many polynomials and interaction term will approximate the PS with a flexible global function, e.g. similar to

series estimation.[24] In that sense, they represent two very different types of approaches. A large literature discusses both methods, establishing theoretical properties (e.g. Hastie, Tibshirani, and Friedman, 2009), as well as modifying them for usage in other types of causal inference problems (e.g. Belloni, Chernozhukov, and Hansen, 2014; Wager and Athey, 2018; Lechner, 2018; Athey, Tibshirani, and Wager, 2019). We compare these two estimators to the true, a random, and a PS based on a parametric ad-hoc (Probit) model, which we then use for estimating the ATET in the RMBA estimator.

Our findings are mixed. LASSO performs well as PS estimator for the usage in radius matching especially in situations in which using Probit and Random Forest do not deliver credible estimates. When there are many covariates compared to observations, Probit does not work well; once the number of observations increases sufficiently, Probit and LASSO perform equally well. Random Forest tends to predict the treatment in sample well, but does not work properly as balancing score estimator. If the share of treated units is low, the Random Forest cannot manage to split deep enough to estimate a PS flexible enough to remove the selection bias. In fact, we find that PS estimated with Random Forest may lead to comparing control units and treated units, which are not sufficiently similar. Thus, whether using specific off-the-shelf machine learning algorithms does help depends on the context of the application. Since knowing which of the methods works a priori appears to be difficult, a plausible alternative is to use Causal Machine Learning methods instead, e.g. 'double machine learning' suggested by Chernozhukov et al. (2018) or the Modified Causal Forest suggested by Lechner (2018), that are optimized specifically for treatment effect estimation (for an overview see e.g. Knaus, Lechner, and Strittmatter, 2020a).

The empirical application that we conducted reflects the sensitivity to method choice. While all methods lead to a positive effect of the training programme, the effects based on PS estimated by Random Forests are about 30 percent larger compared to the estimates using LASSO or Probit as PS estimator.

The structure of the rest of the paper is as follows: In Sections 3.2 and 3.3, we describe the institutional background and the database used for the simulation and

[24] Other potential competitors might be the global estimators Ridge or Elastic Net, as well as the local estimators Boosted Trees, deep or shallow Neural Nets or combinations of those. In case good predictors for the PS lead to better results it might be worth investigating several other competitors coming from the respective class. However, the results below show that this is not the case, though.

application in detail. Section 3.4 introduces the EMCS, as well as the estimators used. Sections 3.5 and 3.6 present the results of the simulations and the empirical application. Section 3.7 concludes. Additional results can be found in the Appendices.

3.2 Institutional background

We analyse these methodological questions with regard to the effects of a German short-term training programme named Determining, Reducing and Removing Employment Impediments (DRR). It is a sub-programme of the Schemes for Activation and Integration (SAI) that consist of different training programmes as well as placement services by private providers.[25]

The SAI programmes, introduced in 2009, replaced a number of earlier programmes with similar basic objectives. They differed from its predecessors in providing greater flexibility to local service providers to better suit their services to the particular needs of different unemployed persons. While there are many sub-programmes within SAI differing in their target groups and detailed goals, we focus only on the "Determining, Reducing and Removing Employment Impediments" sub-programme in order to analyse a rather homogeneous treatment type. The DRR sub-programme focuses on finding out which particular attributes define the individual's disadvantage, improving participants' skills, and providing them with knowledge about suitable occupational fields and individual opportunities on the labour market. The target group is both unemployment insurance and unemployed welfare recipients. The latter are usually long-term unemployed with some prospects of labour market integration. Among the various types of Schemes for Activation and Integration, the relative importance of the "Determining, Reducing and Removing Employment Impediments" sub-programme is considerable. It represents 15 percent of the 428,000 persons entering any type of Schemes for Activation and Integration (SAI) programme in our observation period January to March 2010.[26,27] Due to the flexible programme design, there is no programme duration defined a priori; the average duration is slightly less than two months.

[25] German name of DRR: Feststellung, Verringerung, Beseitigung von Vermittlungshemmnissen; German name of SAI: Maßnahmen zur Aktivierung und beruflichen Eingliederung.

[26] Source: Department of Statistics of the German Federal Employment Agency – Labour Market Programme Statistics.

[27] The inflow of 428,000 people includes both unemployment insurance and unemployment welfare recipients. Our analysis will only consider the unemployed welfare recipients, because the means-tested nature of these benefits results in richer data being available on these individuals, which in turns increases the likelihood that the identifying assumption is fulfilled.

3.3 Data

3.3.1 Dataset

We use a large and rich dataset that not only consists of detailed characteristics on individuals, their labour market history and household situation, but also on the staff structure of the job centres responsible for them.

The data on individuals are based on employer reports to the German social security administration as well as internal records of job centres and labour agencies. They contain socio-demographic characteristics, information on the last job, and almost complete employment and unemployment histories.[28] Moreover, these data include welfare benefit receipt, welfare benefit sanctions, ALMP participation, household composition and income information. The variables are available for the unemployed welfare recipients themselves as well as for their partners.

We augment this dataset with characteristics of the local labour market. They include the unemployment rate, the long-term-unemployment rate, the vacancy-unemployment ratio, the number of registered unemployed people and of unemployment benefit II recipients, and the inflow into various active labour market programmes. Finally, we add information on the staff structure of the job centres. Job centre employee data is available as full-time equivalents. An important piece of information in this context is the average number of welfare recipients for which a job centre employee is responsible. One might interpret it as a measure of the intensity of activation. However, it could also be related to the composition of welfare recipients. If a relatively large share of welfare recipients is characterised by strong or multiple placement impediments, the job centre staff has to invest more time per welfare recipient to achieve a placement. That might be reflected by a relatively low number of welfare recipients per job centre employee. Other available measures in this context are, e.g. the gender distribution of job centre employees, the distribution of contract types, e.g. fixed-term versus open ended or employee versus civil servant, the presence of equal opportunity officers, and the wage distribution among the job centre employees.

[28] The employment data contain periods of marginal employment and employment subject to social security contributions. Periods of self-employment and civil servant employment periods are not represented in our data.

3.3.2 Treatment and sample selection

Our sample design is similar to the one used by Harrer, Moczall, and Wolff (2020), which analysed the effectiveness of the entire SAI. Our treatment group consists of the total inflow from January to March 2010 into the "Determining, Reducing and Removing Employment Impediments" sub-type of the SAI who were unemployed and receiving means-tested benefits on December 31st, 2009. The control group represents a 20 percent random sample of persons likewise unemployed and receiving means-tested welfare benefits on December 31st, 2009, who did not enter any SAI programme from January to March 2010 but may have entered other programme types.

For data quality reasons, we restrict the sample to individuals administered jointly by the Federal Employment Agency and municipalities.[29] Moreover, we only include individuals aged 25 to 55 who are not disabled. For younger welfare recipients, various special rules and group-specific programmes exist so that they are subject to more intense activation than older welfare recipients are. Finally, we dismissed observations from our sample due to missing or obviously wrong values in some of the variables. The remaining final sample of 276,637 observations is analysed in the application in Section 3.6 and our EMCS described in Section 3.4.

3.3.3 Descriptive statistics

Our sample consists of 14,817 treatment group and 261,820 control group observations. For brevity, we only present descriptive statistics of selected variables. Complete descriptive tables for all the covariates are available upon request. The selected variables reflect the aspects covered by the variable groups that in Lechner and Wunsch (2013) were found to be sufficient to remove most biases.

[29] Some job centres are run by municipalities only. Data on unemployment benefit II recipients from these job centres were partly incomplete in particular in the years 2005 and 2006. Therefore, these data are not suitable to construct some of the covariates on past labour market history for our analysis. Moreover, for them no information is available about the full-time equivalents and composition of the job centre staff. Therefore, individuals from these job centres, who represent less than 13 per cent of the unemployed unemployment benefit II recipients in the year 2009, are not included in our analysis.

Table 3.1: Descriptive statistics of selected covariates

Variable	Treated		Controls	
	Mean	SD	Mean	SD
Cumulated duration in regular employment 3-36 months after treatment (Outcome)	218	(314)	162	(282)
Female	0.44		0.46	
Age at sampling date in years	38	(8)	40	(9)
Receives some income from employment (<15h/week)	0.17		0.23	
Cumulated number of days in welfare receipt in year before sampling date	312	(100)	325	(86)
Participated in Schemes by Providers[1)]	0.14		0.06	
Participated in classroom training	0.61		0.45	
Job centre district: Inflow into Schemes by Providers[1)] relative to jobseeker stock in 2009q4	0.02	(0.01)	0.01	(0.01)
Job centre district: Inflow into In-Firm Training[2)] relative to jobseeker stock in 2009q4	0.004	(0.002)	0.004	(0.002)
Days since last employment[3)]	1,904	(1,877)	2,262	(2,045)
Cumulated days in regular employment in five years before sampling date	230	(372)	183	(334)
Secondary schooling degree: None	0.12		0.15	
Secondary schooling degree: Lower	0.47		0.44	
Secondary schooling degree: Intermediate	0.29		0.28	
Secondary schooling degree: University of applied science qualification	0.04		0.04	
Secondary schooling degree: University qualification (A-level equivalent)	0.07		0.08	
Vocational degree: None	0.50		0.51	
Vocational degree: Non-college	0.47		0.44	
Vocational degree: College	0.03		0.04	
Family status: Single	0.42		0.38	
Family status: Married	0.26		0.30	
Family status: Divorced/widowed	0.23		0.25	
Family status: Cohabiting	0.08		0.08	

Notes: SD: standard deviation (only reported for non-binary variables). Descriptive statistics of full set of covariates available upon request. 1) Schemes by Provides are those programmes among the SAI, which are organised by private providers like private placement services or classroom training. 2) In-Firm Training are those programmes among the SAI, which are organised as internships in firm. 3) Mean and SD calculated only for persons who had a last job. Sources: Integrated Employment Biographies and other administrative datasets available at the Institute for Employment Research, regional data of the Statistics Department of the German Federal Employment Agency, own calculations.

However, in contrast to the sample studied in Lechner and Wunsch (2013) our sample consists to a far higher extent of people who did not work for various years.

Therefore, we included in more detail covariates on the labour market history of the last five years. As Table 3.1 shows, treatment and control units, with 218 versus 162 days in regular employment in a three-year period after treatment, differ in terms of our outcome variable of interest. There are also considerable differences in pre-treatment characteristics.

Examples are the days since last employment with 1,904 versus 2,262 days of people who previously were employed, and the cumulated number of days in regular employment in the previous five years at 230 compared with 183 days. This shows that persons with more recent labour market experience are somewhat more likely to participate in DRR. There are no great differences in terms of sex or age. Most striking is the observation that 61 percent of treatment group versus 45 percent of control group individuals had participated in a classroom-training-type programme before. "Classroom training" in this context refers to non-in-firm trainings before the 2009 reform that introduced the SAI programme.

The mean values of education and family status and partner characteristics included in Table 3.1 in most cases do not differ remarkably between treated and control individuals. Nevertheless, these descriptive statistics show that selection into treatment is non-random with respect to some variables. The rest of this paper is therefore concerned with modelling selection on these observable characteristics based on our extensive set of potential confounders.

3.4 Methodology

3.4.1 Target, notation and identification

In the following, we will use the notation for treatment effects estimation using the potential outcome framework of Rubin (1974). Participation in a training programme, as discussed in Section 3.3.2, is indicated with D_i as the binary treatment variable, while $D_i = 1$ indicates that individual i (i=1,…,N) takes part in a training programme and $D_i = 0$, otherwise. The outcome variable Y_i denotes accumulated days in employment of individual i three years after the treatment. Let $Y_i^d \coloneqq Y_i(D_i = d)$ denote the potential outcome if individual i receives treatment $d \in \{0,1\}$.[30] Since each individual can only receive either treatment or non-treatment one potential outcome is

[30] Throughout the work, random variables are indicated by capital letters and realizations of these random variables by lowercase letters.

observable, the other remains counterfactual: $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$. While this implies that individual treatment effects are not directly observable, imposing assumptions may make it possible to identify treatment effects at various aggregation levels, e.g. the average treatment effect (ATE): $\tau = E(Y_i^1 - Y_i^0)$. The focus of this work is on the ATET, i.e. $\theta = E(Y_i^1 - Y_i^0 | D_i = 1)$.

Further, we investigate situations in which treatment assignment is non-randomly determined and empirical researchers opt for a selection-on-observables approach using a matching-type estimator. This is an attractive approach in situations in which all important confounders are arguably available as covariates, denoted by X_i. Confounders are those characteristics jointly affecting selection into treatment as well as potential outcomes. Controlling for those confounding factors lead to potential outcomes, which are independent of the treatment.

In many applications, this set of control variables might be large, like in our empirical setup, leading to a curse of dimensionality in matching-type estimators. Rosenbaum and Rubin (1983) showed the equivalence of conditioning on all *X* and on a one-dimensional balancing score, the so-called propensity score (PS), defined as $p(x) = P[D_i = 1 | X_i = x]$. Matching-type estimators commonly exploit this equivalence. As described in Rubin (2007), the resulting estimator consists of two stages. First, estimate the PS. Second, use this estimated score to compare treated with similar non-treated units.

Throughout we use the following four identifying assumptions, which are standard in the selection-on-observables literature:

A.1: $Y_i^1, Y_i^0 \perp D_i | X_i = x, \quad \forall\, x \in \mathcal{X}$, conditional independence assumption

A.2: $0 < P[D_i = 1 | X_i = x] = p(x) < 1$, common support

A.3: $X_i^0 = X_i^1$, exogeneity of covariates

A.4: $Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i)$, stable unit treatment value assumption

A.1 might be relaxed to $Y_i^0 \perp D_i | X_i = x$ for the case of ATET estimation. This assumption ensures that all confounders are observed and rules out the existence of further (unobserved) confounders jointly influencing the treatment and the potential outcome under non-treatment conditional on the observed *X*, or in this case conditional on the PS. A.2 ensures common support by bounding the treatment probability away from 0 and 1, and can also be relaxed in ATET estimation to $p(x) < 1$. The two latter

assumptions require that covariates are not affected by the treatment (A.3) and that there are no spill over effects between the treatment groups (A.4). Under A.1-A.4, we have:

$$\theta = E[Y_i^1 | D_i = 1] - E[Y_i^1 | D_i = 0]$$

$$= E[Y_i | D_i = 1] - E[E[Y_i | D_i = 0, p(x)] | D_i = 1]$$

Which means that we can identify the (causal) ATET by comparing units in treatment and non-treatment that are comparable with respect to their PS.

3.4.2 Empirical Monte Carlo Simulation

Knowing the true answers of an empirical question is usually not possible. For this reason, evaluation studies tend to do simulation studies in which the researcher specifies the DGP, and therefore all dimensions of the true DGP are known. The drawback of those kinds of studies is that artificially created datasets might not capture the relationships of real applications.

To be as close as possible to applications in the empirical research literature, Huber, Lechner, and Wunsch (2013), and Lechner and Wunsch (2013) developed a so-called Empirical Monte Carlo Study (EMCS). The idea is to use a DGP that exploits the structure of an empirical dataset to its full extent. For example, outcomes and covariates of real data are used. Of course, there are limitations, since the researcher needs to control some features to allow for generalizations, like the sample size or the share of treated in our case. Further, the empirical dataset must be large enough to plausibly presume that the random samples come from an infinite population. This is the case for our data as described in Section 3.3, which is a typical large-scale administrative dataset.

Table 3.2: Empirical Monte Carlo Study

1) The PS is estimated in the full data. The true score is constructed as a combination of the separately estimated scores using the Probit, LASSO and Random Forest as:
$$\hat{p}^{true}(x) = \frac{1}{3}(\hat{p}_{Probit}(x) + \hat{p}_{LASSO}(x) + \hat{p}_{RandomForest}(x))$$
2) Remove all the treated observations from the population.[31]
3) Draw a sample of N units from the (remaining) population of control observations and simulate a placebo treatment in this draw, for which the treatment effect is zero by definition, as:
$$d: Bernoulli(\hat{p}^{true}(x) \times \phi),$$
Where $\phi \in \{1,2,5\}$ is to modify the share of (placebo) treated.[32]
4) Estimate the PS in the sample using the different estimation techniques described in Section 3.4.4 and use those respective PS to estimate the ATET using the RMBA estimator described in Section 3.4.3.
5) Repeat steps 3) & 4) *R* times.
6) Calculate performance measures.

Every EMCS used to evaluate a treatment effects model consists of three basic steps. First, a true PS is estimated in the full population.[33] Second, a sample is drawn from the control units, a placebo treatment is simulated according to the true PS and the effects are estimated in this sample. Last, this is repeated many times and the performance is evaluated.

We look at various performance measures, when evaluating the performance. First, the bias is calculated as mean of the deviation from the true effect, i.e. $bias = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\theta_r} - \theta_0)$. $\widehat{\theta_r}$ is the estimated ATET of the matching step in repetition *r,* θ_0 the true effect (which is equal to zero since we discard all treated units). Most important is the mean squared error (MSE) of the ATET, calculated as $MSE = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\theta_r} - \theta_0)^2$. Other measures we look at are the mean absolute deviation or error (MAD), kurtosis, skewness, the mean of the estimation (standard) error in the matching step, as well as the variance of $\widehat{\theta_r}$. Further common support statistics are reported, as the mean share of observations, as well as the mean share of treated observations remaining in the common support. To investigate the performance of the first-stage estimation, we look

[31] As well as all observations with $\hat{p}^{true} > 0.2$ to ensure that the PS after transformation in step 3 are still between 0 and 1. This accounts for less than 1 percent of all control observations.

[32] While $\phi = 2$ leads to share of treated of about 10 percent, $\phi = 5$ to a share of treated of about 25 percent.

[33] Since our goal is to evaluate different PS estimation techniques, we do not want to favour one specific method. Therefore, the 'true' PS is constructed as a combination of the separately estimated PS using the Probit, LASSO and Random Forest.

at how well the various methods do in the PS estimation. Here we report the mean correlation of the estimated with the true PS, as well as the (in-sample) prediction MSE. Since radius matching compares treated and non-treated units, which are close to each other in terms of PS, the correct ordering of the estimated PS is important. We show two statistics for this, namely the (mean of) Kendall's Tau and the (mean of the) Spearman Rank Correlation coefficient.[34]

According to the procedure presented in Table 3.2, we simulated five different scenarios with three different treatment shares and two different sample sizes (see Table 3.3). We use 5, 10 and 25 percent as treatment shares, because the number of treated is usually much smaller than the number of controls in active labour market programme evaluations. Similarly, samples smaller than our minimum sample size of 4,000 observations rarely occur in observational studies in the labour market context. The maximum of 16,000 observations is chosen due to the increasing computational burden of larger samples.

Another parameter to determine in simulations is the number of repetitions, *R*. Ideally, one would like to set this parameter as large as possible to minimize simulation noise. Since this noise depends on the variance of the estimators, which declines with sample size, we repeated each estimation for the smaller sample 1,000 times and the larger sample 250 times. In case of $\sqrt{N}$ -convergence, this will keep the simulation error approximately constant.

Table 3.3: Summary of DGP's

Scenario	Treatment share	Sample size (N)	Repetitions (R)
A	10 %	4,000	1,000
B	25 %	4,000	1,000
C	5 %	16,000	250
D	10 %	16,000	250
E	25 %	16,000	250

In the following sections, we describe the matching estimator used for the ATET estimation as well as the different "first-stage" PS estimation techniques.

[34] Spearman Rank Correlation is defined as: $r_s = 1 - \frac{6\sum(rank(\widehat{p_i}) - rank(p_i^{true}))^2}{n(n^2-1)}$, Kendall's Tau is defined in the following way: $r_K = \frac{2}{n(n-1)}\sum_{i<j} sign(\widehat{p_i} - \widehat{p_j})sign(p_i^{true} - p_j^{true})$.

3.4.3 Matching estimator

While there are several different matching algorithms available, we use the bias-adjusted-radius-matching-on-the-propensity-score estimator (RMBA) of Lechner, Miquel, and Wunsch (2011). This estimator combines the features of distance-weighted radius matching with bias adjustment to remove biases due to mismatches and performed well in Huber, Lechner, and Wunsch (2013). In radius matching control observations, within a specific radius of the treated unit in terms of the propensity score, can be compared to the latter to obtain the treatment effect as if treated and control units were in an experimental setting. Distance weighting implies a weighting of control units in the specific radius inversely proportional to their distance to the respective treated unit.[35]

It has been shown by Lechner and Strittmatter (2019), among others, that trimming treated observations may be important if there is thin or even lacking support to guard against bias and excessive importance of specific control variables. In the setup of this work trimming does not change the ATET, since the true treatment effect is homogenous (and zero) by construction. The trimming rule used follows the recommendation of Lechner and Strittmatter (2019) and removes too important, i.e. control units with a weight larger than 5 percent, and off-support observations jointly for treated and controls.

3.4.4 Propensity score estimation

For the sake of simplicity, we focus on five different approaches to estimate the PS. One benchmark case, which is usually not observed in observational studies, is provided by the true PS. As another benchmark case, we use a non-information PS consisting of i.i.d. random numbers only.

The other three approaches are choices researchers might use in their work, namely a Probit, a Random Forest, and a LASSO-based estimator. While those methods are known to be good prediction techniques there is little knowledge how they perform in empirical labour market evaluation studies for estimating a causal effect in

[35] The radius is determined data-driven as 1.5 times the maximum pair matching distance as suggested by Lechner, Miquel, and Wunsch (2011). A detailed description of the algorithm can be found in Appendix 3.C.

matching estimators. We describe each of the estimation techniques used in the following in more detail, as well as how they are implemented in the EMCS.

3.4.4.1 Probit

Since the PS is the probability of receiving the treatment conditional on the confounders, the Probit estimation, especially in the past, was the usual choice for this first step estimation.[36] $\hat{p}(x) = \Phi(x\hat{\beta})$ is estimated for each individual, where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution. This parametric, non-linear technique is well suited for those kinds of prediction problems if the following four conditions are satisfied. 1) The true selection equation is well approximated by the Probit link function. 2) The set of confounding characteristics and their relevant measurement (i.e. logs, particular polynomials, etc.) is known. 3) The required functional flexibility of the covariates, in particular with respect to interactions of the variables can be well approximated by the researcher. 4) The final set of covariates (incl. all terms that enter the linear index in the Probit link function) is not too large with respect to sample size.

Usually in observational studies ensuring conditions 1) to 3) is subject to a credible line of argumentation, and in most cases, even with a strong intuition, hard to specify correctly. Further, including every variable and functional transformation thereof contradicts the fourth condition in most settings. Too many covariates may decrease the precision of the estimator or may make estimation numerically infeasible.[37]

[36] Similarly, one might choose the Logit estimator, which is omitted here for the sake of brevity.

[37] Too many covariates might not only decrease precision, but also reduce the common support (compare D'Amour et al., 2017) as in-sample predictive power increases.

3.4.4.2 LASSO

The LASSO as proposed by Tibshirani (1996) is a shrinkage estimator, which works like an OLS estimator with penalized coefficients. Since we are estimating a probability-like quantity, we oppose this potential issue of predicting values below 0 or above 1 by using a Logit version of the LASSO.[38] Therefore, the following minimization function is used:

$$\min_{\beta} \{\sum_{i=1}^{N}[-y_i x_i \beta + \log(1 + \exp(x_i \beta))] + \lambda \sum_{j=1}^{k} |\beta_j|\} \qquad (3.1)$$

and the PS is obtained as $\hat{p}(x) = \frac{\exp(x\widehat{\beta})}{1+\exp(x\widehat{\beta})}$.

The last term in equation (3.1) is to penalize the size of the $j = 1, \ldots, k$ coefficients, with k being the number of covariates. λ represents the penalty term. The larger this penalty term, the more the coefficients are pushed towards zero and variable selection takes place, i.e. coefficient become exactly zero. The idea behind this procedure is to shrink the coefficients of those covariates to zero that contain no or little predictive information about the dependent variable.[39]

Determining the size of the penalty term is therefore crucial. This choice represents a trade-off between bias, which λ increases, and the variance, which decreases when λ increases. Here, the penalty term is chosen by 5-fold cross-validation minimizing the out-of-sample mean squared error (MSE). Note that for a specific application it might be beneficial to define a set of indispensable variables, e.g. theoretically relevant variables for selection into treatment, which are excluded from the variables selection to improve the performance of the LASSO PS.

3.4.4.3 Random Forest

In the machine learning literature, the Random Forests algorithm developed by Breiman (2001) is a widely used non-parametric and non-linear estimation technique. It is built as an ensemble of Regression Trees, which are to some extent randomly constructed. A Regression Tree recursively splits the covariate space into separate non-overlapping areas as it minimizes the MSE of the prediction of the outcome. The resulting structure is reminiscent of a rotated tree, as one observes the trunk with all

[38] Compare Hastie, Tibshirani, and Friedman (2009, p. 125).

[39] A 'double-selection' alternative is proposed by Belloni, Chernozhukov, and Hansen (2014), in which additionally variables are captured that are highly correlated to the outcome and mildly related to the treatment selection. To be consistent with the other methods in this work we focus on using the LASSO capturing treatment selection.

the observations in the beginning, which is split into finer branches the further you go down. The tree predictions are the average of the outcome of those observations falling into the same end-nodes, so called leaves.

Like in LASSO, there is a trade-off between bias and variance: Deeply grown trees have lower bias and higher variance compared to shallow trees. This trade-off is controlled by specifying the minimum number of observations in each leaf.[40] For a Random Forest several deep, low-bias trees are estimated on random subsamples and the predictions are averaged over those trees.[41] In our simulations, 600 trees are built for each forest. The more trees are estimated, the smoother the prediction become, but computation time increases.[42] Further, to de-correlate the trees only a random subset of covariates is considered at every split point within the tree building process. [43]

Finally, we use the so-called honest splitting rule, as proposed by Athey and Imbens (2016). Using independent samples for building the tree and for making the predictions contributes to higher prediction accuracy. This comes with the price to pay in terms of reduced sample sizes. As an example, in the N=4,000 setting only 1,000 observations are used to build the tree structures, another 1,000 to do the predictions.[44]

3.4.4.4 Set of covariates

The Methods described above are able to work with different kinds of variables (as described in Section 3.3) in other ways, and therefore the sets of covariates in the PS estimation differ for each method. Probit and LASSO cannot distinguish between ordered and unordered categorical variables. Unordered variables are therefore split into binary variables for each category.[45] This results in 309 covariates for the Probit estimation.

[40] In our simulations, we used a minimum leaf size of five observations.

[41] The random subsamples can be generated by either bootstrapping or subsampling. We follow the recommendation of Wager and Athey (2018) to use subsampling. In the simulations and application, the subsampling size is a share of 50 percent of the sample size.

[42] In pre-tests we found the results not to be sensitive on the number of trees once each forest has more than 500 trees.

[43] In the simulations and application, the number of covariates is chosen to be 50.

[44] Subsampling 50 percent of the sample, as well as using half of it for the tree building and the other half for predicting. Lowering the sample size at first decreasing accuracy, as the variance is higher in smaller samples. Still, this honest split should reduce the bias coming from overfitting.

[45] Examples for which there are no natural ordering are family status, last occupation or nationality.

Since the LASSO has a variable selection property, it is able to solve the objection of including too many covariates up to a certain degree.[46] To be more flexible, we are able to increase the set of potential confounding variables by including second-order polynomials and interactions of all continuous variables resulting in a full set of 1,011 covariates available for the LASSO. Of course, ideally, one would like to include interactions up to a higher degree, to be as flexible as possible, but since the potential set of covariates increases exponentially computational resources are quickly exhausted.

The Random Forest is able to work with unordered categorical variables, while in the other methods dummies are used instead.[47] Further, there is no need to include transformation of variables, like polynomials and interactions in the LASSO, as the tree structures are able to incorporate any interactive and non-linear nature of the covariate structure. Therefore, this method ensures a very large degree of flexibility, as it can, at least asymptotically, pick-up any non-linearity. The set of covariates is therefore substantially lower, i.e. 109 covariates, compared to the other methods. Still, this is only another way to work with the same information and there should be no advantage or disadvantage compared to the other methods. To reduce the computational burden binary variables representing less than 2 percent of the observations are removed in all the methods, as well as we only keep one if there are multiple covariates, which show correlations of more than ± 0.98 in the respective sample.

3.5 Simulation

We evaluate the performance of the various PS methods in the estimation of the ATET using radius matching. For the sake of brevity, summaries of the full results are discussed here, while detailed and additional results tables are presented in Appendices 3.A and 3.B.

Before discussing the results, as this may be an important issue in applied research, we like to point out convergence problems of the Probit estimation in the small sample. We report the results for all repetitions in the main results. Further, we report the results for only converged approaches in the Appendix, as common practice

[46] In fact, increasing the number of covariates also decreases the speed of convergence, which might harm the estimator at some point more than it helps.

[47] For information how this works and how it is implemented see Hastie, Tibshirani, and Friedman, (2009, p.310).

in the literature is rather to modify the specification of the Probit instead of using a non-converged PS in applied work. The results do slightly differ, but the general conclusions are equivalent, however, this points to difficulties in using the Probit in settings with low number of observations and a large set of confounders, especially if the share of treated is low.[48]

Table 3.4: Summary of simulation results, propensity score estimation

	Spearman Rank Correlation					MSE				
	N=4,000		N=16,000			N=4,000		N=16,000		
Treatment	10 %	25 %	5 %	10 %	25 %	10 %	25 %	5 %	10 %	25 %
Probit	0.36	0.60	0.49	0.73	0.87	8.50	17.25	4.62	8.56	16.60
RF	0.72	0.82	0.75	0.81	0.86	8.19	16.72	4.37	8.16	15.92
LASSO	0.77	0.86	0.82	0.87	0.92	8.62	16.58	4.64	8.64	16.56
True	-	-	-	-	-	8.60	16.53	4.64	8.63	16.54
Random	0.00	0.00	0.00	0.00	0.00	9.30	20.90	4.81	9.33	20.90

Notes: Figures shown are the mean of the Spearman Rank Correlation of the estimated PS compared to the true PS, as well as the (in-sample) MSE (times 100) of the prediction over 1,000, respectively 250 simulation repetitions. The full results can be found in Tables 3.10, 3.12, 3.14, 3.16 and 3.18 in the Appendix. True and Random indicates the true, respectively the randomized PS.

To investigate the performance of the PS estimation in Table 3.4 we find the (in-sample) prediction MSEs to show the Random Forest predicting best. More important is the ordering of the PS determining which control units are matched to the respective treated units. The results of the Spearman Rank Correlation with the true PS are depicted in Table 3.4. We find every method to perform better in those settings with higher treatment shares and/or more observations. The Random Forest and the LASSO both reach the highest rank correlations, while the Probit is doing rather poor in the small samples. With more observations, i.e. effectively a lower number of covariates relative to observations, the Probit becomes more competitive and reaches a higher Spearman Rank Correlation compared to the Random Forest in the higher treatment share. This may indicate that the underlying model is well approximated by the Probit functional form. Further, as expected, the random PS obtains values of (close to) zero.

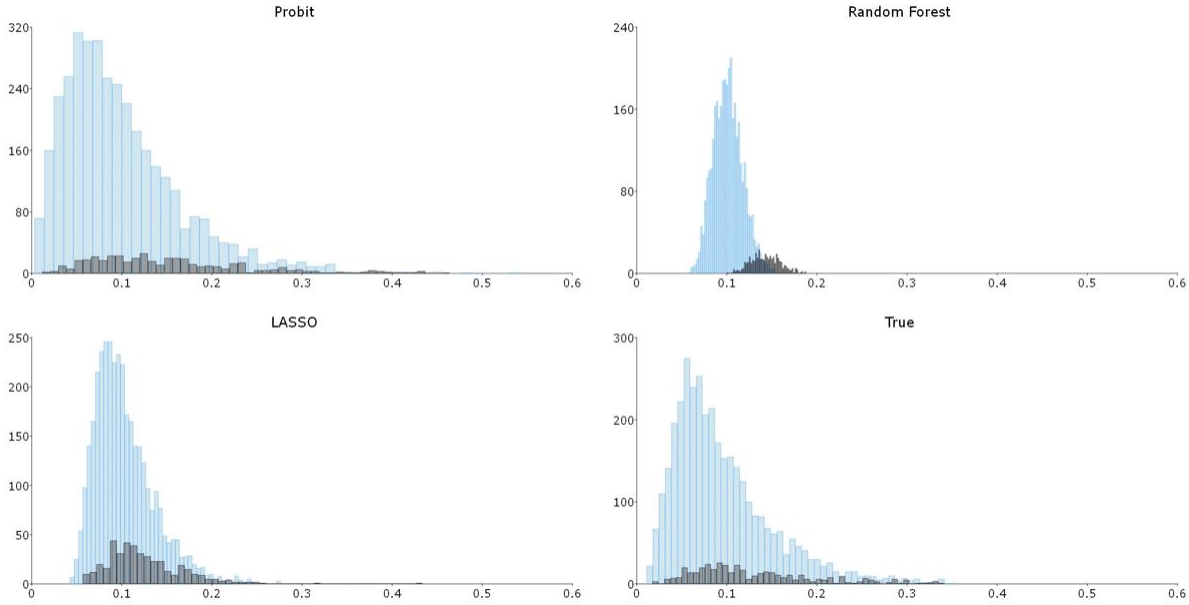
For the Random Forest, having a low treatment share may contribute to splitting less deeply than it should.[49] Therefore having a higher treatment share enables the

[48] For N=4,000 and 10 percent treated about 35 percent of replications, for 25 percent treated about 4 percent of replications did not converge. In the larger samples, this problem is only present for the case of 5 percent treated with about 25 percent of replications not converging. Compare Tables 3.9, 3.11, 3.13, 3.15, and 3.17 in the Appendix.

[49] Having a low share of treated, i.e. a large number of zeros and a low number of ones, in the outcome variables makes it more likely that there cannot be any improvement in terms of MSE by splitting a certain leaf, leading to large final leaves after few

growing of deeper trees, which might be necessary for balancing the covariates in the matching estimator. Figure 3.1 provides some insights into the estimated PS of the respective methods, as well as the true PS for the small sample and low treatment share.[50]

Figure 3.1: Propensity scores by treatment status, N=4,000, 10% treated



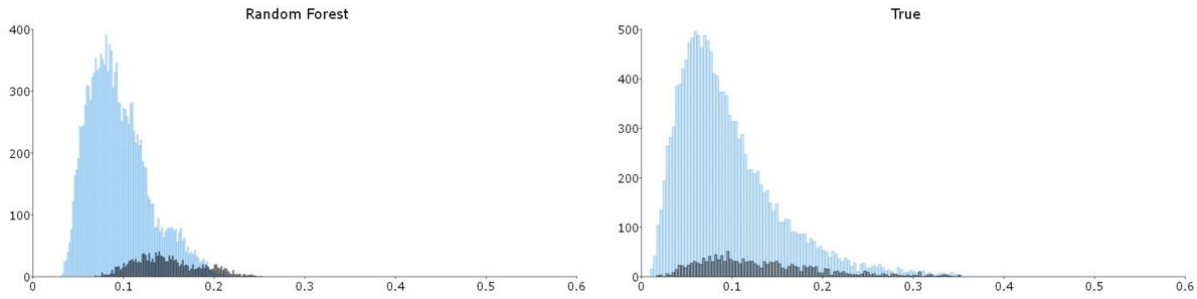
Notes: Histograms with PS on the horizontal axis. Top left is the Probit PS, top right Random Forest, bottom left and right the LASSO estimated and true PS. Each from the same one simulation with N=4,000 and 10% treatment share. Control units are light, treated units dark shaded.

First to note is that the Random Forest in the top right graph looks quite different to the other estimates, as well as the true PS. Not being able to split deep enough leads to a narrower distribution of estimated PS and treated and controls are more separated compared to the other methods. On the one hand, this reduced overlap leads to lower common support. On the other hand, this might lead to matching "wrong" control to the respective treated units. Despite there might be a tendency towards a wider spread of the Random Forest in the larger sample, Figure 3.2 shows generally a similar pattern.

splits. In fact, the average leaf size for the Random Forest is larger in both simulations with the low treatment share compared to the higher treatment share.

[50] Figures for the other simulation scenarios can be found in Appendix 3.B.

Figure 3.2: Propensity scores by treatment status, N=16,000, 10% treated



Notes: Histograms with PS on the horizontal axis. Left is the PS estimated by the Random Forest, right the true PS. Each from the same one simulation with N=16,000 and 10% treatment share. Control units are light, treated units dark shaded. LASSO and Probit PS can be found in Appendix 3.B.2

Table 3.5: Matching-quality

	$q_{0.1}$	$q_{0.3}$	$q_{0.5}$	$q_{0.7}$
Panel A:		N=4,000, 10% treated		
Probit	0.31 (0.05)	0.59 (0.05)	0.76 (0.03)	0.89 (0.02)
Random Forest	0.80 (0.54)	0.91 (0.36)	0.96 (0.22)	0.99 (0.11)
LASSO	0.27 (0.03)	0.55 (0.03)	0.74 (0.02)	0.88 (0.01)
True	0.26 (0.00)	0.54 (0.00)	0.74 (0.00)	0.88 (0.00)
Panel B:		N= 4,000, 25% treated		
Probit	0.27 (0.02)	0.56 (0.04)	0.74 (0.04)	0.88 (0.04)
Random Forest	0.46 (0.17)	0.69 (0.10)	0.83 (0.04)	0.94 (0.03)
LASSO	0.30 (0.02)	0.61 (0.02)	0.79 (0.01)	0.91 (0.01)
True	0.29 (0.00)	0.59 (0.00)	0.79 (0.00)	0.91 (0.00)
Panel C:		N=16,000, 5% treated		
Probit	0.32 (0.08)	0.60 (0.08)	0.77 (0.06)	0.89 (0.04)
Random Forest	0.90 (0.65)	0.96 (0.43)	0.99 (0.26)	1.00 (0.13)
LASSO	0.26 (0.02)	0.54 (0.02)	0.73 (0.02)	0.87 (0.01)
True	0.25 (0.00)	0.53 (0.00)	0.73 (0.00)	0.87 (0.00)
Panel D:		N= 16,000, 10% treated		
Probit	0.28 (0.05)	0.56 (0.06)	0.73 (0.06)	0.85 (0.05)
Random Forest	0.66 (0.40)	0.82 (0.27)	0.91 (0.17)	0.97 (0.10)
LASSO	0.26 (0.01)	0.55 (0.01)	0.74 (0.01)	0.88 (0.01)
True	0.26 (0.00)	0.54 (0.00)	0.74 (0.00)	0.88 (0.00)
Panel E:		N= 16,000, 25% treated		
Probit	0.30 (0.01)	0.60 (0.01)	0.79 (0.00)	0.91 (0.00)
Random Forest	0.47 (0.18)	0.69 (0.09)	0.85 (0.07)	0.94 (0.02)
LASSO	0.30 (0.01)	0.60 (0.01)	0.79 (0.00)	0.91 (0.00)
True	0.29 (0.00)	0.59 (0.00)	0.79 (0.00)	0.91 (0.00)

Notes: This table shows which quantiles of the control samples are matched to the respective quantiles of the treated units. Mean values over all 1,000, respectively 250 repetitions, are reported. Mean absolute deviation to the quantiles of the true PS method are reported in parentheses. q_x stands for the x-percent quantile of the treated.

To investigate this further, we provide matching quality measures in Table 3.5 showing which quantiles of the distribution of control units' PS are matched in the simulation to the respective quantiles of the distribution of the treated PS. As can be seen in Table 3.5 in every panel the Random Forest estimates lead to the most distinct matching of quantiles. This is most pronounced in the scenarios of low treatment shares. Of course, the matching quantiles of the true PS is not necessarily the best, but a valid benchmark. While the LASSO is in most situations the closest to the true PS matching quantiles, the Random Forest is, especially in the 10 percent quantile far away from the true PS results. Despite the "matching-quality" becoming closer to the true PS for the higher quantiles, the estimates of the Random Forest do not seem to work well, especially in low treatment shares, in the context of matching-type estimators.

Table 3.6 shows the observed final performance of the estimated PS in the RMBA estimator. In column (6) of Table 3.6, we observe the absolute mean standardized bias in covariate balancing (SB), which is one rough measure of how well the covariates are balanced using the respective PS estimate.[51] While the balancing ability of the Probit increased considerably in Panels B-E compared to Panel A, the seemingly good Random Forest prediction led to rather worse covariate balancing. For the higher treatment shares the balancing statistics is acceptable. The true and the LASSO PS showed good balancing properties throughout the results.

[51] As there is no clear guidance, commonly used ad-hoc rules suggest that balancing bias should not exceed 20 (e.g. Imbens and Rubin, 2015), or in more restrictive settings 10 (e.g. Normand et al., 2001). Further, despite Cannas and Arpino (2019) found this score to predict the bias of causal estimators well, there are two other reasons why one should not take balancing measures too serious (compare Ho et al., 2007). 1.) The SB only looks at balancing of variables in their baseline form. A good SB might therefore be necessary, but not sufficient for a low bias in the matching step. 2.) There is no distinction between the strength of the confounders. While for the first issue there is up to our knowledge no credible solution proposed in the literature, as the true confounding is unknown. To oppose this second issue, we only look at the ten most important confounders determined as those variables selected in both LASSO procedures, Y on X and D on X, in the full dataset.

Table 3.6: Summary of simulation results, matching

	(1) Bias	(2) MSE	(3) Variance	(4) CS (%)	(5) CS (%), treated	(6) SB
Panel A:			N=4,000, 10% treated			
Probit	21.95	885.52	403.54	92.8	64.7	8.20
RF	-26.15	2,258.05	1,574.16	56.7	90.2	28.18
LASSO	5.03	398.29	372.95	98.1	99.4	5.49
True	-0.39	341.25	341.10	98.3	99.6	5.33
Random	20.47	773.07	353.89	99.6	99.9	16.34
Panel B:			N=4,000, 25% treated			
Probit	11.68	310.55	174.05	98.0	95.6	3.13
RF	-2.18	275.33	270.57	94.2	97.4	9.41
LASSO	3.63	213.93	200.73	98.9	99.1	4.03
True	-0.32	226.28	226.18	99.0	99.0	4.06
Random	24.29	762.48	172.80	99.9	99.9	19.46
Panel C:			N=16,000, 5% treated			
Probit	6.73	271.81	226.47	98.0	75.9	5.41
RF	24.93	2,221.60	1,600.35	35.9	89.1	35.39
LASSO	3.44	207.21	195.37	98.9	99.9	3.73
True	0.17	234.63	234.60	98.9	99.9	3.68
Random	20.62	636.58	211.38	99.8	97.6	15.21
Panel D:			N=16,000, 10% treated			
Probit	1.56	109.86	107.45	99.1	95.1	2.47
RF	-12.40	440.31	286.46	74.9	96.8	17.89
LASSO	1.40	86.05	84.08	99.4	99.9	2.67
True	-0.19	95.90	95.86	99.5	99.9	2.70
Random	20.63	507.72	82.22	99.9	99.2	16.09
Panel E:			N=16,000, 25% treated			
Probit	2.63	49.80	42.87	99.7	99.3	1.56
RF	1.10	72.73	71.50	95.3	98.6	8.50
LASSO	1.15	42.34	41.03	99.7	99.8	2.19
True	-0.72	53.62	53.10	99.7	99.4	2.04
Random	24.52	641.45	40.42	99.9	99.9	19.39

Notes: Figures shown are the mean of the respective measure over 1,000 (in Panel A&B) or 250 (in Panel C-E) replications. RF stands for Random Forest. Random indicates the randomized PS. Bias is the mean bias over all simulation repetitions. MSE is the mean squared error. CS and CS, treated is the common support (for the treated) and SC is the (mean) absolute standardized bias in covariate balancing of the ten most important confounders. The full results can be found in Tables 3.9, 3.11, 3.13, 3.15 & 3.17 in the Appendix.

Although it is not clear how low the SB should be and if this translates directly into good final ATET estimates this is indicative for the bad performance of the Random Forest in the matching step with 10 percent treated as can be seen in Panels A and D of Table 3.6, columns (1)-(3). The LASSO PS is only slightly biased and the

resulting MSE is the lowest despite the true PS results in Panel A and even lower than the true PS in Panel D (compare Abadie and Imbens, 2016, for this phenomenon). Panels B and E are giving some insights into the simulations with the higher treatment share. All estimation techniques performed better compared to the lower treatment share, with the LASSO outperforming the other PS in terms of MSE and MAD. More observations, as can be seen in Panels D and E, generally improves the performance of every method. Estimating the PS with the Probit is benefiting from the larger sample especially by reducing the mean bias compared to the low observation scenarios. The Random Forest PS works decently well with 25 percent treated units, i.e. the bias is closest to zero, but is biased with a lower share of treated and has the highest variance in every scenario. For the lowest treatment share of 5 percent in Panel C we find that Probit and LASSO perform well in terms of bias and MSE, while the Random Forest is doing much worse than the other methods.

Columns (4) and (5) report the share of observations remaining in common support (CS), overall, as well as for treated only. Here we find the Probit and the Random Forest to have the lowest overlap in Panel A and C, as well as, but less extreme, in Panel D. Less severely, this is also observed in Panels B and E in the simulation for 25 percent treated. No huge support problems are observed for the LASSO, as well as the true and the random PS.

3.6 Empirical application

We evaluate the effect of participating in the training programme, "Determining, Reducing and Removing Employment Impediments", using the full sample of 14,817 treated and 261,820 control units as described in Section 3.3. The ATET is estimated using the three PS methods, i.e. Random Forest, LASSO and Probit, in the RMBA estimator. The results can be found in Table 3.7.

Table 3.7: Empirical treatment effect estimation, matching

Propensity score method used	Treatment effect	Standard error	P-value	SB	Common support
Probit	26.59	4.34	0.00	0.89	99.9%
LASSO	27.92	2.00	0.00	2.07	99.9%
Random Forest	36.62	3.13	0.00	6.62	99.0%

Notes: Average treatment effect on the treated. N=276,637. The outcome is days in employment in the three years after treatment. Inference based on bootstrapping (299 replications) p-values. SB is the absolute mean standardized bias in covariate balancing of the ten most important cofounders.

Although LASSO (and Probit) performed well in our simulation exercise as PS estimation technique for 16,000 observations, this gives us only little indication how this performance translates into this larger sample.[52]

We find that participation in the investigated training programme leads to about 27 days more in employment compared to not being assigned to the programme. The effect estimated using the Probit PS, with 26.6 days and the effect using the LASSO PS, with 27.9 days, are roughly equal. The estimates using the Random Forest for estimating the PS suggest an effect of about 36.6 days, which is compared to the LASSO estimate around 30 percent higher. Worth noting is the fact that the estimated standard error is remarkably lower if the PS is estimated using LASSO compared to the other methods. The common support and the SB for all the methods is found to be similar to the findings in the simulation.

Table 3.8: Covariate balancing in application

	Before Matching	Probit	Random Forest	LASSO
Female[1)]	-2.50	0.20	0.20	0.60
Age	-22.13	1.69	-7.41	2.35
Receives some income from employment[1)]	22.60	0.40	-4.10	0.10
Cumulated number of days in welfare receipt in year before	-13.35	0.66	-5.32	0.46
Participated in Schemes by Providers[1)]	7.50	0.20	-2.50	-0.50
Participated in classroom training[1)]	15.30	-0.30	-5.30	-1.70
Job centre district: Inflow into Schemes by Providers relative to jobseeker stock in 2009q4	48.54	1.01	-15.03	-4.89
Job centre district: Inflow into In-Firm Training relative to jobseeker stock in 2009q4	19.75	-0.32	-1.12	-4.00
Days since last employment	-13.63	-1.68	-3.61	0.39
Cumulated days in regular employment in last five years	12.89	-0.99	3.88	-2.14

Notes: Covariate balancing after matching in the application using the three different PS estimation methods. N=276,673. Mean bias in percent for binary, standardized bias in percent for non-binary variables. [1)]binary variable.

[52] In Appendix B.5, we show the distributions of the PS are very similar for the Probit and the LASSO, while the Random Forest estimates a slightly narrower distribution.

In Table 3.8, we provide the covariate balancing statistics for the ten most important confounders. While the Probit is balancing every covariate well, the Random Forest PS shows deficits in balancing some of the, especially non-binary, variables.[53] In conclusion, the choice of the first stage estimator does matter in practical research and choosing a non-appropriate method could lead to wrong policy conclusions.

3.7 Conclusion

In this work, we investigated through simulations and an application whether predicting the PS by machine learning methods helps to increase credibility of programme evaluation studies based on propensity score matching. Having an arguably realistic DGP using a rich, high-dimensional administrative dataset for German long-term unemployed, we simulated the finite sample performance of various PS estimation techniques in a matching-type estimator estimating the ATET. We considered two very different methods from the machine learning literature, namely the Random Forest and the LASSO. We compared their performance to a "classical" Probit approach with an ad-hoc specification of covariates, as well as to the true and a randomized PS.

While the choice of "first-stage" estimator is highly relevant for settings with a low number of observations and few treated, the methods become more similar in terms of performance with more observations and/or more treated units. We find that LASSO is doing especially well, being close or even better than using true PS in matching. Our evidence suggest the usage of Random Forest for this purpose might lead to misleading results, especially if the share of treated is low, and using it in similar setups has to be considered with caution. This could be the case because in these situations the Random Forest is not able to split deep enough to balance the covariates properly. The target of the PS in matching is to balance confounding factors to obtain a quasi-random situation. Therefore, the Random Forest was not able to replicate the spread of the PS, which led to comparing control units to treated units, which were potentially not sufficiently similar in terms of confounding influences. Also, if the tree structures are not able to split deep enough they cannot estimate the tails well. Athey and Imbens (2019) point out the fact that forests are likely to have bias in the tails, because the single trees cannot centre their leaves near the boundary. This might be more

[53] To balance non-binary variables trees potentially need more splits compared to binary variables. Having a low share of treated the single trees might not be able to split deep enough to balance especially non-binary variables.

pronounced the lower the treatment share. Further research would be helpful to understand this phenomenon in our context more deeply.

In our application we see this sensitivity again: LASSO and Probit as PS estimator used in radius matching lead to similar point estimates, but with lower variance for the LASSO. The estimator based on a Random Forest estimated PS shows a substantial deviation in the magnitude of the effect compared to the other methods.

The conclusion of these exercises is that estimating the propensity score by machine learning is not clearly beneficial compared to current conventional matching methods. Instead, the methods of the new causal machine literature that are directly optimized for treatment effect estimation may be a more promising alternative, although this is beyond the scope of this paper (see Knaus, Lechner, and Strittmatter, 2020, and Lechner, 2018, for various proposals and comparisons).

Of course, as the machine learning methods rely on different tuning parameters, more tailored implementations might improve the performance and reliability. Despite relying on a realistic DGP, it remains unclear if the results hold for studies outside the labour market context and further research might be useful here, especially considering the case of low (high) shares of treated units. Further, recent developments in the literature of doubly robust alternatives (compare e.g. Antonelli et al., 2018; Chernozhukov et al., 2018) might be helpful for increasing the credibility of empirical researches.

Appendices

Appendix 3.A: Full result tables

In this Appendix, we show the full results tables of the EMCS presented in Section 3.3.5. The following four subsections refer to the four simulation scenarios. Summaries of those tables are found in the text.

Appendix 3.A.1: Scenario A: N=4,000, 10% treated

Table 3.9: Simulation results for N=4,000 and ~10% share of treated

Measure	Probit	Probit (conv.)	Random Forest	LASSO	True	Random
			Treatment effects			
Mean treatment effect / bias	21.95	16.60	-26.15	5.03	-0.39	20.47
Mean SE of matching[1)]	19.65	21.76	31.03	20.07	20.17	19.41
MAD	25.11	21.44	36.12	15.91	14.83	23.12
MSE	885.52	706.23	2,258.05	398.29	341.25	773.07
SE	20.09	20.75	39.68	19.31	18.47	18.81
Variance	403.54	430.75	1,574.16	372.94	341.10	353.89
Skewness	-0.27	0.002	-0.78	-0.16	-0.02	0.08
Kurtosis	3.05	2.98	4.75	2.95	2.77	3.32
			Common Support			
Mean share remaining in CS	0.93	0.91	0.57	0.98	0.98	0.99
Mean share treated remaining in CS	0.65	0.99	0.90	0.99	0.99	0.99
			Balancing of covariates as standardized differences			
Mean abs. stand. mean bias	8.20	3.74	28.18	5.49	5.33	16.34
Mean abs. stand. max. bias	19.14	8.65	106.29	12.47	12.51	37.54
Sample size	4,000	4,000	4,000	4,000	4,000	4,000
Replications	1,000	653	1,000	1,000	1,000	1,000
Share of treated	0.0993	0.0935	0.0993	0.0993	0.0993	0.0993

Notes: SE: standard error. CS stands for common support. In column 2, only those repetitions are taken into account in which the Probit was able to converge correctly. Balancing of covariates according to the ten most important confounders, determined as those variables selected in both LASSO procedures, Y on X and D on X, in the full dataset. [1)]Estimated as weight-based variance as described in Huber, Lechner, and Steinmayr (2015).

Table 3.10: Propensity score results for N=4,000 and ~10% treated

Measure	Probit	Probit (conv.)	Random Forest	LASSO	Random
Mean correlation	0.36	0.56	0.70	0.75	0.00
Mean Kendall's Tau	0.26	0.39	0.53	0.58	0.00
Mean Spearman Rank	0.36	0.56	0.72	0.77	0.00
Sample size	4,000	4,000	4,000	4,000	4,000
Replications	1,000	653	1,000	1,000	1,000
Share of treated	0.0993	0.0935	0.0993	0.0993	0.0993

Notes: In column 2, only those repetitions are taken into account in which the Probit was able to converge correctly. The formulas for Kendall's Tau and the Spearman Rank Correlation can be found in the main text.

Appendix 3.A.2: Scenario B: N=4,000, 25% treated

Table 3.11: Simulation results for N=4,000 and ~25% share of treated

Measure	Probit	Probit (conv.)	Random Forest	LASSO	True	Random
			Treatment effects			
Mean treatment effect / bias	11.68	11.27	-2.18	3.63	-0.32	24.29
Mean SE of matching[1)]	14.51	14.66	16.63	14.84	15.02	13.10
MAD	14.52	14.22	13.14	11.50	12.10	24.63
MSE	310.55	299.99	275.33	213.92	226.28	762.48
SE	13.19	13.15	16.45	14.17	15.04	13.15
Variance	174.05	172.99	270.57	200.73	226.18	172.80
Skewness	-0.08	-0.05	-0.07	-0.17	0.01	0.00
Kurtosis	3.17	3.21	3.03	3.37	2.89	2.87
			Common Support			
Mean share remaining in CS	0.98	0.98	0.94	0.99	0.99	0.99
Mean share treated remaining in CS	0.96	0.99	0.97	0.99	0.99	0.99
			Balancing of covariates as standardized differences			
Mean abs. stand. mean bias	3.13	2.66	9.41	4.03	4.06	19.46
Mean abs. stand. max. bias	7.90	6.36	31.75	10.07	9.73	46.47
Sample size	4,000	4,000	4,000	4,000	4,000	4,000
Replications	1,000	961	1,000	1,000	1,000	1,000
Share of treated	0.2493	0.2485	0.2493	0.2493	0.2493	0.2493

Notes: SE: standard error. CS stands for common support. In column 2, only those repetitions are taken into account in which the Probit was able to converge correctly. Balancing of covariates according to the ten most important confounders, determined as those variables selected in both LASSO procedures, Y on X and D on X, in the full dataset. [1)]Estimated as weight-based variance as described in Huber, Lechner, and Steinmayr (2015).

Table 3.12: Propensity score results for N=4000 and ~25% treated

Measure	Probit	Probit (conv.)	Random Forest	LASSO	Random
Mean correlation	0.61	0.64	0.80	0.86	0.00
Mean Kendall's Tau	0.43	0.44	0.62	0.67	0.00
Mean Spearman Rank	0.60	0.62	0.82	0.86	0.00
Sample size	4,000	4,000	4,000	4,000	4,000
Replications	1,000	961	1,000	1,000	1,000
Share of treated	0.2493	0.2485	0.2493	0.2493	0.2493

Notes: In column 2, only those repetitions are taken into account in which the Probit was able to converge correctly. The formulas for Kendall's Tau and the Spearman Rank Correlation can be found in the main text.

Appendix 3.A.3: Scenario C: N=16,000, 5% treated

Table 3.13: Simulation results for N=16,000 and ~5% share of treated

Measure	Probit	Probit (conv.)	Random Forest	LASSO	True	Random
			Treatment effects			
Mean treatment effect / bias	6.73	2.77	24.93	3.44	0.17	20.62
Mean SE of matching[1)]	13.06	13.81	27.13	13.76	13.82	13.44
MAD	13.45	11.46	35.25	11.34	11.89	21.84
MSE	271.81	201.09	2,221.60	207.21	234.63	636.58
SE	15.05	13.91	40.00	13.98	15.32	14.54
Variance	226.47	193.39	1,600.35	195.37	234.60	211.38
Skewness	-0.15	-0.23	-0.80	-0.18	-0.42	-0.12
Kurtosis	3.20	2.89	4.29	3.23	3.36	3.00
			Common Support			
Mean share remaining in CS	0.98	0.97	0.36	0.99	0.99	0.99
Mean share treated remaining in CS	0.76	0.99	0.89	0.99	0.99	0.98
			Balancing of covariates as standardized differences			
Mean abs. stand. mean bias	5.41	2.37	35.39	3.73	3.68	15.21
Mean abs. stand. max. bias	12.59	5.52	135.67	8.80	8.54	36.08
Sample size	16,000	16,000	16,000	16,000	16,000	16,000
Replications	250	190	250	250	250	250
Share of treated	0.0498	0.0499	0.0498	0.0498	0.0498	0.0498

Notes: SE: standard error. CS stands for common support. In column 2, only those repetitions are taken into account in which the Probit was able to converge correctly. Balancing of covariates according to the ten most important confounders, determined as those variables selected in both LASSO procedures, Y on X and D on X, in the full dataset. [1)]Estimated as weight-based variance as described in Huber, Lechner, and Steinmayr (2015).

Table 3.14: Propensity score results for N=16,000 and ~5% treated

Measure	Probit	Probit (conv.)	Random Forest	LASSO	Random
Mean correlation	0.48	0.64	0.71	0.80	-0.00
Mean Kendall's Tau	0.35	0.47	0.56	0.62	-0.00
Mean Spearman Rank	0.49	0.65	0.75	0.82	-0.00
Sample size	16,000	16,000	16,000	16,000	16,000
Replications	250	190	250	250	250
Share of treated	0.0498	0.0499	0.0498	0.0498	0.0498

Notes: In column 2, only those repetitions are taken into account in which the Probit was able to converge correctly. The formulas for Kendall's Tau and the Spearman Rank Correlation can be found in the main text.

Appendix 3.A.4: Scenario D: N=16,000, 10% treated

Table 3.15: Simulation results for N=16,000 and ~10% share of treated

Measure	Probit	Random Forest	LASSO	True	Random
		Treatment effects			
Mean treatment effect / bias	1.56	-12.40	1.40	-0.19	20.63
Mean SE of matching[1)]	9.98	13.39	10.04	10.06	9.70
MAD	8.12	16.82	7.63	7.71	20.63
MSE	109.86	440.31	86.05	95.90	507.72
SE	10.37	16.93	9.17	9.79	9.07
Variance	107.45	286.46	84.08	95.90	507.72
Skewness	0.48	-0.24	-0.03	0.07	0.17
Kurtosis	3.65	3.50	2.49	2.97	2.67
		Common support			
Mean share remaining in CS	0.99	0.75	0.99	0.99	0.99
Mean share treated remaining in CS	0.95	0.97	0.99	0.99	0.99
		Balancing of covariates as standardized differences			
Mean abs. stand. mean bias	2.47	17.89	2.67	2.70	16.09
Mean abs. stand. max. bias	5.80	71.11	6.70	6.33	37.62
Sample size	16,000	16,000	16,000	16,000	16,000
Replications	250	250	250	250	250
Share of treated	0.0997	0.0997	0.0997	0.0997	0.0997

Notes: SE: standard error. CS stands for common support. Balancing of covariates according to the ten most important confounders, determined as those variables selected in both LASSO procedures, Y on X and D on X, in the full dataset. [1)]Estimated as weight-based variance as described in Huber, Lechner, and Steinmayr (2015).

Table 3.16: Propensity score results for N=16,000 and ~10% treated

Measure	Probit	Random Forest	LASSO	Random
Mean correlation	0.73	0.79	0.86	0.00
Mean Kendall's Tau	0.54	0.62	0.68	0.00
Mean Spearman Rank	0.73	0.81	0.87	0.00
Sample size	16,000	16,000	16,000	16,000
Replications	250	250	250	250
Share of treated	0.0997	0.0997	0.0997	0.0997

Notes: The formulas for Kendall's Tau and the Spearman Rank Correlation can be found in the main text.

Appendix 3.A.5: Scenario E: N=16,000, 25% treated

Table 3.17: Simulation results for N=16,000 and ~25% share of treated

Measure	Probit	Random Forest	LASSO	True	Random
		Treatment effects			
Mean treatment effect / bias	2.63	1.10	1.15	-0.72	24.52
Mean SE of matching[1)]	7.37	8.59	7.53	7.59	6.55
MAD	5.55	6.76	5.14	5.80	24.52
MSE	49.80	72.73	42.34	53.62	641.45
SE	6.55	8.46	6.41	7.29	6.36
Variance	42.87	71.50	41.03	53.10	40.42
Skewness	0.28	0.01	-0.21	0.03	0.29
Kurtosis	4.18	2.98	3.37	3.31	2.86
		Common support			
Mean share remaining in CS	0.99	0.95	0.99	0.99	0.99
Mean share treated remaining in CS	0.99	0.99	0.99	0.99	0.99
		Balancing of covariates as standardized differences			
Mean abs. stand. mean bias	1.56	8.50	2.19	2.04	19.39
Mean abs. stand. max. bias	3.70	27.94	6.14	4.78	46.35
Sample size	16,000	16,000	16,000	16,000	16,000
Replications	250	250	250	250	250
Share of treated	0.2500	0.2500	0.2500	0.2500	0.2500

Notes: SE: standard error. CS stands for common support. Balancing of covariates according to the ten most important confounders, determined as those variables selected in both LASSO procedures, Y on X and D on X, in the full dataset. [1)]Estimated as weight-based variance as described in Huber, Lechner, and Steinmayr (2015).

Table 3.18: Propensity score results for N=16,000 and ~25% treated

Measure	Probit	Random Forest	LASSO	Random
Mean correlation	0.86	0.86	0.92	0.00
Mean Kendall's Tau	0.69	0.67	0.76	0.00
Mean Spearman Rank	0.87	0.86	0.92	0.00
Sample size	16,000	16,000	16,000	16,000
Replications	250	250	250	250
Share of treated	0.2500	0.2500	0.2500	0.2500

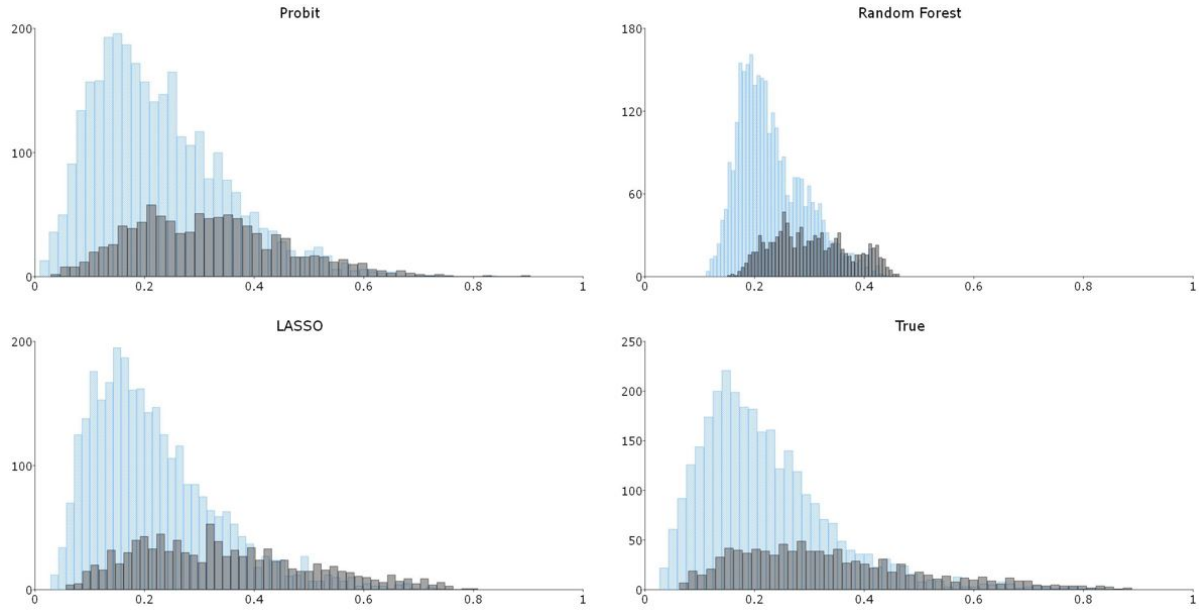
Notes: The formulas for Kendall's Tau and the Spearman Rank Correlation can be found in the main text.

Appendix 3.B: Estimated propensity score by treatment status

The distributions of the same one PS estimation for each scenario from the EMCS in Section 3.5 is presented in the appendices 3.B.1 – 3.B.4. Scenario A can be found in the main text. The distribution of the PS of the Probit, Random Forest and LASSO from the application in Section 3.6 are depicted in 3.B.5.

Appendix 3.B.1: Scenario B: N=4,000, 25% treated

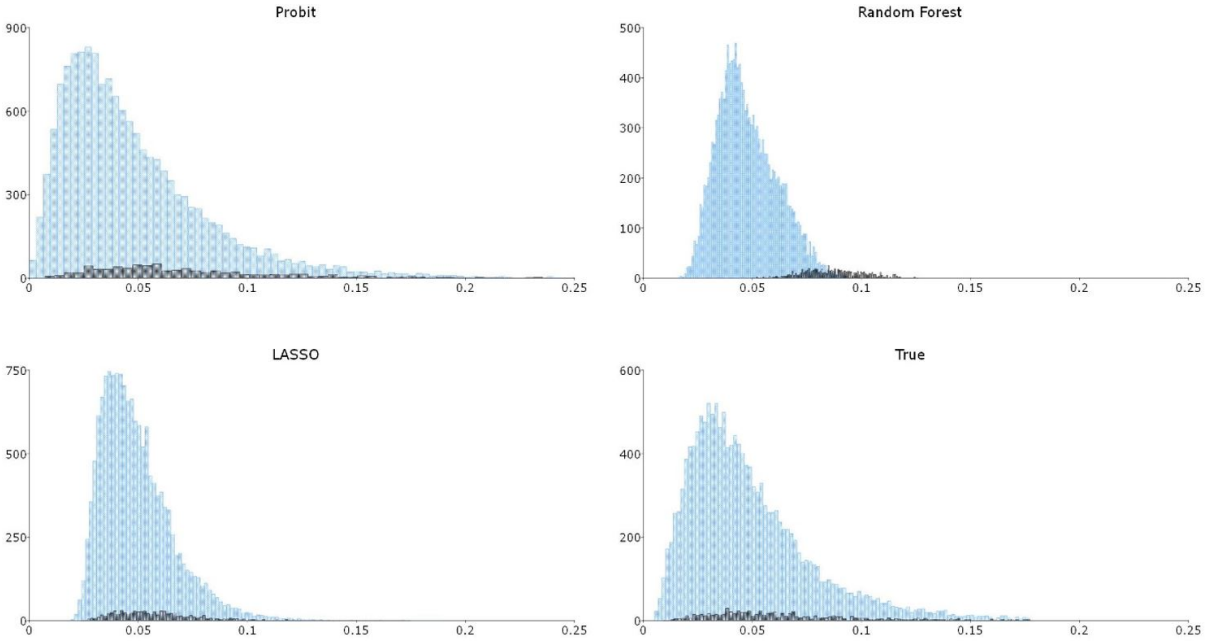
Figure 3.3: Propensity scores by treatment status, N=4,000, 25% treated



Notes: Histograms with PS on the horizontal axis. Top left is the Probit PS, top right Random Forest, bottom left and right the LASSO estimated and true PS. Each from the same one simulation with N=4,000 and 25% treatment share. Control units are light, treated units dark shaded.

Appendix 3.B.2: Scenario C: N=16,000, 5% treated

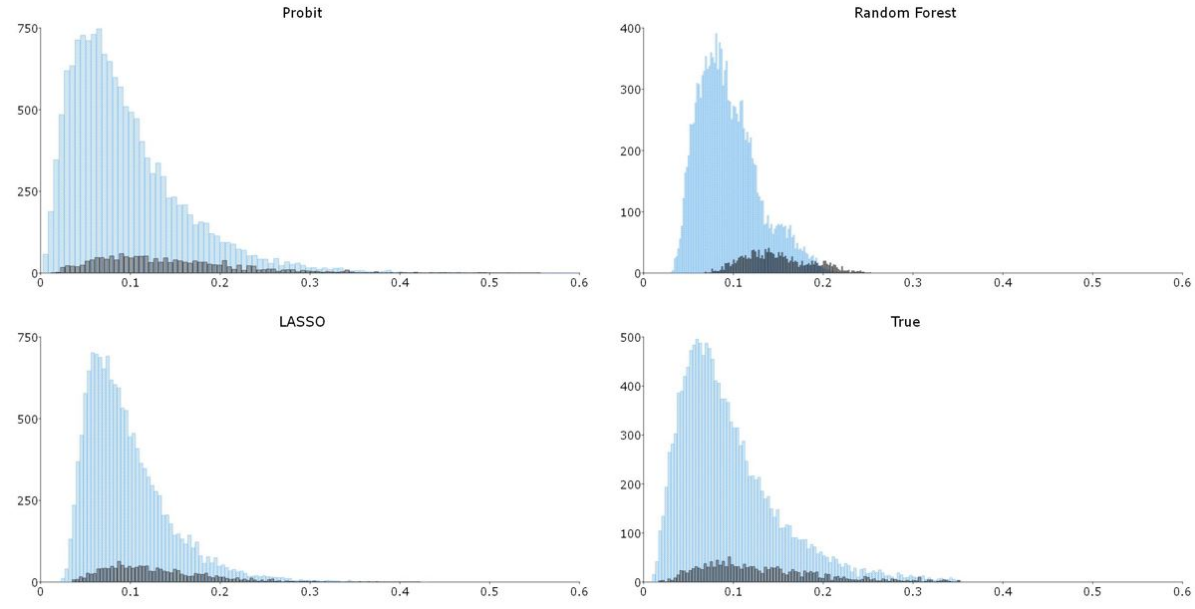
Figure 3.4: Propensity scores by treatment status, N=16,000, 5% treated



Notes: Histograms with PS on the horizontal axis. Top left is the Probit PS, top right Random Forest, bottom left and right the LASSO estimated and true PS. Each from the same one simulation with N=16,000 and 5% treatment share. Control units are light, treated units dark shaded.

Appendix 3.B.3: Scenario D: N=16,000, 10% treated

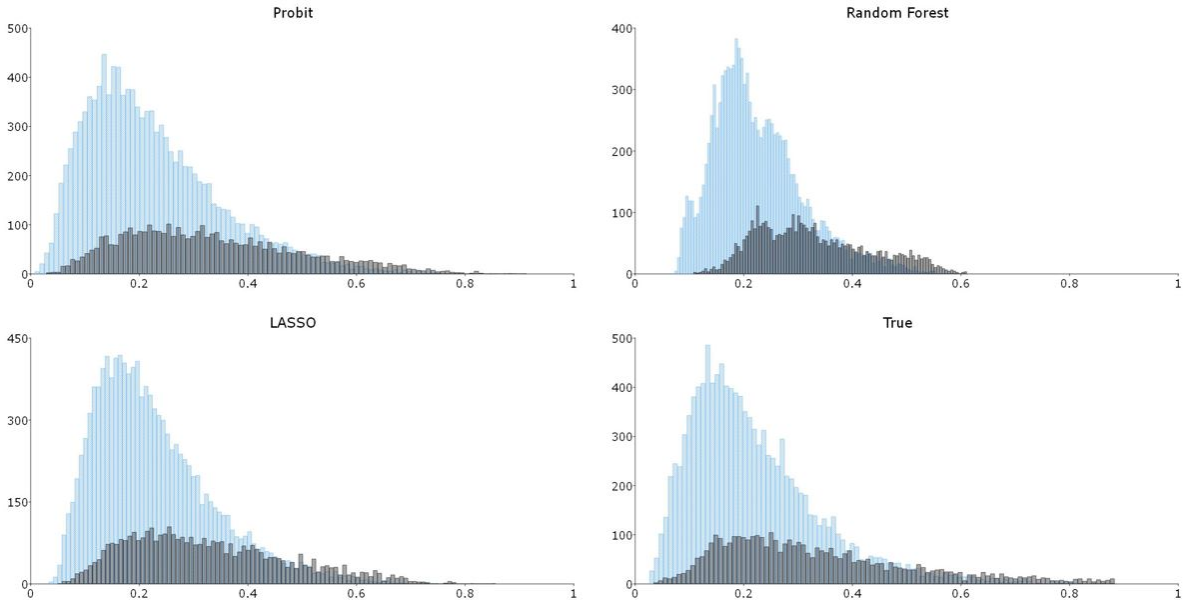
Figure 3.5: Propensity scores by treatment status, N=16,000, 10% treated



Notes: Histograms with PS on the horizontal axis. Top left is the Probit PS, top right Random Forest, bottom left and right the LASSO estimated and true PS. Each from the same one simulation with N=16,000 and 10% treatment share. Control units are light, treated units dark shaded.

Appendix 3.B.4: Scenario E: N=16,000, 25% treated

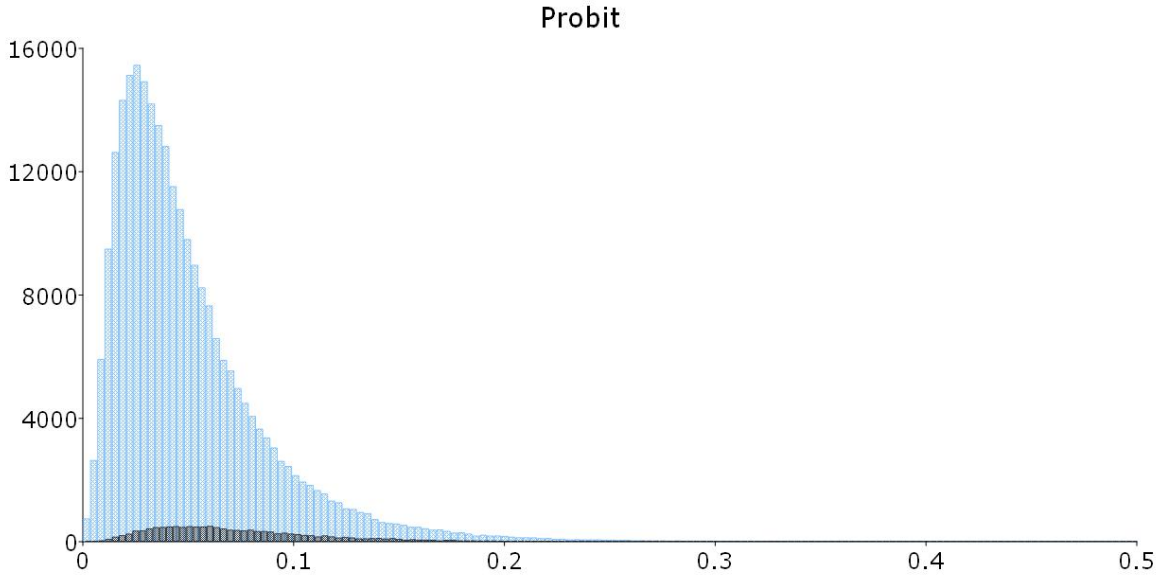
Figure 3.6: Propensity scores by treatment status, N=16,000, 25% treated



Notes: Histograms with PS on the horizontal axis. Top left is the Probit PS, top right Random Forest, bottom left and right the LASSO estimated and true PS. Each from the same one simulation with N=16,000 and 25% treatment share. Control units are light, treated units dark shaded.

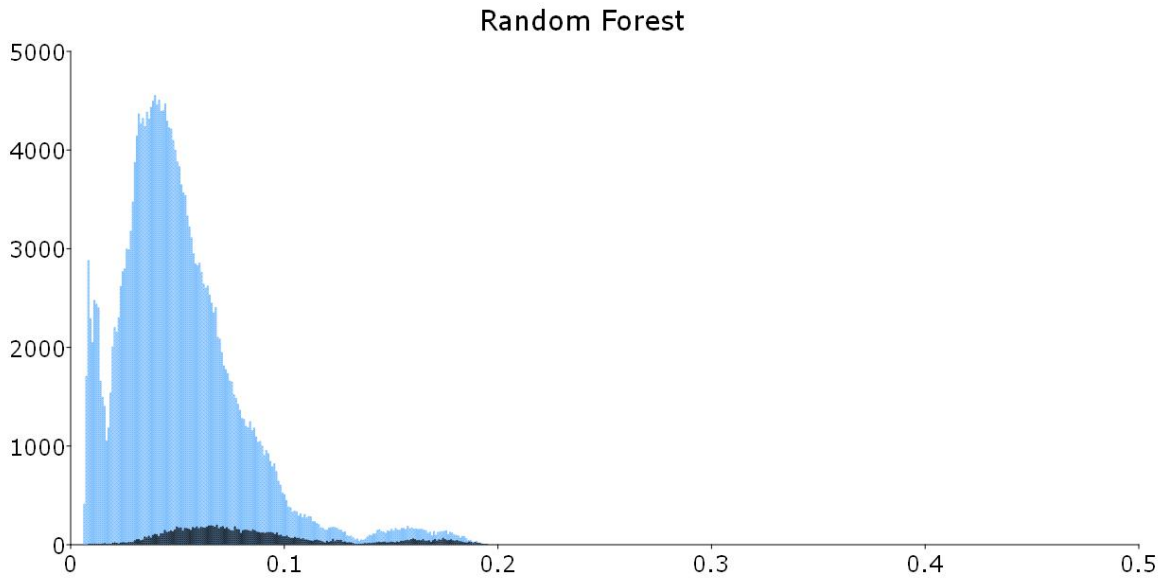
Appendix 3.B.5 Application

Figure 3.7: Propensity scores by treatment status, Probit



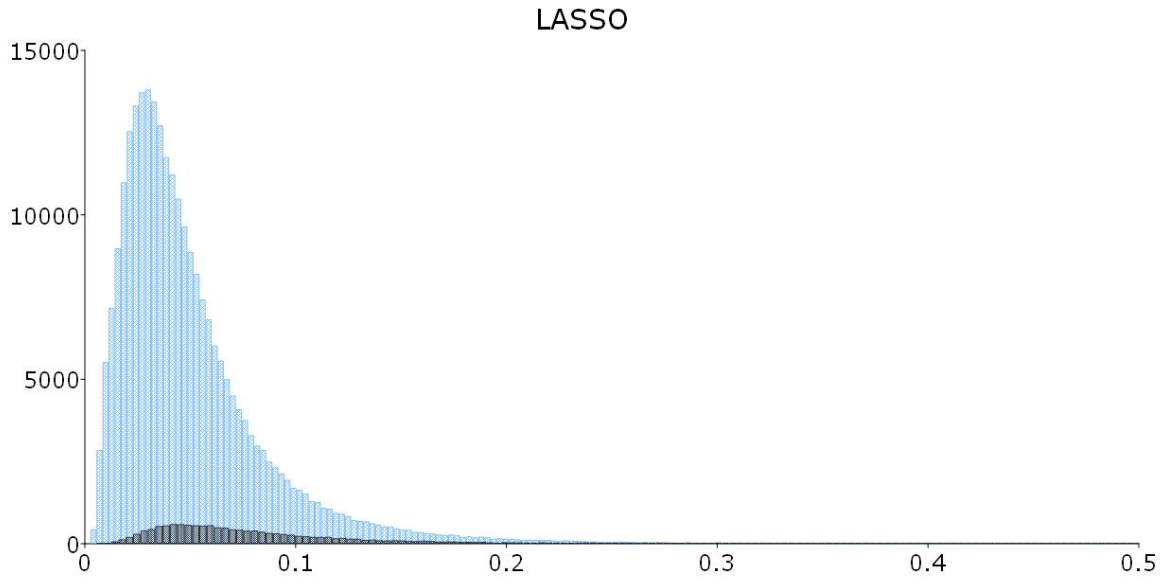
Notes: Histogram with PS on the horizontal axis estimated using the Probit. From the application in Section 3.6 with N=276,637 and about 5% treatment share. Control units are light, treated units dark shaded.

Figure 3.8: Propensity scores by treatment status, Random Forest



Notes: Histogram with PS on the horizontal axis estimated using the Random Forest. From the application in Section 3.6 with N=276,637 and about 5% treatment share. Control units are light, treated units dark shaded.

Figure 3.9: Propensity scores by treatment status, LASSO



Notes: Histogram with PS on the horizontal axis estimated using the LASSO. From the application in Section 3.6 with N=276,637 and about 5% treatment share. Control units are light, treated units dark shaded.

Appendix 3.C: Radius matching with bias adjustment estimator

The Radius matching with bias adjustment estimator of Lechner, Miquel, and Wunsch (2011) combines the features of a distance-weighted radius matching estimator with a regression based (bias) adjustment, which removes potential biases due to mismatches. Distance weighting implies a weighting of control units in the specific radius inversely proportional to their distance to the respective treated unit. Therefore, in contrast to

other matching algorithms control units do not get the same weight, but according to their distance, in terms of confounding influences, to the treated unit. The weighted control observations can finally be compared to the respective treated unit as if they were in an experimental setting.

Table 3.19 gives a detailed matching protocol used for estimating the ATET. For further details on implementation and properties of this estimator the reader is referred to Huber, Lechner, and Wunsch (2013) and Huber, Lechner, and Steinmayr (2015).

Table 3.19: Radius matching protocol

Step A-1	Choose one observation in the subsample defined by d=1 and delete it from that pool.
Step B-1	Find an observation in the subsample defined by d=0 that is as close as possible to the one chosen in Step A-1 in terms of (x). Closeness is based on the Mahalanobis distance.
Step C-1	Repeat A-1 and B-1 until no observation with d=1 is left.
Step D-1	Compute the maximum distance (dist) obtained for any comparison between a member of the reference distribution and matched comparison observations.
Step A-2	Repeat A-1.
Step B-2	Repeat B-1. If possible, find other observations in the subsample of d=0 that are at least as close as R*dist to the one chosen in step A-2. Do not remove these observations, so that they can be used again. Compute weights for all chosen comparisons observations that are proportional to their distance. Normalise the weights such that they add to one.
Step C-2	Repeat A-2 and B-2 until no participant in d=1 is left.
Step D-2	For any potential comparison observation, add the weights obtained in A-2 and B-2.
Step E	Using the weights $w(x_i)$ obtained in D-2, run a weighted linear regression of the outcome variable on the variables used to define the distance (and an intercept).
Step F-1	Predict the potential outcome $Y^0(x_i)$ of every observation using the coefficients of this regression $\hat{Y}^0(x_i)$.
Step F-2	Estimate the bias of the matching estimator for $E(Y^0\|D=1)$ as: $\sum_{i=1}^{N} \frac{(1-d_i)w_i\hat{y}^0(x_i)}{N_0} - \frac{d_i\hat{y}^0(x_i)}{N_1}$.
Step G	Using the weights obtained by weighted matching in D-2, compute a weighted mean of the outcome variables in d=0. Substract the bias from this estimate to get $E(Y^0\|D=1)$.

Note: R is the distance measure and set to 150%.

4 Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues

Daniel Goller and Alex Krumer

Abstract: Balancing the allocation of games in sports competitions is an important organizational task that can have serious financial consequences. In this paper, we examine data from 10,142 soccer games played in the top German, Spanish, French, and English soccer leagues between 2007/2008 and 2016/2017. Using a machine learning technique for variable selection and applying a semi-parametric analysis of radius matching on the propensity score, we find that all four leagues have a lower attendance in games that take place on four non-frequently played days than those on three frequently played days. We also find that, in all leagues, there is a significantly lower home advantage for the underdog teams on non-frequent days. Our findings suggest that the current schedule favours underdog teams with fewer home games on non-frequent days. Therefore, to increase the fairness of the competitions, it is necessary to adjust the allocation of the home games on non-frequent days in a way that eliminates any advantage driven by the schedule. These findings have implications for the stakeholders of the leagues, referees' and calendar committees as well as for coaches and players.

Keywords: Performance, schedule effects, soccer

JEL classification: D00, L00, D20, Z20

4.1 Introduction

In recent decades, top European soccer leagues have become large business corporations. Each of the top leagues receives more than 1 billion Euros from television revenues alone.[54] A large part of these amounts is redistributed to teams based on their performance. In addition, the highest ranked teams of the top leagues earn the right to participate in the UEFA Champions League and Europa League. According to UEFA (the governing body of soccer in Europe), in the 2016/2017 season, more than 1.3 billion Euros was shared among the clubs in the Champions League and almost 400 million Euros among the clubs in the Europa League.[55] Since an unbalanced schedule may have serious financial consequences, the leagues face an important organizational task in creating a schedule that will not discriminate against or favour specific teams.

Top European soccer leagues use a double round-robin structure, where each team competes against each other team twice during the season. Operational research literature has intensively investigated different issues of round-robin structures, such as balanced distribution of home and away matches (Della Croce and Oliveri, 2006; Durán, Guajardo, and Sauré, 2017), break optimization (Ribeiro and Urrutia, 2007), police requirements (Kendall et al., 2010), stakeholders' requirements (Goossens and Spieksma, 2009), minimizing traveling distance (Kendall, 2008; Durán et al., 2019), optimizing the number of prizes (Krumer, Megidish, and Sela, 2020), and allocation of rescheduled games (Yi, Goossens, and Nobibon, 2020).[56] However, the operational research literature has neglected another important issue: the allocation of (non-rescheduled) games between days that are not the usual days in a league's calendar. This may play an important role, because fans may have different preferences toward certain days of the week (Wang, Goossens, and Vandebroek, 2018) or even the timings of the game (Krumer, 2020). For example, if fans are not used to attending games on a certain day, their routine may prevent them from attending games on those days. In such cases, lower attendance may be expected on these days, which may reduce the home advantage (Downward and Jones, 2007; Nevill et al., 2002; Page and Page, 2010; Pettersson-Lidbom and Priks, 2010). In the 2017/2018 season, for example, the

[54] From https://www.euronews.com/2018/05/30/footbal-broadcast-rights-ligue-1-championship-booms-into-the-billion and https://www.soccerex.com/insight/articles/2018/laliga-s-new-tv-rights-distribution-model-a-level-playing-field. Last accessed on 31.01.2019.

[55] From http://www.uefa.com/uefachampionsleague/news/newsid=2398575.html and http://www.uefa.com/uefaeuropaleague/news/newsid=2398584.html . Last accessed on 31.01.2019.

[56] See the comprehensive reviews of Rasmussen and Trick (2008), Kendall et al. (2010), Goossens and Spieksma (2012), and Van Bulck et al. (2020) for additional references.

decision to schedule Monday games in the German Bundesliga 1 led to large protests. Moreover, German team 1.FSV Mainz 05 officially complained to the German Federation (DFL) since it had to play eight non-weekend games, six of which were at home. The DFL finally decided to abolish Monday games.[57]

The present paper is closely related to the study of Krumer and Lechner (2018), who investigated games in the German Bundesliga 1 and found a significantly lower attendance and also a lower home advantage in midweek days compared to weekend days (Friday, Saturday, Sunday − the most frequently played days in this league). In other leagues, however, the three most frequent days, which account for approximately 90 percent of all matches, differ from those in Bundesliga 1. For example, in England and France, the three most frequent days for games are Saturday, Sunday, and Wednesday, whereas the respective days in Spain are Saturday, Sunday, and Monday.[58]

In this paper, we ask a simple question: Does playing on non-frequent days have any effect on the various aspects of soccer games? More specifically, using different definition of non-frequent days and applying data from the four above-mentioned European soccer leagues between 2007/2008 and 2016/2017, we compare the games that were played on frequent and on non-frequent days with regard to their attendance and home advantage. More specifically, unlike Krumer and Lechner (2018), we separately investigated games with a home advantage for the favourite team and games with a home advantage for the underdog team. This is another contribution to the operational research literature, which has not yet considered the possibility of such a heterogeneous effect.[59] To the best of our knowledge, the only paper to study such an effect in a context of schedule is Krumer (2020), which investigated the UEFA Europa League games with kick-off times at 19:00 CET and 21:05 CET. That paper documented a lower attendance in games that started at 21:05 CET and a significantly lower home advantage for the underdog teams in these later games.

It is important to note that the allocation of the match days is not entirely random, and might be based on different schedule-related features such as public holidays,

[57] For additional information, see: https://www.mainz05.de/news/brief-an-die-dfl-kritik-an-terminierungen/ (in German) and https://www.dw.com/en/bundesliga-monday-games-to-be-discontinued-as-fan-protests-persist/a-46390559. Last accessed on 03.11.2019.

[58] According to the UEFA association club coefficients, these are four of the five most successful leagues in Europe. We do not use data on the fifth (the Italian Serie A) since it suffered from various scandals and club insolvencies in the underlying period. See, for example, Buraimo, Migali and Simmons (2016), who found a significantly lower crowd attendance after the Calcipoli scandal in the 2005/2006.

[59] We discuss this type of heterogeneity in detail in the data section.

international breaks, European tournaments, police requirements, broadcasters' and clubs' interests, months of the year, and even teams' values. Therefore, we need to control for these deviations from random selection into treatment (that is, non-frequent days) using a selection-on-observables approach. Specifically, we estimated the average treatment effect of playing on the non-frequent days by using the distance-weighted radius matching approach with bias adjustment suggested by Lechner, Miquel, and Wunsch (2011). This estimator is constructed to be more robust than other matching-type estimators, as it combines the features of distance-weighted radius matching with a bias adjustment to remove sample biases due to mismatches (Huber, Lechner and Wunsch, 2013). In addition, having a rich database in terms of potential confounding variables, we use a machine learning technique for variable selection as proposed by Belloni, Chernozhukov, and Hansen (2014).

Based on analysis of 10,142 games from the top four European leagues over 10 seasons, we found a significantly lower attendance on four non-frequent days in all four leagues. In addition, all of the leagues had a reduced home advantage on four non-frequent days for the underdog teams, which is in line with Krumer (2020). Our results suggest that the difference in the number of points between the favourite and the underdog teams, when the game takes place on non-frequent days compared to frequent days, is 0.49 in Ligue 1, 0.43 in La Liga, 0.42 in the Premier League, and 0.63 in Bundesliga 1. To put these numbers into perspective, a favourite with home advantage gains about 1.1 points more than the underdog, on average.[60,61]

Such a reduced home advantage for weaker teams in games with a lower attendance is in line with the literature on the effect of the density of the crowd and its noise on referees' bias in favour of the home team. For example, using laboratory settings, Nevill et al. (2002) determined that crowd noise had a significant effect on the probability of a referee issuing a yellow card against a home team. Downward and Jones (2007) showed a positive relationship between the size of the crowd and the likelihood of a player receiving a yellow card in the English FA Cup. Similarly, Pettersson-Lidbom and Priks (2010) found a significant home bias of referees in games in which spectators were present compared to games with no spectators at all in the

[60] Note that the winning team receives three points, while the losing team gets no points. In case of a draw, each team gets one point.

[61] We also investigated the effect of playing on the two most frequent days, which are always Saturday and Sunday versus five non-frequent days (5/2 days split). We found that the third most frequent day is more similar to weekend games than to the other midweek games confirming our choice to refer to the 4/3 days split as to our preferred specification. For additional details, see the results section.

Italian Serie A. In addition, Ponzo and Scoppa (2018) showed a significantly larger number of cards against away teams in Serie A games between the teams from the same city that shared the same stadium. Finally, Page and Page (2010) found that the home advantage effect differs significantly among referees, and that this relationship is moderated by the size of the crowd. Therefore, a possible mediator of the difference in the home advantage of the underdog teams in games that take place on non-frequent days is lower crowd noise compared to games on frequent days.[62]

However, except for the case of Bundesliga 1, whose favourite teams that played at home suffer from the highest reduction of the crowd on non-frequent days (almost 16 percent smaller crowds), we found no difference in home advantage between different days when favourite teams play at home. As Krumer (2020) proposed, it is possible that the underdog teams depends more on crowd support or even pre-performance routine than the favourite teams because the latter are likely to win due to their higher ability regardless of home support or some other psychological factors. Therefore, underdog teams seem to lose more points in games with lower crowd density compared to the favourite teams.

Our results suggest that since some underdog teams play more home games on non-frequent days than other underdog teams, the current structure favours underdogs that play fewer home games on non-frequent days and favourites that play more away games on non-frequent days. To illustrate a possible relationship between an unbalanced schedule and the resulting monetary rewards, Krumer and Lechner (2018) gave an example of *SC Paderborn 07*, which was relegated from the Bundesliga 1 in the 2014/2015 season. This team played more home games on non-frequent days than its closest rival until the very last game in the relegation fight, *Hamburger SV*, which eventually remained in the top division. Moreover, one of these games was against *Hamburger SV*, which the latter won. According to Krumer and Lechner (2018), if *SC Paderborn 07* had survived in the Bundesliga 1, its additional revenue from TV alone would have been at least 10.3 million Euros (not counting all other revenues from ticketing, advertising, and so on).

The remainder of the paper is organized as follows. Section 4.2 describes the schedule of the different leagues. The data and some descriptive results are presented

[62] One additional explanation may be related to the usual pre-performance routine that has been found to be positively associated with performance in sports (see, for example, Lonsdale and Tam, 2008, and Mesagno and Mullane-Grant, 2010). We discuss this possibility in the results section.

in Section 4.3. Section 4.4 presents the empirical strategy. The results are contained in Section 4.5 and we offer concluding remarks in Section 4.6.

4.2 Description of the leagues' schedules

4.2.1 General structure of the leagues

While there are specific features for different leagues, the structure of all four leagues we investigate is largely similar. The leagues are organized as double round-robin tournaments, with each round consisting of $\frac{n}{2}$ games, where *n* is the number of teams in the league. In total, each team plays each other team twice, once at its home field in the first half of the season, and once away in the second half of the season (or vice versa). In total, every team has $2(n-1)$ games. In the French, Spanish, and English leagues, there are 20 teams, resulting in 38 games for each team. In the German Bundesliga 1, there are 18 teams, resulting in 34 games for each team. In addition, except for the English Premier League, the leagues have a winter break of several weeks without games.

The schedule of the leagues should also take into account international tournaments between nations, with the requirement to release participating players earlier and allow them a longer vacation. The main tournaments are the FIFA World Cup and the UEFA European Championship (held alternately every two years in June and July). Other tournaments that have the requirement to release players are the African Cup of Nations and the Asian Cup. Those take place during wintertime in parallel to the European leagues' matches. League games usually take place on weekends, but since there are not enough weekends in the season, some rounds take place on other days.

At the end of a season, the final table determines which teams participate in the following season's European club tournaments; these include the Champions League, which is the most prestigious club tournament in Europe, and the Europa League, which also yields significant monetary rewards. In addition, the two or three worst-ranked clubs are relegated to the second division, implying that the different outcomes have substantial financial consequences for the clubs.

Following Krumer and Lechner (2018), who investigated the effect of playing on Fridays, Saturdays, and Sundays (the three most frequently played days in the Bundesliga 1), we identified the three most frequently played days separately for each

league, as depicted in Table 4.1. Although less pronounced than in the Bundesliga 1, the three most frequent days for Ligue 1 and Premier League are quite clear. For the La Liga, however, the reduction in frequency between the third and fourth days is less significant. We follow the choice of Krumer and Lechner (2018) by taking the three most frequent days and need the remaining four days, which are defined as non-frequent, to have enough observations. Nevertheless, we will also investigate the effect of playing on the two most frequent days, which are always Saturday and Sunday, and compare between the results by discussing whether the third most frequent day is more similar to weekend games or to the other midweek games. In the following, we discuss special settings and uniqueness of schedule of the games that are described below for each league separately.

Table 4.1: Frequency of matches by days

Weekday	Ligue 1	La Liga	Premier League	Bundesliga
Monday	**11**	199	**204**	**1**
Tuesday	**57**	**71**	**182**	**72**
Wednesday	251	**185**	287	**96**
Thursday	**14**	**70**	**26**	**3**
Friday	**186**	**140**	**30**	275
Saturday	2188	1285	2152	1994
Sunday	1093	1850	919	619

Notes: The total number of matches on the specific weekdays, including matches that were excluded from the final analysis. Bold numbers represent the days of the "non-frequent" specification for each league.

4.2.2 The French Ligue 1

The three most frequently played match days are Saturday, Sunday, and Wednesday. The seasonal tournament in France takes place from August to the beginning of May. The top three teams advance to the Champions League (or for the Champions League playoffs). Teams in the fourth to sixth positions play in the Europa League (this may also depend on the outcome of an elimination French Cup tournament, called the *Coupe de France*). In addition, the two worst-ranked clubs are relegated to the lower division and the 18th-ranked team has to participate in a relegation playoff against the team that won the second division playoff for the right to play in the Ligue 1 in the following year.[63]

[63] Prior to 2016/17, the three bottom-ranked teams were directly relegated to the lower division.

4.2.3 The Spanish La Liga

The three most frequently played match days are Saturday, Sunday, and Monday. The seasonal tournament in Spain runs from the end of August or beginning of September until May of the following year. The top four teams advance to the Champions League (or the playoffs). Teams finishing fifth to seventh play in the Europa League (this may also depend on the outcome of an elimination Spanish Cup tournament, called the *Copa del Rey*). In addition, the three worst-ranked clubs are relegated to the lower division.

4.2.4 The German Bundesliga 1

The three most frequently played match days are Saturday, Sunday, and Friday. The seasonal tournament in Germany takes place from August to May. The top four teams advance to the Champions League (or to the playoffs). Teams finishing fifth to seventh play in the Europa League (this may also depend on the outcome of an elimination German Cup tournament, called the *DFB-Pokal*). In addition, the two worst-ranked clubs are relegated to the lower division and the 16th-ranked team must participate in the relegation playoffs against the third-ranked team in the Bundesliga 2 for the right to play in the Bundesliga 1 in the following year.[64]

4.2.5 The English Premier League

The three most frequently played match days are Saturday, Sunday, and Wednesday. The seasonal tournament in England takes place from August until May. This is the only one of the four leagues discussed here that does not have a long winter break. Several rounds take place during the Christmas holidays, usually involving local derbies to avoid fans having to travel long distances on those days (Kendall, 2008). The best-known round takes place on Boxing Day, which is a part of the Commonwealth tradition. We expect this to play a role in the scheduling process in the underlying period and account for this issue, as described in the next section.

The top four teams advance to the Champions League (or the playoffs). Teams finishing fifth to seventh play in the Europa League (this may also depend on the outcome of two elimination English Cup tournaments: the FA Cup and the League Cup). In addition, the three worst-ranked clubs are relegated to the lower division.

[64] The relegation playoff format was introduced in the 2008/2009 season. Prior to that, three teams were directly relegated to the Bundesliga 2.

Compared to the other three leagues, the Premier League has the highest amount of rescheduled games, because its clubs potentially have the highest number of games to play in their national cups (the FA Cup and the League Cup). The reason for this is that, in most stages of these competitions, a drawn match necessitated a repeated second game.[65] This partly interfered with the initial schedule proposed by the calendar committee.

4.3 Data and descriptive results

4.3.1 Database

We used data on four major European football leagues: the French Ligue 1, German Bundesliga 1, the Spanish La Liga, and the English Premier League.[66] For each of the leagues, we collected data on all the games starting from the start of the 2007/2008 season until the end of the 2016/2017 season. This represents a total of 14,460 games. However, we disregarded games in which a home team did not play at its usual home stadium. For example, Bayer Leverkusen from Germany did not play the second half of the 2008/2009 season at its home stadium due to reconstruction. RC Lens from France experienced a similar situation in the 2014/2015 season. In addition, Montpellier, Caen (both 2014/2015), and Lille (2007/2008 and 2008/2009) from France played some home games in alternative stadiums. We also removed matches in which one of the teams had already been relegated or had already won the championship title.[67] In addition, teams that play in the Champions League or Europa League may strategically adjust their squads in the domestic leagues games that take place just before or after the European cups (for example, they may save their best players before the European games to avoid a risk of injury or let them rest after). Therefore, we also removed games that involved teams playing just before or just after the continental competitions.[68] Finally, we also removed rescheduled games, since those may differ with regard to media attention as they are detached from the rest of

[65] In the case of a draw after the second game, overtime is played and, if needed, penalty shootouts determine the winner.

[66] See Appendix 4.D for the full list of sources.

[67] For example, in 2013/2014 season Bayern Munich from Germany had won the Bundesliga 1 title after 27 rounds. However, in the next three games they only gained one point out of nine and were accused of lacking motivation. See Kendall and Lenten (2017) for additional discussion on the usage of squads in remaining games after winning a title.

[68] See Rohde and Breuer (2017), who showed that teams adjust their efforts in domestic league just before or after games in European tournaments.

the matches.[69] Removing those games left 10,142 matches, 9,010 of which took place on frequent days and 1,132 on non-frequent days.

For every game, we collected information on the identity of teams, referees, exact day, attendance, distance between the cities, and the final score. We also used data from the *Transfermarkt* website to proxy the market value of each player of each team in every season. This data also includes personal information of each player, such as his age, height, and preferred foot. Finally, we have data on the dates of the beginning and the end of each coach's tenure, as well as data on the capacity of each stadium.

4.3.2 Definition of heterogeneity

There can be different types of heterogeneity in sports competition, such as home versus away or the favourite versus underdog teams.[70] We choose the favourite-underdog type of heterogeneity because it is intuitive that probabilities of winning (or the expected number of points) are largely driven by the differences in the teams' abilities, whereas the home-away factor plays a secondary role in increasing or decreasing the gap between the teams' probabilities of winning. While home advantage is a well-established phenomenon, the literature has largely neglected the heterogeneous effect for favourites versus underdogs. More importantly, beyond the above-mentioned intuition, standard economic theory predicts probabilities of winning based on contestants' innate abilities. For example, the Tullock contest (Tullock, 1980) is a well-known model in economic theory that has been applied in many fields, from political races (e.g. Klumpp and Polborn, 2006) to sports tournaments (e.g., Szymanski, 2003). The most popular versions of this model are lottery and all-pay contest. In the lottery version, a contestant with a lower effort still has a positive probability of winning, whereas an all-pay contest is fully discriminatory, where a contestant with a lower effort is certain to lose.

Now, assume a contest between two heterogeneous contestants 1 and 2, whose values (or the ability types) are $V_1 > V_2$, implying that contestant 1 is a stronger (or a higher-ranked) contestant. In the lottery model, contestants' efforts (x_i) are given by $x_1 = \frac{v_1^2 v_2}{(v_1+v_2)^2}$, and $x_2 = \frac{v_2^2 v_1}{(v_1+v_2)^2}$, and their probabilities of winning (p_i) are given by

[69] For additional discussion on rescheduled games, see Yi, Goossens, and Nobibon (2020).

[70] Other types of heterogeneities might be found in teams that replaced their coach, matches played on artificial versus natural grass, televised versus non-televised matches, traditional rivalry versus non-rivalry matches, etc.

$p_1 = \frac{v_1}{v_1+v_2}$ and $p_2 = \frac{v_2}{v_1+v_2}$. In the all-pay case, contestants' efforts are given by $x_1 = \frac{v_2}{2}$, and $x_2 = \frac{v_2^2}{2v_1}$, and their probabilities of winning are given by $p_1 = 1 - \frac{v_2}{2v_1}$ and $p_2 = \frac{v_2}{2v_1}$.[71] Therefore, the favourite-underdog type of heterogeneity is the one that fits the economic theory when investigating probabilities of winning (or the number of the gained points per game, in the case of soccer).

4.3.3 Variables and descriptive statistics

To estimate the effect of playing on non-frequent days on attendance and the number of gained points by the teams, we coded a dummy variable that equals 1 if a match was played on a non-frequent day in a certain league, and zero otherwise. We also used a rich set of variables that characterize team value and players' ability, game attendance, and the international and national schedule. In the following, we present some of the most important measures (a more comprehensive list of variables appears in Appendix 4.A).

Our approach is closely related to that of Krumer and Lechner (2018). Following their study, we used data on players' values from a popular soccer website, *Transfermarkt,* which are supposed to reflect teams' abilities. Since these values increase every season, we standardized them for each league and season so that they take the within-season variation into account.[72] The teams' values measure strongly correlates with teams' performance, suggesting that we have measured teams' abilities quite well.[73]

[71] For the lottery case, see, for example, Megidish and Sela (2014), who studied two-stage contests that are frequently used in sports competitions. For the all-pay case, see, for example, Krumer and Lechner (2017), who showed that in six out of seven possible cases in Olympic wrestling competitions, the all-pay mode predicted correctly the identity of a wrestler with a higher probability of winning.

[72] According to Bryson, Frick and Simmons (2013), the coverage of Transfermarkt is quite "impressive with information on 190,000 players across 330 football competitions" (p. 611). Players' values are estimated by industry experts and take into account salaries, signing fees, bonuses, and transfer fees. Franck and Nüesch (2012) found that the correlation between values evaluated by Transfermarkt and Kicker, another highly-respected sport magazine in Germany, is as high as 0.89.

[73] The results of the relevant regression analysis are available upon request from the authors.

Table 4.2: Descriptive statistics of the main variables

Variable	Favourite playing home		Underdog playing home	
	Frequent days	Nonfrequent days	Frequent days	Nonfrequent days
Favourite points	1.93 (1.25)	1.98 (1.25)	1.33 (1.29)	1.58 (1.30)
Underdog points	0.82 (1.13)	0.78 (1.13)	1.39 (1.30)	1.14 (1.25)
Favourite win	0.56	0.58	0.35	0.43
Underdog win	0.19	0.18	0.37	0.29
Visitors	32,743 (17,533)	34,583 (18,871)	24,031 (12,640)	24,943 (12,584)
Stadium capacity	40,825 (17,510)	44,864 (19,521)	29,924 (12,974)	31,410 (13,588)
Share of capacity	0.78 (0.19)	0.75 (0.20)	0.79 (0.18)	0.79 (0.18)
Ln(Attendance)	10.24 (0.58)	10.29 (0.60)	9.96 (0.53)	10.01 (0.50)
Distance (in km)	487.5 (342)	501.6 (365)	485.1 (338)	461.5 (402)
Fav. stand. team value	0.22 (1.01)	0.46 (1.14)	0.23 (1.00)	0.51 (1.20)
Und. stand. team value	-0.58 (0.35)	-0.51 (0.33)	-0.57 (0.36)	-0.52 (0.39)
Fav. ratio of top 3 to ranked 9-11 most valuable players	2.41 (0.71)	2.49 (0.82)	2.41 (0.73)	2.47 (0.74)
Und. ratio of top 3 to ranked 9-11 most valuable players	2.33 (0.75)	2.40 (0.82)	2.33 (0.75)	2.38 (0.75)
Fav. plays in Europa or Champions League	0.33	0.46	0.34	0.44
Und. plays in Europa or Champions League	0.09	0.13	0.08	0.11
African Cup of Nations	0.07	0.02	0.07	0.04
Asian Cup	0.02	0.01	0.02	0.01
Public holiday	0.04	0.09	0.04	0.09
2 months before European Championship	0.06	0.08	0.06	0.07
2 months after European Championship	0.04	0.07	0.04	0.05
2 months before World Cup	0.04	0.04	0.04	0.03
2 months after World Cup	0.03	0.05	0.03	0.04
Observations	4,462	569	4,548	563

Notes: This table presents average values and standard deviations for the three (four) most (non-)frequent days for the main variables from all leagues combined. Descriptive statistics for all available variables appear in Appendix 4.A.

For each game, the favourite is defined as the team with the higher standardized *Transfermarkt* value and the underdog is the team with the lower standardized *Transfermarkt* value. Unlike with betting odds, where favourite and underdog can be a function of the day of the week and the home advantage, the *Transfermarkt* values are determined without considering those factors. Therefore, these definitions are exogenous.

Following Krumer (2020), we divided the data into games that take place at the favourite's and the underdog's home fields. In Table 4.2, which presents descriptive statistics for the pooled data, we can see that when a favourite plays at its home stadium, the average number of points it gains on frequent days is 1.93. When the game is on a non-frequent day, the favourite team gains a very similar number of points (1.98). However, when an underdog team hosts the game, it gains an average of 1.39 points on frequent days and 1.14 points on non-frequent days, suggesting a lower home advantage on non-frequent days for the underdog teams only.

Tables 4.3 and 4.4 presents the descriptive statistics divided into four different leagues, where we can see a similar pattern. However, we also observe that when an underdog team plays at home, in all leagues except La Liga, there are much stronger favourites on non-frequent days compared to frequent days. The direction is the same in La Liga, but the difference between standardized values of favourites on frequent and non-frequent days is less pronounced. This descriptive evidence indicates that there is non-random selection into treatment; that is, non-frequent days. We will discuss how to solve this issue in the next section.

Table 4.3: Descriptive statistics, Ligue 1 & La Liga

Variable	Favourite playing home		Underdog playing home	
	Frequent days	Nonfrequent days	Frequent days	Nonfrequent days
	Ligue 1			
Favourite points	1.91 (1.23)	1.90 (1.22)	1.31 (1.29)	1.69 (1.27)
Underdog points	0.81 (1.09)	0.79 (1.07)	1.41 (1.30)	1.00 (1.18)
Ln(Attendance)	9.86 (0.57)	10.05 (0.59)	9.60 (0.44)	9.82 (0.41)
Distance (in km)	657.2 (335)	662.1 (308)	658.8 (331)	608.2 (339)
Fav. stand. team value	0.31 (1.10)	0.77 (1.31)	0.30 (1.09)	1.13 (1.52)
Und. stand. team value	-0.54 (0.36)	-0.41 (0.29)	-0.54 (0.37)	-0.43 (0.24)
Fav. plays in Europa or Champions League	0.27	0.38	0.29	0.29
Und. plays in Europa or Champions League	0.08	0.11	0.07	0.05
African Cup of Nations	0.06	0.04	0.05	0.04
Public holiday	0.03	0.05	0.03	0.02
Weekdays	Wed, Sat, Sun	Mon, Tue, Thu, Fri	Wed, Sat, Sun	Mon, Tue, Thu, Fri
Observations	1365	111	1397	92
	La Liga			
Favourite points	2.01 (1.25)	2.07 (1.23)	1.31 (1.30)	1.45 (1.28)
Underdog points	0.78 (1.14)	0.72 (1.10)	1.43 (1.31)	1.25 (1.26)
Ln(Attendance)	10.15 (0.61)	10.08 (0.64)	9.76 (0.48)	9.72 (0.48)
Distance (in km)	631.2 (401)	659.0 (407)	621.5 (389)	688.2 (504)
Fav. stand. team value	0.19 (1.01)	0.23 (1.07)	0.19 (1.01)	0.22 (1.08)
Und. stand. team value	-0.52 (0.26)	-0.50 (0.24)	-0.52 (0.25)	-0.51 (0.23)
Fav. plays in Europa or Champions League	0.33	0.38	0.32	0.39
Und. plays in Europa or Champions League	0.09	0.12	0.08	0.09
African Cup of Nations	0.07	0.00	0.09	0.02
Public holidays	0.04	0.01	0.05	0.02
Weekdays	Mon, Sat, Sun	Tue, Wed, Thu, Fri	Mon, Sat, Sun	Tue, Wed, Thu, Fri
Observations	1,078	226	1,105	195

Notes: This table presents average values and standard errors for the main variables for each league separately. For each league, we also provide the used frequent and non-frequent weekdays.

Table 4.4: Descriptive statistics, Premier League & Bundesliga

Variable	Favourite playing home		Underdog playing home	
	Frequent days	Nonfrequent days	Frequent days	Nonfrequent days
	Premier League			
Favourite points	1.97 (1.24)	2.03 (1.22)	1.39 (1.29)	1.60 (1.29)
Underdog points	0.79 (1.12)	0.72 (1.07)	1.32 (1.28)	1.11 (1.24)
Ln(Attendance)	10.49 (0.36)	10.60 (0.37)	10.20 (0.32)	10.21 (0.36)
Distance (in km)	235.8 (133)	203.3 (131)	233.6 (133)	208.0 (127)
Fav. stand. team value	0.21 (0.99)	0.54 (1.00)	0.25 (0.98)	0.55 (1.10)
Und. stand. team value	-0.65 (0.41)	-0.60 (0.41)	-0.64 (0.42)	-0.56 (0.55)
Fav. plays in Europa or Champions League	0.39	0.52	0.40	0.51
Und. plays in Europa or Champions League	0.09	0.14	0.10	0.15
African Cup of Nations	0.06	0.06	0.06	0.07
Public holiday	0.04	0.28	0.04	0.24
Weekdays	Wed, Sat, Sun	Mon, Tue, Thu, Fri	Wed, Sat, Sun	Mon, Tue, Thu, Fri
Observations	1111	153	1124	190
	Bundesliga			
Favourite points	1.83 (1.28)	1.73 (1.37)	1.33 (1.30)	1.70 (1.38)
Underdog points	0.91 (1.18)	1.09 (1.32)	1.40 (1.30)	1.14 (1.35)
Ln(Attendance)	10.62 (0.38)	10.65 (0.45)	10.42 (0.41)	10.40 (0.42)
Distance (in km)	369.8 (183)	403.3 (172)	365.0 (185)	350.4 (185)
Fav. stand. team value	0.12 (0.87)	0.54 (1.23)	0.14 (0.89)	0.41 (1.04)
Und. stand. team value	-0.60 (0.35)	-0.51 (0.41)	-0.60 (0.36)	-0.52 (0.38)
Fav. plays in Europa or Champions League	0.35	0.66	0.36	0.56
Und. plays in Europa or Champions League	0.09	0.17	0.08	0.11
African Cup of Nations	0.07	0.00	0.07	0.00
Public holidays	0.02	0.00	0.03	0.00
Weekdays	Fri, Sat, Sun	Mon, Tue, Wed, Thu	Fri, Sat, Sun	Mon, Tue, Wed, Thu
Observations	908	79	922	86

Notes: This table presents average values and standard errors for the main variables for each league separately. For each league, we also provide the used frequent and non-frequent weekdays.

The players' values are used to create additional measures such as the distribution of values between and within teams. More specifically, for each team we compute the standard deviation of players' values − the Herfindahl-Hirschman Index (HHI) − which is defined as the sum of the squares of the values shares of each player

within the team. We also created other within-team inequality-related variables such as the ratio of different players' values according to their ranking order in the team. For example, one measure is the ratio between the top three players to players ranked 9−11 according to their values within a team.[74] In addition to players' values, we also use several other variables that may reflect the level of ability, such as a dummy variable for a team's first season in the top division after being promoted from the lower division, whether a team dismissed its coach during a season, and the age of the coach.[75] We also use data on the size of the squad, share of foreign players in the squad, height of the players, share of left-footed players, age of oldest/youngest players, etc.

Based on the large body of the literature on the effect that the crowd has on home advantage, we created a measure to reflect the attendance in a match. Our first measure − attendance as share of the capacity of the stadium − is the ratio between the number of viewers in a match and the maximal possible capacity of the respective stadium. We also applied a different measure of attendance, namely a natural logarithm of attendance (*Ln(Attendance))* that is also used in the literature on attendance demand (Buraimo and Simmons, 2015; Buraimo, Tena and de la Piedra, 2018; Krumer, 2020). In addition, following the studies of Boyko, Boyko, and Boyko (2007) and Page and Page (2010), who described the existence of individual differences among referees in terms of the home advantage in different soccer leagues, we created dummy variables for each individual referee in our data. Further, there is also information about the distance between cities, in kilometres for the shortest traveling distance.

We also obtained information on other schedule-related variables in international competitions, such as two pre- and post-World Cup and European Championships months, as well as the months in which the African Cups of Nations and Asian Cup took place. We also took different months of the season and public holidays into account.

[74] See Coates, Frick, and Jewell (2016) for discussion on the relationship between players' inequality in salaries and teams' performance.

[75] See Tena and Forrest (2007), and Flores, Forrest, and Tena (2012) for discussions on the effects of coach dismissals on team performance.

4.4 Empirical strategy

4.4.1 Selection into treatment

We studied the effect of playing on a non-frequent day compared to a frequent day on the performance of a team. Here, the challenge for identifying a causal effect lies in the non-random determination of the teams that play at home on non-frequent days. In order to obtain an unbiased causal effect, it is essential to disentangle the effect coming with the selection from the effect caused by playing on non-frequent days. In other words, there is the need to take selection effects into account. Decisions regarding which teams play on which days are made by the calendar committees of the respective leagues and might be driven by teams' characteristics and other schedule-related features, such as public holidays, international breaks, TV broadcasters' interests, European association tournaments, etc..[76]

The rich database presented in the previous section enabled us to opt for a selection-on-observables approach; that is, controlling for the reasons for the deviations from random treatment assignment. Having information on teams and game characteristics, European cups scheduling, national teams' tournaments, etc., enabled us to capture all confounding factors related to team, location, and timing to create a quasi-experimental setup. This allows us to identify the causal effect of playing on non-frequent days on performance if there are no unobserved characteristics that simultaneously affect both the probability of playing on a non-frequent day and the outcome.

An important issue that is worth a separate discussion relates to broadcasting issues, which may affect the allocation of games into frequent and non-frequent days. For instance, a broadcaster may influence the decision to allocate strategically more or less attractive games into different days. However, it is problematic to have a dummy variable of whether a game was broadcast or not since it is potentially part of our treatment. We try to solve this issue by capturing the selection effects, which might be partially due to broadcasters' interests, by controlling for a huge range of variables. Among those are the approximated strength of a team and its popularity (in the form of teams' values and teams' dummies, among many other things), as well as controls for other obligations of the teams, so as whether they play in the European tournaments and are therefore shifted to non-frequent days. Therefore, if a broadcaster has a

[76] See Goossens and Spieksma (2009) for examples of different requirements the calendar committee must consider.

preference for certain teams to play on frequent or non-frequent days (or even if a team itself has such a preference), these teams' dummy variable will capture this potential issue. Thus, it is very likely that selection into treatment is captured by our very large set of potential control variables satisfying the conditional independence assumption needed for credibly estimating a causal effect in our study.

Finally, it is also important to note that the share of capacity (or *Ln(Attendance)*) is an endogenous variable since it is an outcome variable that depends on the day of the game. Therefore, we do not include it as a covariate in our estimation, but only use it as an outcome variable.

4.4.2 Estimation

4.4.2.1 Estimator

In order to have a flexible approach and overcome the restrictive assumptions of classical statistical linear models, we used a statistical matching approach. More specifically, we applied the radius-matching-on-the-propensity score estimator with bias adjustment (Lechner, Miquel and Wunsch, 2011).[77] Not only was this estimator found to be very competitive among a range of matching-type estimators, but Huber, Lechner and Wunsch (2013) also showed its superior finite sample and robustness properties in a large-scale Empirical Monte Carlo Study. This estimator combines the features of distance-weighted radius matching with a bias adjustment, which removes potential biases due to mismatches.[78] Control observations, which are close to the treated unit in terms of the confounding influences, can be compared to the latter to obtain the treatment effect as if treated and control units were in an experimental setting. Therefore, it is crucial to capture all confounding influences; we explain how we do this in more detail below.

4.4.2.2 Propensity score

The propensity score, which is the probability of playing on a non-frequent day, condenses the information from all relevant confounding variables to a one-dimensional score, determining which observations are similar in terms of confounding influences. In their pioneering work, Rosenbaum and Rubin (1983) showed that

[77] The variance is estimated as weight-based variance as described in Huber, Lechner and Steinmayr (2015), which Bodory et al. (2018) showed to lead to conservative standard errors.

[78] Distance weighting leads to a weighting of non-treated observations within the radius inversely proportional to their distance to the respective treated unit.

controlling for the propensity score removes selection bias. Therefore, treated and non-treated observations with similar propensity scores are compared to each other in the matching estimator.

If the exact relation of confounding variables and the treatment assignment is known, the variables to include in the propensity score estimation can be specified ad-hoc. In our case, we have a set of 385 potentially confounding variables, in addition to the referees and team dummies, as described in the previous section. Despite prior knowledge about the selection process, we cannot specify ad-hoc exactly which of the many potential confounders to use in the propensity score estimation. Further, including everything would lead to an unfeasible estimation, less precise or instable estimates. Therefore, for the specification we rely on a machine learning algorithm.

Using machine learning for causal inference is not a trivial exercise and the literature is still under development. Since those algorithms are designed for prediction and not for doing inference in treatment effects estimation, we follow the approach of Belloni, Chernozhukov, and Hansen (2014) to make machine learning algorithms useful in this setup. Those authors suggested using the LASSO procedure developed by Tibshirani (1996) as a variable selection tool, twice.[79,80] In the first step, we selected a set of variables confounding the treatment. In the second step, we selected those variables correlated with the respective outcome.[81] The reason for this double-selection procedure, as opposed to only looking at the treatment selection equation, is to additionally capture variables that are highly correlated to the outcome and mildly related to the treatment selection. The same line of argumentation holds for only looking at the outcome equation. Ignoring those kinds of variables would lead to potentially biased results, as described in Belloni et al. (2014). The union of variables selected by the two separate LASSO procedures is our final set of variables for the propensity score estimation. We repeat this selection procedure for all of the estimations presented in the next section.

[79] The LASSO procedure is a shrinking estimator, which works like an OLS estimator with penalized coefficients. Penalizing the coefficients leads to variables selection as the coefficients of not too informative covariates are forced to zero.

[80] Goller et al. (2020) compared different (machine learning and "classical" Probit) estimation procedures for the propensity score in matching estimation and found the LASSO delivered the most credible results in a setup, which is comparable to ours, with many potentially confounding variables and a low share of treated units.

[81] Using the LASSO method requires a penalty term, which is data-driven determined using 10-fold cross-validation. For the current analysis, we chose the penalty term that minimized the mean squared error.

4.5 Results

4.5.1 Pooled data

First, our aim was to study whether playing on non-frequent days has an effect on performance when using the data on all the leagues together. To accomplish this goal, we first estimated the propensity score, based on variables that were chosen in the double-selection LASSO procedure described in the previous section. It is important to note that the purpose of the propensity score estimation is purely technical, to allow the easy purging of the results from the selection effects. Therefore, the respective marginal effects of the propensity score estimation cannot be interpreted in a causal sense, but rather in their contribution to the probability of being treated. Appendix 4.B provides an example of one of the propensity score estimations we had to execute, which is the propensity score estimation for the number of favourite's points. Generally, as is already apparent from Tables 4.2 and 4.3, selection effects are driven by team values as well as by schedule-related features such as public holidays and international tournaments. In addition, several individual referees and teams were picked by the double-selection LASSO procedure. While Appendix 4.B only presents the results of the propensity score estimation for the number of points of a favourite team, a separate propensity score estimation for each matching estimation presented in the paper is available upon request.

In addition, we wish to show the sensitivity of our results to the presence or absence of referees' dummies. Such an importance is driven by findings of Boyko, Boyko, and Boyko (2007) and Page and Page (2010), who showed that some referees might be more affected by the home crowd and therefore may serve as possible mediator of a difference in a home advantage. Therefore, we present our results with and without the inclusion of referees' dummies. More specifically, in the specification that includes referees' dummies, we only use the dummies that were chosen in the double-selection LASSO procedure.

Table 4.5: Levels and effects of playing on four non-frequent days, all data

Dependent Variables	Exp. value on NF days	Exp. value on frequent days	Effect of playing on NF days	p-value	Common Support (in %)
	(1)	(2)	(3)	(4)	(5)
Panel A: Incl. Referee Dummies					
		Favourite team playing home			
Fav. Points	1.911	1.940	-0.029	0.64	98.5
Und. Points	0.826	0.809	0.017	0.76	96.9
Share of capacity	0.745	0.783	-0.038***	0.00	93.3
Ln(Attendance)	10.185	10.258	-0.073**	0.02	97.5
		Underdog team playing home			
Fav. Points	1.499	1.329	0.170**	0.01	99.9
Und. Points	1.251	1.376	-0.125***	0.04	99.4
Share of capacity	0.775	0.793	-0.018**	0.05	99.9
Ln(Attendance)	9.903	9.961	-0.058**	0.02	99.8
Panel B: Excl. Referee Dummies					
		Favourite team playing home			
Fav. Points	1.916	1.935	-0.019	0.76	98.5
Und. Points	0.838	0.801	0.038	0.50	97.1
Share of capacity	0.745	0.782	-0.037***	0.00	93.3
Ln(Attendance)	10.213	10.259	-0.046*	0.10	97.8
		Underdog team playing home			
Fav. Points	1.494	1.336	0.159**	0.01	99.9
Und. Points	1.244	1.376	-0.132**	0.04	99.5
Share of capacity	0.780	0.795	-0.015*	0.08	99.9
Ln(Attendance)	9.926	9.965	-0.038*	0.10	99.9

Notes: The results represent all of the data. Columns (1) and (2) represent the expected values for four most non-frequent (NF) and three most frequent days, respectively. Columns (3) and (4) report the average treatment effect and the respective p-value. Standard errors are calculated as weight-based standard errors and clustered at the season per league level. Column (5) states the share of observations in common support in the radius matching. ** and *** denote the 5% and 1% significance levels, respectively.

Panel A of Table 4.5 shows the effect of playing on four non-frequent days compared to the three frequent days by pooling all the leagues together. We can see that when a favourite team plays at home, the effect of playing on non-frequent days on the number of points is very close to zero (-0.029) and highly insignificant (p-value=0.64). However, when testing the effect of playing on non-frequent days on the number of points of the favourite when an underdog plays at home, we find that a favourite gains 0.17 points significantly more on non-frequent days than on the

frequent days. An underdog gains about 0.13 points less when hosting a game on non-frequent compared to frequent days, making the difference between favourite and underdog about 0.30 points. The share of capacity on non-frequent days is 3.8 percentage points less when a favourite team plays at home and 1.8 percentage points less when an underdog team hosts the game. Similarly, *Ln(Attendance)* is significantly lower on non-frequent days. Panel B of Table 4.5, where we present the results without referees' dummies shows a similar pattern for all the outcome variables.

Table 4.6: Effects of different definitions of non-frequent days, all data

	Incl. referee dummies		Excl. referee dummies	
Days	Fav. Points	Und. Points	Fav. Points	Und. Points
	Favourite team playing home			
4-7 vs. 1-3	-0.029 (0.64)	0.017 (0.76)	-0.019 (0.76)	0.038 (0.50)
3-7 vs. 1,2	-0.015 (0.76)	0.000 (0.99)	0.008 (0.87)	0.010 (0.87)
4-7 vs. 1,2	-0.007 (0.91)	-0.002 (0.97)	-0.043 (0.50)	0.031 (0.59)
4-7 vs. 3	-0.085 (0.35)	0.041 (0.61)	-0.106 (0.35)	0.074 (0.37)
3 vs. 1,2	0.018 (0.79)	-0.005 (0.94)	0.093 (0.18)	-0.033 (0.59)
	Underdog team playing home			
4-7 vs. 1-3	0.170** (0.01)	-0.125** (0.04)	0.159** (0.01)	-0.132** (0.04)
3-7 vs. 1,2	0.137** (0.01)	-0.116** (0.02)	0.111** (0.04)	-0.104** (0.05)
4-7 vs. 1,2	0.146** (0.01)	-0.159** (0.01)	0.169** (0.01)	-0.140** (0.01)
4-7 vs. 3	0.215** (0.03)	-0.256***(0.00)	0.199** (0.04)	-0.216** (0.03)
3 vs. 1,2	0.051 (0.50)	-0.040 (0.59)	0.085 (0.26)	-0.063 (0.40)

Notes: The results represent the effects of playing on different definitions of non-frequent days for all the data. The most frequent day is defined as day 1 and the least frequent day is defined as day 7. P-values of the effects are presented in parentheses. Standard errors are calculated as weight-based standard errors and clustered at the seasonal level. *, **, and *** denote the 10%, 5% and 1% significance levels, respectively.

One possible concern is that our results are driven by the definition of the frequent days. As presented in Table 1, Saturdays and Sundays are the most frequently played days in all the leagues. Therefore, an alternative comparison could be between the five midweek days to the two weekend days. In Table 4.6, we present the effect of playing on the five non-frequent days compared to the two frequent days (defined as 3-7 vs. 1,2 in Table 4.6). We can see that, as previously, there is no significant effect on teams' points when a favourite team plays at home. However, when an underdog team plays at home, the effect on teams' points is lower than in the case of the 4/3 days split as presented in Table 4.5 (also defined as 4-7 vs. 1,2,3 in Table 4.6). In addition, we excluded the third frequent day and compared the effect of playing on the four remaining weekdays compared to the two weekend days (defined as 4-7 vs. 1,2 in Table 4.6). In the case of the underdog's home advantage, we can see that the effects are much closer to the 4/3 days split rather than to the 5/2 days split. We also show that

the third most frequent day is significantly different from days 4-7 in terms of teams' points (defined as 4-7 vs. 3 in Table 4.6), namely underdog teams that play at home on the four non-frequent days gain significantly less points than when they play at home on the third most frequent day. Finally, we show that there is no significant difference in terms of points between the two weekend days and the third frequent day (defined as 3 vs. 1,2 in Table 4.6). These results suggest that the third frequent day is more similar to the weekend days rather than to the other weekdays.

4.5.2 Individual leagues

To investigate whether our results are driven by pooling the data, in Tables 4.7 and 4.8, we present the results for each league separately by using the 4/3 days split. When including referees' dummies, as presented in Panel A, we can see that, for all the leagues and for both cases of home advantage, we find a lower attendance as a share of capacity and a lower *Ln(Attendance)* on non-frequent days compared to frequent days. When a favourite team hosts the game, the effect of playing on non-frequent days on share of capacity ranges from 1.6 percentage points less in English Premier League to 7.4 percentage points less in German Bundesliga 1. When looking at *Ln(Attendance)*, we see that a reduction in attendance ranges from 6.5 percent lower attendance in English Premier League to 15.9 percent lower attendance in German Bundesliga 1 on non-frequent days. When an underdog team hosts the game, the effect of playing on non-frequent days on the share of capacity ranges from 3.4 percentage points less in the English Premier League to 5.9 percentage points less in Spanish La Liga. When looking at *Ln(Attendance)*, we see that reduction in attendance ranges from 6.0 percent lower attendance in English Premier League and French Ligue 1 to 10.2 percent lower attendance in German Bundesliga 1. Panel B of Table 4.8 shows a very similar pattern. These results replicate the finding of Krumer and Lechner (2018), who also found a lower attendance in the German Bundesliga 1.

When using referees dummies, as presented in Panel A of Table 4.7, in all the leagues apart from Bundesliga 1, playing on non-frequent days has no effect on the home advantage of the favourite teams. However, in all the four leagues there is a reduced home advantage on non-frequent days for the underdog teams. Our results suggest that the difference in the number of points between the favourite and the underdog teams, when the game takes place on non-frequent days compared to frequent days, is 0.49 in Ligue 1, 0.43 in La Liga, 0.42 in Premier League, and 0.63 in Bundesliga 1. This difference is quite large given that, in our dataset, a favourite with home advantage gains on average about 1.1 points more than the underdog. In addition,

in a tight league, one point could make the difference between relegation and survival or between qualification to the UEFA Champions League and the less prestigious UEFA Europa League.

Table 4.7: Effects of playing on four non-frequent days, leagues; Panel A

Outcomes	Ligue 1	La Liga	Premier L.	Bundesliga	Wald-test p-value
Panel A: Incl. referee dummies					
			Favourite team playing home		
Fav. Points	-0.062	0.015	-0.093	-0.276	0.55
	(0.62)	(0.89)	(0.43)	(0.10)	
Und. Points	-0.064	0.135	-0.046	0.310*	0.17
	(0.57)	(0.15)	(0.65)	(0.06)	
Share of capacity	-0.042*	-0.066***	-0.016*	-0.074***	0.00
	(0.10)	(0.00)	(0.06)	(0.00)	
Ln(Attendance)	-0.102*	-0.116**	-0.065*	-0.159***	0.50
	(0.07)	(0.03)	(0.06)	(0.00)	
			Underdog team playing home		
Fav. Points	0.228*	0.242**	0.225**	0.329**	0.96
	(0.07)	(0.02)	(0.05)	(0.05)	
Und. Points	-0.266**	-0.188*	-0.197*	-0.303*	0.91
	(0.05)	(0.06)	(0.07)	(0.06)	
Share of capacity	-0.036**	-0.059***	-0.034***	-0.036**	0.44
	(0.05)	(0.00)	(0.00)	(0.02)	
Ln(Attendance)	-0.060*	-0.073*	-0.060**	-0.102**	0.90
	(0.09)	(0.07)	(0.03)	(0.04)	

Notes: The results represent the effects of playing on four non-frequent days compared to the three most frequent days for each league separately. P-values of the effects are presented in parentheses. Standard errors are calculated as weight-based standard errors and clustered at the seasonal level. *, **, and *** denote the 10%, 5%, and 1% significance levels, respectively. Common support for each of the matching estimations is at least 88.1% (Ligue 1), 92% (La Liga), 91.4% (Premier League), and 83.7% (Bundesliga).

Table 4.8: Effects of playing on four non-frequent days, leagues; Panel B

Outcomes	Ligue 1	La Liga	Premier L.	Bundesliga	Wald-test p-value
Panel B: Excl. referee dummies					
		Favourite team playing home			
Fav. Points	0.118	0.021	0.019	-0.230	0.46
	(0.33)	(0.85)	(0.87)	(0.20)	
Und. Points	-0.088	0.030	-0.053	0.496***	0.02
	(0.44)	(0.74)	(0.60)	(0.00)	
Share of capacity	-0.043**	-0.065***	-0.016*	-0.081***	0.00
	(0.05)	(0.00)	(0.09)	(0.00)	
Ln(Attendance)	-0.121**	-0.158***	-0.063*	-0.097*	0.44
	(0.03)	(0.00)	(0.05)	(0.08)	
		Underdog team playing home			
Fav. Points	0.294*	0.204**	0.078	0.410**	0.36
	(0.07)	(0.04)	(0.47)	(0.01)	
Und. Points	-0.407***	-0.310***	0.032	-0.305**	0.07
	(0.00)	(0.00)	(0.75)	(0.05)	
Share of capacity	-0.034*	-0.041***	-0.034***	-0.059***	0.62
	(0.05)	(0.00)	(0.00)	(0.00)	
Ln(Attendance	-0.065	-0.088**	-0.050*	-0.103**	0.75
	(0.12)	(0.02)	(0.08)	(0.03)	

Notes: The results represent the effects of playing on four non-frequent days compared to the three most frequent days for each league separately. P-values of the effects are presented in parentheses. Standard errors are calculated as weight-based standard errors and clustered at the seasonal level. *, **, and *** denote the 10%, 5%, and 1% significance levels, respectively. Common support for each of the matching estimations is at least 88.1% (Ligue 1), 92% (La Liga), 91.4% (Premier League), and 83.7% (Bundesliga).

Interestingly, when excluding referees' dummies, as presented in Panel B of Table 4.8, we find no significant effect of playing on non-frequent days in the Premier League. This result may suggest that certain referees may serve as possible mediators of a reduced home advantage of an underdog team on non-frequent days, which is in line with the findings in Boyko, Boyko, and Boyko (2007) and Page and Page (2010), who described the existence of individual differences among referees in terms of the home advantage. This is also in line with the recent speculations that some referees have a reputation for being biased in favour or against specific teams in the English Premier League.[82]

Furthermore, we conducted a WALD test for the equality of each estimate obtained in the analyses for the single leagues. More specifically, we tested the

[82] See, for example, https://www.sportskeeda.com/slideshow/football-most-biased-referees-top-teams-premier-league. Last accessed on 02.11.2019.

hypothesis that the effects are equal for all the leagues and reported the p-values from the resulting chi square test statistic in the last column of Tables 4.7 and 4.8. We found that the effects are not statistically different between the leagues for the favourites and underdogs points (except for underdog points without referees' dummies), or for the *Ln(Attendance).* Although the only robust significant difference between the leagues is for the share of capacity in the case when a favourite team plays at home, this emphasizes the need to conduct the analysis for each league separately.

4.5.3 Robustness tests

Despite our findings presented in Table 4.6 that on the aggregate level, the third frequent day is likely to resemble the two weekend days, we nevertheless conduct separate analyses for the 5/2 days split for each league separately. The results in Table 4.11 in the appendix show a similar pattern as with 4/3 days split presented in Tables 4.7 and 4.8, though not always significant at conventional level. More specifically, we find that the share of capacity in case when an underdog plays at home is always negative and significant for all the leagues with and without referees' dummies. These results are in line with the findings of Buraimo and Simmons (2015), Buraimo (2008), and Forrest and Simmons (2006), all of whom reported that weekend games attract larger crowds and larger TV ratings in the English Premier League. However, *Ln(Attendance)* is sensitive to inclusion of referees' dummies in Ligue 1 and also not significant in three out of four cases in the Bundesliga 1, but confirms the general picture. In addition, we can see a higher number of points for the favourite team (lower number of points for the underdog team) when underdog team plays at home, however with a lower magnitude compared to the results in Tables 4.7 and 4.8. When including referees' dummies, the results are significant for La Liga and Bundesliga 1, and are in similar direction as with the 4/3 days split, but with p-val in the range between 0.14-0.17 in Ligue 1 and Premier League. When excluding the referees' dummies, as presented in Panel B of Table 4.11, the Premier League has a zero effect, which is in line with the results presented in Panel B of Table 4.8. The results in Ligue 1 and La Liga are just insignificant with p-val in the range between 0.11-0.14.

Finally, as previously, we excluded the third frequent day and tested the effect of playing on four non-frequent days versus the two most frequently played days. Table 4.12 in the appendix presents the results, where we can see that the share of capacity is always negative and significant, and *Ln(Attendance)* is not significant for Ligue 1 without referees' dummies and for one case in the Premier League. When excluding the third most frequent day in Bundesliga 1 (Friday), the *Ln(Attendance)* became

significant again, suggesting that Friday is much more similar to Saturday and Sunday than to the other weekdays. Finally, the effects on teams' points when the underdog team plays at home are much larger and more significant than in the 5/2 days split, presented in Table 4.11 in the appendix. As already discussed, these results suggest again that the third most frequent day is more associated with the weekend games rather than with midweek games, and therefore we refer to the 4/3 days split as to our preferred specification.

4.5.4 Discussion

Our finding regarding the lower home advantage of the underdog on non-frequent days is in line with the literature on the effect of the density of the crowd and its noise on referees' bias in favour of the home team (Downward and Jones, 2007; Nevill et al., 2002; Page and Page, 2010; Pettersson-Lidbom and Priks, 2010). Therefore, a possible mediator of the difference in the home advantage of the underdog teams in games that take place on non-frequent days is lower crowd noise compared to games on frequent days. Crowd noise might be lower on non-frequent days not only due to lower attendance, but also magnified by a lower amount of alcohol consumption on non-frequent days. In addition, if a crowd is less passionate and supportive on non-frequent days, this may also negatively affect players' motivation, which may result in lower performance.

Another possible explanation to the lower home advantage of the underdog on non-frequent days relates to the literature on pre-performance routine. Previous research has suggested that having the usual pre-performance routine has a positive relationship with performance in sports. This may be due to lowered anxiety, increased task-relevant focus or reduced negative self-assessment. For example, Mesagno and Mullane-Grant (2010) showed that usual pre-performance routine helped Australian football players performing better under increased pressure. Similarly, Lonsdale and Tam (2008) demonstrated that NBA basketball players' free throw shooting was better if they followed their dominant behavioural sequence.[83] In that spirit it is possible that players of the home team have a different pre-performance routine on non-frequent day (for example, different within family routine, different traffic before the game, different media coverage, etc.), which may negatively affect their anxiety and self-efficacy levels. In that regard, away teams are less exposed to negative effects of

[83] For additional references on the effect of pre-performance routine on performance, see a recent study of Wergin et al. (2020).

deviating from their routine, since they are not in their usual settings anyway, and therefore prepare to that game in the usual away game routine.

One possible explanation for the finding that the home advantage of the favourite teams is not affected by the day of the game is that these teams are likely to win because of their higher abilities, regardless of home support or pre-performance routine. This would suggest that underdog teams depend more on crowd support or some other psychological factors than favourite teams. The negative effect on the home advantage of the favourite in Bundesliga 1 may be explained by the fact that this league differs from the other leagues as it has 18 teams, not 20. In addition, according to Yi, Goossens, and Nobibon (2020), Bundesliga 1 has the most reliable league schedule, with the lowest number of rescheduled games and a higher quality realized schedule than the other leagues. Indeed, as Table 4.1 shows, Bundesliga 1 has the lowest amount of games on non-frequent days. In addition, as already mentioned, Bundesliga 1 has the highest reduction of attendance among all the leagues in the 4/3 days split when a favourite team plays at home (15.9 percent less attendance). This suggests that fans are not used to attending the stadium on non-frequent days and also that the players of the favourite teams may be less motivated to play in front of smaller crowds that are also likely to be less supportive on non-frequent days.

4.6 Conclusion

According to Wright (2014), the main objective of his survey on operational research in sports was fairness, which is probably one of the most important features in sports competitions. In the context of scheduling of the soccer leagues, a schedule would be considered fair if ex-ante all teams have the same probability to convert the home advantage into success, given their individual characteristics, regardless of the day of the game. In this regard, our findings suggest that in all four leagues that we investigated (Bundesliga 1, La Liga, Premier League and Ligue 1), we find that the current schedule structure favours underdog teams that play fewer home games on non-frequent days and favourite teams that play more away games on these days. However, at least in the Premier League, the possible mediator may be a referee bias in favour or against certain teams. Therefore, our results may be of interest to the referees' committee whose goal is to reduce any bias.

Our results also suggest that all four leagues suffer from lower attendance rates on non-frequent days. Therefore, our findings may also be of interest to the calendar committees of the relevant leagues, whose task is to allocate games in a way that

eliminates any advantage driven by schedule. More specifically, it is important that home games on non-frequent days are allocated evenly on the level of a single team per season. Having said that, we are aware that the calendar committees have to deal with a large amount of constraints, such that it is sometimes almost impossible to satisfy all of them. On top of it, schedule effects that might drive a referee bias should be a big concern for the referees' committees rather than for calendar committees. However, the case of 1.FSV Mainz 05 that was mentioned in the introduction provides an example where the calendar committee disregarded the issue of fair allocation between different days. Furthermore, our findings might be worth an additional restriction, not only in constructing the initial schedule, but also for implementing proactive or reactive strategies for rescheduled games, as Yi, Goossens, and Nobibon (2020) recently proposed.

In addition, the results of this paper may also help coaches and players prepare to play on different days. According to our results, underdog teams may be expected to have a lower home advantage on non-frequent days and should therefore consider adjusting their preparation to these games. Furthermore, teams may adjust their ticket sales strategy. For example, tickets for games on non-frequent days, for which there is less demand, could be sold for a lower price to attract larger crowds and increase home advantage.

Finally, we call for additional empirical research on different schedule effects in sports leagues that may potentially affect on-pitch performance as well as financial outcomes.

Appendices

Appendix 4.A: Descriptive statistics

Table 4.9: Descriptive statistics for all available variables

	Favourite playing home		Underdog playing home	
Variable	Frequent days	NF days	Frequent days	NF days
Game outcomes				
Fav. points	1.930	1.981	1.333	1.575
Und. points	0.817	0.784	1.391	1.144
Fav. win	0.559	0.582	0.352	0.432
Und. win	0.188	0.183	0.371	0.288
Game characteristics				
Stadium capacity	40,825	44,864	29,924	31,410
Visitors	32,743	34,583	24,031	24,943
Share of capacity	0.781	0.750	0.792	0.794
Ln(Attendance)	10.242	10.292	9.955	10.005
Distance	487.502	501.569	485.106	461.455
Teams characteristics				
Fav. team value*	147.424	194.466	148.308	204.742
Und. team value*	63.947	73.319	64.615	78.651
Fav. standardized team value	0.218	0.464	0.228	0.510
Und. standardized team value	-0.577	-0.508	-0.570	-0.515
Fav. mean team value*	4.389	5.635	4.407	5.893
Und. mean team value*	1.933	2.209	1.955	2.336
Fav. median team value*	2.999	3.738	3.012	3.964
Und. median team value*	1.473	1.632	1.479	1.760
Fav. ratio top3 to ranked 9-11 most val. player	2.410	2.492	2.408	2.469
Und. ratio top3 to ranked 9-11 most val. player	2.330	2.403	2.332	2.381

Table 4.9, continued

	Favourite playing home		Underdog playing home	
Variable	Frequent days	NF days	Frequent days	NF days
Fav. plays in Champions or Europa League	0.331	0.455	0.340	0.440
Und. plays in Champions or Europa League	0.086	0.127	0.084	0.108
Fav. plays in Europa League	0.176	0.221	0.183	0.211
Und. plays in Europa League	0.070	0.112	0.066	0.083
Fav. plays in Champions League	0.184	0.290	0.188	0.281
Und. play in Champions League	0.023	0.026	0.024	0.041
Newcomer	0.046	0.054	0.292	0.242
Fav. coaches age	49.280	49.708	49.325	49.917
Und. coaches age	48.422	48.815	48.446	49.110
Fav. new coach (first match)	0.024	0.018	0.027	0.014
Und. new coach (first match)	0.024	0.011	0.021	0.021
Fav. new coach (second match)	0.025	0.021	0.023	0.021
Und. new coach (second match)	0.023	0.039	0.025	0.023
Fav. new coach (third match)	0.022	0.019	0.024	0.020
Und. new coach (third match)	0.024	0.016	0.025	0.023
Fav. new coach (fourth match)	0.021	0.025	0.021	0.036
Und. new coach (fourth match)	0.020	0.023	0.021	0.027
Fav. new coach (fifth match)	0.023	0.035	0.024	0.032
Und. new coach (fifth match)	0.016	0.026	0.019	0.032

Table 4.9, continued

	Favourite playing home		Underdog playing home	
Variable	Frequent days	NF days	Frequent days	NF days
Fav. size squad	33.363	34.186	33.384	34.561
Und. size squad	32.753	32.889	32.704	33.286
Fav. number of foreigners	17.525	17.427	17.540	18.078
Und. number of foreigners	15.822	15.202	15.886	16.089
Fav. share of foreigners	0.520	0.503	0.520	0.515
Und. share of foreigners	0.478	0.457	0.479	0.476
Fav. share right-footed	0.696	0.692	0.696	0.680
Und. share right-footed	0.694	0.709	0.695	0.695
Fav. share left-footed	0.213	0.206	0.212	0.219
Und. share left-footed	0.222	0.218	0.219	0.225
Fav. share both-footed	0.073	0.074	0.073	0.089
Und. share both-footed	0.061	0.054	0.063	0.063
Fav. mean height[§]	181.816	181.453	181.785	181.677
Und. mean height[§]	181.778	181.534	181.786	181.534
Fav. min height[§]	171.736	171.170	171.715	171.391
Und. min height[§]	172.026	171.525	171.982	171.666
Fav. max height[§]	191.623	191.817	191.586	191.933
Und. max height[§]	191.005	191.111	191.039	191.206
Fav. sd height	6.288	6.562	6.295	6.520
Und. sd height	5.986	6.116	6.003	6.147
Fav. HHI	0.060	0.061	0.060	0.060
Und. HHI	0.058	0.059	0.058	0.058
Fav. sd HHI	0.031	0.032	0.031	0.032
Und. sd HHI	0.028	0.029	0.028	0.029

Table 4.9, continued

	Favourite playing home		Underdog playing home	
Variable	Frequent days	NF days	Frequent days	NF days
Fav. ratio top3 to ranked 12-14 most val. player	3.065	3.181	3.065	3.135
Und. ratio top3 to ranked 12-14 most val. player	2.927	3.044	2.926	3.002
Fav. ratio top11 to ranked 12-23 most val. player	2.879	2.939	2.881	2.879
Und. ratio top11 to ranked 12-23 most val. player	2.704	2.779	2.702	2.690
Fav. median age	23.896	23.927	23.890	23.915
Und. median age	24.356	24.502	24.360	24.571
Fav. mean age	24.258	24.282	24.250	24.265
Und. mean age	24.663	24.762	24.666	24.813
Fav. sd age	4.470	4.502	4.477	4.533
Und. sd age	4.453	4.430	4.455	4.434
Fav. mean age 11 most val. player	25.868	25.825	25.878	25.780
Und. mean age 11 most val. player	26.106	26.024	26.118	26.027
Fav. age ratio top 11 to 12-23 most val. player	1.006	1.000	1.006	0.995
Und. age ratio top 11 to 12-23 most val. player	1.003	0.994	1.003	0.994
Fav. mean age if aged above 20	26.069	26.111	26.076	26.151
Und. mean age if aged above 20	26.239	26.265	26.240	26.297
Fav. min age	17.239	17.146	17.225	17.091
Und. min age	17.401	17.420	17.413	17.414

Table 4.9, continued

	Favourite playing home		Underdog playing home	
Variable	Frequent days	NF days	Frequent days	NF days
Fav. max age	34.232	34.330	34.239	34.433
Und. max age	34.455	34.591	34.501	34.671
Schedule-related				
African Cup of Nations	0.066	0.023	0.066	0.037
Asian Cup	0.022	0.009	0.021	0.007
Public holiday	0.035	0.091	0.035	0.091
Christmas holiday (24.01. - 01.01.)	0.015	0.095	0.015	0.105
After inter. break	0.123	0.035	0.116	0.048
Before UEFA European Championship	0.064	0.084	0.063	0.073
After UEFA European Championship	0.042	0.065	0.041	0.048
Before FIFA World Cup	0.040	0.039	0.044	0.027
After FIFA World Cup	0.028	0.049	0.028	0.044
Observations	4462	569	4548	563

Notes: This table presents average values for all available variables from all league combined for the three most frequent and four non-frequent days. *Und.* and *Fav.* represent the coefficients for the underdog and favourite team, respectively. *sd* represents the standard deviation. * in mill. Euro. § in centimetres.

Appendix 4.B: Propensity score estimation

Table 4.10: Estimation of propensity score

Variables	Favourite playing home	Underdog playing home
Schedule and match characteristics		
Season 2009/10	-0.029*	
Season 2013/14	0.027*	
Season 2014/15	0.044***	0.035**
Season 2015/16	0.031**	0.055***
Ligue 1	-0.056**	-0.068***
Premier League		0.045**
March	0.011	
August	-0.065***	-0.082***
November		-0.034*
After international break	-0.104***	
Public holiday	0.103***	0.085***
Referee 28	0.131	
Referee 125		-0.042
Referee 178		-0.011
Referee 220		0.105
Team characteristics		
Fav. standardized team value	0.007	0.016
Und. standardized team value	0.046**	0.045*
Und. team value (in mill. €)	-0.000	-0.000
Fav. average team value (in mill. €)	-0.004	0.003
Fav. sd team value	0.006	
Und. ratio of value top 3 to ranked 9-11 most val. player		0.007
Und. ratio of value top 11 to ranked 12-23 most val. player		-0.005
Und. Athletic Bilbao	-0.009	
Und. Getafe CF	0.039	
Und. EA Guingamp	-0.020	
Und. RC Mallorca	0.014	
Fav. OG Nice	0.042	
Und. VfB Stuttgart	0.016	
Und. Deportivo Alaves		-0.001

Table 4.10 continued

Variables	Favourite playing home	Underdog playing home
Fav. Tottenham Hotspurs		-0.044
Und. Xerez CD		-0.048
Und. Eintracht Frankfurt		0.042
Fav. UD Las Palmas		0.157
Fav. Manchester United		-0.045
Und. Manchester United		0.235*
Und. Ajaccio GFCO	-0.033	
Fav. Cordoba CF	0.113	
Fav. plays in Europa League	0.026**	
Und. plays in Europa League	0.020	
Und. plays in Champions League	-0.056*	-0.000
Und. sd height of 11 most valuable players	0.005	
Fav. age of 11 most valuable players	-0.001	-0.003
Und. max. height of 11 most valuable players	-0.002	
Und coach age	0.001	0.001**
Fav. squad size		0.006***
Fav. number of foreigners in squad		-0.003**
Und. share of foreigners	-0.086**	-0.061*
Number of observations	5031	5111

Notes: Dependent variable is whether a game is played on one of the four non-frequent days. Probit average marginal effects are presented. The results are based on the union of variables selected by the two-step LASSO variable selection for playing on non-frequent days and the number of favourites' points. *Und*. and *Fav*. represent the underdog and favourite teams, respectively. *sd* is the standard deviation. *, **, and *** represent the 10%, 5%, and 1% significance levels, respectively.

Appendix 4.C: Additional Results

Table 4.11: Effects of playing on five non-frequent days for each league

Outcomes	Ligue 1	La Liga	Premier L.	Bundesliga	Wald-test p-value
Panel A: Incl. referee dummies					
			Favourite team playing home		
Fav. Points	-0.095	0.043	-0.010	-0.045	0.78
	(0.34)	(0.64)	(0.92)	(0.68)	
Und. Points	0.043	-0.031	0.114	-0.048	0.48
	(0.63)	(0.71)	(0.16)	(0.64)	
Share of capacity	-0.030*	-0.060***	-0.016**	-0.031**	0.02
	(0.06)	(0.00)	(0.02)	(0.01)	
Ln(Attendance)	-0.084*	-0.125***	-0.039	-0.055	0.36
	(0.07)	(0.00)	(0.14)	(0.12)	
			Underdog team playing home		
Fav. Points	0.151	0.154*	0.148	0.274**	0.83
	(0.17)	(0.10)	(0.15)	(0.02)	
Und. Points	-0.150	-0.227**	-0.143	-0.236**	0.87
	(0.15)	(0.01)	(0.14)	(0.04)	
Share of capacity	-0.025*	-0.070***	-0.037***	-0.028**	0.06
	(0.09)	(0.00)	(0.00)	(0.02)	
Ln(Attendance)	-0.054*	-0.110***	-0.074***	-0.054	0.67
	(0.09)	(0.00)	(0.00)	(0.14)	
Panel B: Excl. referee dummies					
			Favourite team playing home		
Fav. Points	-0.018	0.020	-0.002	0.006	0.99
	(0.86)	(0.83)	(0.98)	(0.96)	
Und. Points	0.114	-0.099	-0.047	-0.046	0.33
	(0.20)	(0.23)	(0.59)	(0.66)	
Share of capacity	-0.033**	-0.065***	-0.018**	-0.029**	0.01
	(0.03)	(0.00)	(0.01)	(0.01)	
Ln(Attendance)	-0.053	-0.097**	-0.047*	-0.060*	0.81
	(0.28)	(0.03)	(0.07)	(0.07)	
			Underdog team playing home		
Fav. Points	0.153	0.155	-0.008	0.331***	0.24
	(0.14)	(0.11)	(0.94)	(0.00)	
Und. Points	-0.198*	-0.143	0.028	-0.202*	0.32
	(0.06)	(0.13)	(0.77)	(0.08)	
Share of capacity	-0.043***	-0.054***	-0.035***	-0.020*	0.20
	(0.00)	(0.00)	(0.00)	(0.08)	
Ln(Attendance)	-0.027	-0.125***	-0.073***	-0.033	0.15
	(0.41)	(0.00)	(0.00)	(0.32)	

Notes: The results represent the effects of playing on five non-frequent days compared to the two most frequent days for each league separately. P-values of the effects are presented in parentheses. Standard errors are calculated as weight-based standard errors and clustered at the seasonal level. *, **, and *** denote the 10%, 5% and 1% significance levels, respectively. Common support for each of the matching estimations is at least 92.8% (Ligue 1), 93.4% (La Liga), 89.5% (Premier League), and 94.6% (Bundesliga)

Table 4.12: Effects excluding the third most frequent day for each league

Outcomes	Ligue 1	La Liga	Premier L.	Bundesliga	Wald-test p-value
Panel A: Incl. referee dummies					
			Favourite team playing home		
Fav. Points	-0.000	-0.038	0.052	-0.311*	0.36
	(1.00)	(0.72)	(0.67)	(0.07)	
Und. Points	-0.052	0.040	-0.082	0.311*	0.20
	(0.65)	(0.69)	(0.45)	(0.05)	
Share of capacity	-0.045*	-0.067***	-0.018**	-0.094***	0.00
	(0.06)	(0.00)	(0.02)	(0.00)	
Ln(Attendance)	-0.107*	-0.103**	-0.060*	-0.124**	0.70
	(0.09)	(0.04)	(0.06)	(0.02)	
			Underdog team playing home		
Fav. Points	0.270*	0.155	0.195*	0.302*	0.87
	(0.09)	(0.15)	(0.09)	(0.08)	
Und. Points	-0.301**	-0.257**	-0.176*	-0.315**	0.85
	(0.03)	(0.01)	(0.10)	(0.05)	
Share of capacity	-0.050***	-0.060***	-0.037***	-0.032**	0.46
	(0.00)	(0.00)	(0.00)	(0.04)	
Ln(Attendance)	-0.069*	-0.104***	-0.067**	-0.083*	0.86
	(0.08)	(0.00)	(0.02)	(0.10)	
Panel B: Excl. referee dummies					
			Favourite team playing home		
Fav. Points	0.095	-0.090	-0.007	-0.221	0.48
	(0.47)	(0.39)	(0.95)	(0.19)	
Und. Points	-0.112	-0.085	-0.060	0.346**	0.13
	(0.31)	(0.36)	(0.56)	(0.05)	
Share of capacity	-0.056**	-0.063***	-0.019**	-0.092***	0.00
	(0.01)	(0.00)	(0.04)	(0.00)	
Ln(Attendance)	-0.093	-0.158***	-0.053	-0.111**	0.35
	(0.11)	(0.00)	(0.12)	(0.04)	
			Underdog team playing home		
Fav. Points	0.329**	0.266**	-0.046	0.482***	0.03
	(0.04)	(0.01)	(0.67)	(0.00)	
Und. Points	-0.378**	-0.311***	-0.004	-0.293*	0.10
	(0.01)	(0.00)	(0.97)	(0.07)	
Share of capacity	-0.049***	-0.043***	-0.033***	-0.051***	0.68
	(0.00)	(0.00)	(0.00)	(0.00)	
Ln(Attendance)	-0.049	-0.093**	-0.054*	-0.099*	0.74
	(0.24)	(0.01)	(0.06)	(0.06)	

Notes: The results represent the effects of playing on four non-frequent days compared to the two most frequent days for each league separately. P-values of the effects are presented in parentheses. Standard errors are calculated as weight-based standard errors and clustered at the seasonal level. *, **, and *** denote the 10%, 5% and 1% significance levels, respectively. Common support for each of the matching estimations is at least 87.3% (Ligue 1), 92.5% (La Liga), 91.4% (Premier League), and 83.2% (Bundesliga)

Appendix 4.D: List of sources for database

www.uefa.com

www.fifa.com

www.transfermarkt.com

www.football-data.co.uk

www.rsssf.com

www.espnfc.com

www.fcal.ch

https://en.wikipedia.org/wiki/Bundesliga

www.weltfussball.com

www.google.com/maps

www.kicker.de

References

Abadie, A., & Imbens, G. (2016). Matching on the Estimated Propensity Score. *Econometrica, 84*(2), 781-807.

Antonelli, J., Cefalu, M., Palmer, N., & Agniel, D. (2018). Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics, 74*(4), 1171-1179.

Apesteguia, J., & Palacios-Huerta, I. (2010). Psychological pressure in competitive environments: Evidence from a randomized natural experiment. *American Economic Review, 100*(5), 2548-2564.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large Stakes and Big Mistakes. *The Review of Economic Studies, 76*(2), 451-469.

Arlegi, R., & Dimitrov, D. (2020). Fair elimination-type competitions. *European Journal of Operational Research, 287*(2), 528-535.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science, 355*, 483-485.

Athey, S., & Imbens G. (2019). Machine Learning Methods Economists Should Know About. *arXiv:1903.10075*.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353-7360.

Athey, S., & Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. *arXiv:1902.07409*.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics, 47*(2), 1148-1178.

Baumeister, R. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46*(3), 610-620.

Baumeister, R., Hamilton, J., & Tice, D. (1985). Public Versus Private Expectancy of Success: Confidence Booster or Performance Pressure? *Journal of Personality and Social Psychology, 48*(6), 1447-1457.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies, 81*(2), 608-650.

Biewen, M., Fitzenberger, B., Osikominu, A., & Paul, M. (2014). The Effectiveness of Public Sponsored Training Revisited: The Importance of Data and Methodological Choices. *Journal of Labor Economics, 32*(4), 837-897.

Bodory, H., Camponovo, L., Huber, M., & Lechner, M. (2018). The Finite Sample Performance of Inference Methods for Propensity Score Matching and Weighting Estimators. *Journal of Business and Economic Statistics*, 1-43.

Boyko, R., Boyko, A., & Boyko, M. (2007). Referee bias contributes to home advantage in English Premiership football. *Journal of Sports Sciences, 25*(11), 1185-1194.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Brown, K., Merrigan, P., & Royer, J. (2018). Estimating Average Treatment Effects With Propensity Scores Estimated With Four Machine Learning Procedures: Simulation Results in High Dimensional Settings and With Time to Event Outcomes. *SSRN Electronic Journal*.

Bryson, A., Frick, B., & Simmons, R. (2013). The Returns to Scarce Talent: Footedness and Player Remuneration in European Soccer. *Journal of Sports Economics, 14*(6), 606-628.

Buraimo, B. (2008). Stadium attendance and television audience demand in English League Football. *Managerial and Decision Economics, 29*(6), 513-523.

Buraimo, B., & Simmons, R. (2015). Uncertainty of Outcome or Star Quality? Television Audience Demand for English Premier League Football. *International Journal of the Economics of Business, 22*(3), 449-469.

Buraimo, B., Migali, G., & Simmons, R. (2016). An Analysis of Consumer Response to Corruption: Italy's Calciopoli Scandal. *Oxford Bulletin of Economics and Statistics, 78*(1), 22-41.

Buraimo, B., Tena, J., & de la Piedra, J. (2018). Attendance demand in a developing football market: the case of the Peruvian first division. *European Sport Management Quarterly, 18*(5), 671-686.

Butler, J., & Baumeister, R. (1998). The trouble with friendly faces: Skilled performance with a supportive audience. *Journal of personality and social psychology, 75*(5), 1213.

Caliendo, M., Mahlstedt, R., & Mitnik, O. (2017). Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labor market policies. *Labour Economics, 46*, 14-25.

Calónico, S., & Smith, J. (2017). The Women of the National Supported Work Demonstration. *Journal of Labor Economics, 35*(S1), 65-97.

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal, 61*(3), 1-24.

Cao, Z., Price, J., & Stone, D. (2011). Performance under pressure in the NBA. *Journal of Sports Economics, 12*(3), 231-252.

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association, 16*(3), 894-931.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal, 21*(1), C1-C68.

Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018). The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *Econometrica, 86*(6), 1911-1938.

Coates, D., Frick, B., & Jewell, T. (2016). Superstar Salaries and Soccer Success: The Impact of Designated Players in Major League Soccer. *Journal of Sports Economics, 17*(7), 716-735.

Cockx, B., Lechner, M., & Bollens, J. (2019). Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *arXiv: 1912.12864*.

Cohen-Zada, D., Krumer, A., & Shapir, O. (2018). Testing the effect of serve order in tennis tiebreak. *Journal of Economic Behavior and Organization, 146*, 106-115.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2017, 11 7). Overlap in Observational Studies with High-Dimensional Covariates. *arXiv:1711.02582v3*.

Davis, J., & Heller, S. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review: Papers & Proceedings, 107*(5), 546-550.

Dehejia, R., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151-161.

Della Croce, F., & Oliveri, D. (2006). Scheduling the Italian Football League: An ILP-based approach. *Computers and Operations Research, 33*(7), 1963-1974.

Doerr, A., Fitzenberger, B., Kruppe, T., Paul, M., & Strittmatter, A. (2017). Employment and earnings effects of awarding training vouchers in Germany. *Industrial and Labor Relations Review, 70*(3), 767-812.

Dohmen, T. (2008). Do professionals choke under pressure? *Journal of Economic Behavior and Organization, 65*, 636-653.

Downward, P., & Jones, M. (2007). Effects of crowd size on referee decisions: Analysis of the FA Cup. *Journal of Sports Sciences, 25*(14), 1541-1545.

Durán, G., Durán, S., Marenco, J., Mascialino, F., & Rey, P. (2019). Scheduling Argentina's professional basketball leagues: A variation on the Travelling Tournament Problem. *European Journal of Operational Research, 275*(3), 1126-1138.

Durán, G., Guajardo, M., & Sauré, D. (2017). Scheduling the South American Qualifiers to the 2018 FIFA World Cup by integer programming. *European Journal of Operational Research, 262*(3), 1109-1115.
Ehrenberg, R., & Bognanno, M. (1990). Do tournaments have incentive effects? *Journal of Political Economy, 98*(6), 1307-1324.
Faltings, R., Krumer, A., & Lechner, M. (2019). Rot-Jaune-Verde. Language and Favoritism: Evidence from Swiss Soccer. *Economic Working Paper Series, 1915*.
Fan, Q., Hsu, Y.-C., Lieli, R., & Zhang, Y. (2019). Estimation of Conditional Average Treatment Effects with High-Dimensional Data. *arXiv:1908.02399*.
Flores, R., Forrest, D., & Tena, J. (2012). Decision taking under pressure: Evidence on football manager dismissals in Argentina and their consequences. *European Journal of Operational Research, 222*(3), 653-662.
Forrest, D., & Simmons, R. (2006). New Issues in Attendance Demand: The Case of the English Football League. *Journal of Sports Economics, 7*(3), 247-266.
Franck, E., & Nüesch, S. (2012). Talent and/or popularity: What does it take to be a superstar? *Economic Inquiry, 50*(1), 202-216.
Ginsburgh, V., & Van Ours, J. (2003). Expert opinion and compensation: Evidence from a musical competition. *American Economic Review, 93*(1), 289-296.
Goller, D., & Krumer, A. (2020). Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues. *European Journal of Operational Research, 286*(2), 740-754.
Goller, D., Knaus, M., Lechner, M., & Okasa, G. (2018). Predicting Match Outcomes in Football by an Ordered Forest Estimator. *Economic Working Paper Series, No. 1811*.
Goller, D., Lechner, M., Moczall, A., & Wolff, J. (2020). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics, 65*.
González-Díaz, J., Gossner, O., & Rogers, B. (2012). Performing best when it matters most: Evidence from professional tennis. *Journal of Economic Behavior and Organization, 84*(3), 767-781.
Goossens, D., & Spieksma, F. (2009). Scheduling the belgian soccer league. *Interfaces, 39*(2), 109-118.
Goossens, D., & Spieksma, F. (2012). Soccer schedules in Europe: An overview. *Journal of Scheduling, 15(5)*, 641-651.
Harb-Wu, K., & Krumer, A. (2019). Choking Under Pressure in Front of a Supportive Audience: Evidence from Professional Biathlon. *Journal of Economic Behavior and Organization, 166*, 246-262.
Harrer, T., Moczall, A., & Wolff, J. (2020). Free, free, set them free? Are programmes effective that allow job centres considerable freedom to choose the exact design? *International Journal of Social Welfare, 29*(2), 154-167.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data mining, inference, and prediction.* 2nd. ed. New York: Springer.
Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research, 46*(3), 477-513.
Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*, 199-236.
Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945-960.
Huber, M., Lechner, M., & Steinmayr, A. (2015). Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour. *Empirical Economics, 49*(1), 1-31.

Huber, M., Lechner, M., & Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics, 175*(1), 1-21.
Hurley, W. (2009). Equitable birthdate categorization systems for organized minor sports competition. *European Journal of Operational Research, 192*(1), 253-264.
Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29.
Imbens, G. (2015). Matching Methods in Practice: Three Examples. *Journal of Human Resources, 50*(2), 373-419.
Imbens, G., & Rubin, D. (2015). *Causal inference: For statistics, social, and biomedical sciences an introduction.* Cambridge University Press.
Kendall, G. (2008). Scheduling English football fixtures over holiday periods. *Journal of the Operational Research Society, 59*(6), 743-755.
Kendall, G., & Lenten, L. (2017). When sports rules go awry. *European Journal of Operational Research, 257*(2), 377-394.
Kendall, G., Knust, S., Ribeiro, C., & Urrutia, S. (2010). Scheduling in sports: An annotated bibliography. *Computers and Operations Research, 37*(1), 1-19.
Kirkegaard, R. (2012). Favoritism in asymmetric contests: Head starts and handicaps. *Games and Economic Behavior, 76*(1), 226-248.
Klein Teeselink, B., Potter Van Loon, R., Van Den Assem, M., & Van Dolder, D. (2020). Incentives, Performance and Choking in Darts. *Journal of Economic Behavior and Organization, 169*, 38-52.
Klumpp, T., & Polborn, M. (2006). Primaries and the New Hampshire Effect. *Journal of Public Economics, 90*(6-7), 1073-1114.
Knaus, M. (2020). Double Machine Learning based Program Evaluation under Unconfoundedness. *arXiv:2003.03191v1*.
Knaus, M., Lechner, M., & Strittmatter, A. (2020a). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *The Econometrics Journal, utaa014*.
Knaus, M., Lechner, M., & Strittmatter, A. (2020b). Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach. *Journal of Human Resources, 0718-9615R1*.
Konrad, K. (2002). Investment in the absence of property rights; the role of incumbency advantages. *European Economic Review, 46*, 1521-1537.
Krumer, A. (2013). Best-of-two contests with psychological effects. *Theory and Decision, 75*(1), 85-100.
Krumer, A. (2020). Testing the effect of kick-off time in the UEFA Europa League. *European Sport Management Quarterly, 20*(2), 225-238.
Krumer, A., & Lechner, M. (2017). First in first win: Evidence on schedule effects in round-robin tournaments in mega-events. *European Economic Review, 100*, 412-427.
Krumer, A., & Lechner, M. (2018). Midweek effect on soccer performance: Evidence from the German Bundesliga. *Economic Inquiry, 56*(1), 193-207.
Krumer, A., Megidish, R., & Sela, A. (2020). The optimal design of round-robin tournaments with three players. *Journal of Scheduling, 23*(3), 379-396.
Lazear, E. (2000). The Power of Incentives. *The American Economic Review: Papers and Proceedings, 90*(2), 410-414.
Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. *IZA Discussion Paper Series, No. 12040*.
Lechner, M., & Strittmatter, A. (2019). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews, 38*(2), 193-207.
Lechner, M., & Wunsch, C. (2009). Are Training Programs More Effective When Unemployment Is High? *Journal of Labor Economics, 27*(4), 653-692.

Lechner, M., & Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics, 21*, 111-121.
Lechner, M., Miquel, R., & Wunsch, C. (2011). Long-run effects of public sector sponsored training in West Germany. *Journal of the European Economic Association, 9*(4), 742-784.
Lee, B., Lessler, J., & Stuart, E. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346.
Levitt, S., & List, J. (2008). Homo economicus evolves. *Science, 319*(5865), 909-910.
Liebscher, S., & Kirschstein, T. (2017). Predicting the outcome of professional darts tournaments. *International Journal of Performance Analysis in Sport, 17*(5), 666-683.
Lonsdale, C., & Tam, J. (2008). On the temporal and behavioural consistency of pre-performance routines: An intra-individual analysis of elite basketball players' free throw shooting accuracy. *Journal of Sports Sciences, 26*(3), 259-266.
Masters, R. (1992). Knowledge, knerves and know-how: The role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure. *British Journal of Psychology, 83*(3), 343-358.
Megidish, R., & Sela, A. (2014). Sequential contests with synergy and budget constraints. *Social Choice and Welfare, 42*(1), 215-243.
Meirowitz, A. (2008). Electoral contests, incumbency advantages, and campaign finance. *Journal of Politics, 70*(3), 681-699.
Mesagno, C., & Mullane-Grant, T. (2010). A Comparison of Different Pre-Performance Routines as Possible Choking Interventions. *Journal of Applied Sports Psychology, 22*(3), 343-360.
Musch, J., & Grondin, S. (2001). Unequal competition as an impediment to personal development: A review of the relative age effect in sport. *Developmental Review, 21*(2), 147-167.
Nevill, A., Balmer, N., & Williams, A. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise, 3*(4), 261-272.
Nie, X., & Wager, S. (2017). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *arXiv:1712.04912v3*.
Nitzan, S. (1994). Modelling rent-seeking. *European Journal of Political Economy, 10*, 41-60.
Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryan, T., Cleary, P., & McNeil, B. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*(4), 387-398.
Ötting, M., Deutscher, C., Schneemann, S., Langrock, R., Gehrmann, S., & Scholten, H. (2020). Performance under pressure in skill tasks: An analysis of professional darts. *PLoS ONE, 15*(2), 1-21.
Page, K., & Page, L. (2010). Alone against the crowd: Individual differences in referees' ability to cope under pressure. *Journal of Economics Psychology, 31*, 192-199.
Page, L., & Page, K. (2007). The second leg home advantage: Evidence from European football cup competitions. *Journal of Sports Sciences, 25*(14), 1547-1556.
Pettersson-Lidbom, P., & Priks, M. (2010). Behavior under social pressure: Empty Italian stadiums and referee bias. *Economics Letters, 108*(2), 212-214.
Pirracchio, R., Petersen, M., & Van Der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using Super Learner. *American Journal of Epidemiology, 181*(2), 108-119.
Ponzo, M., & Scoppa, V. (2018). Does the Home Advantage Depend on Crowd Support? Evidence From Same-Stadium Derbies. *Journal of Sports Economics, 19*(4), 562-582.
Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature, 37*(1), 7-63.
Rasmussen, R., & Trick, M. (2008). Round robin scheduling - a survey. *European Journal of Operational Research, 188*(3), 617-636.
Ribeiro, C., & Urrutia, S. (2007). Heuristics for the mirrored traveling tournament problem. *European Journal of Operational Research, 179*(3), 775-787.

Robins, J., Rotnitzky, A., & Zhao, L. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association, 89*(427), 846-866.
Robins, J., Rotnitzky, A., & Zhao, L. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association, 90*(429), 106-121.
Rohde, M., & Breuer, C. (2017). Financial Incentives and Strategic Behavior in European Professional Football: A Match Day Analysis of Starting Squads in the German Bundesliga and UEFA Competitions. *International Journal of Sport Finance, 12*(2).
Rosen, S. (1986). Prizes and Incentives in Elimination Tournaments. *The American Economic Review, 76*(4), 701-715.
Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrica, 70*(1), 41-55.
Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.
Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.
Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*, 20-36.
Segev, E., & Sela, A. (2014). Sequential all-pay auctions with head starts. *Social Choice and Welfare, 43*(4), 893-923.
Semenova, V., & Chernozhukov, V. (2017). Simultaneous Inference for Best Linear Predictor of the Conditional Average Treatment Effect and Other Structural Functions. *arXiv:1702.06240v3*.
Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., & Cook, E. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*(6), 546-555.
Shapiro, C., & Stiglitz, J. (1984). Equilibrium Unemployment as a Worker Discipline Device. *The American Economic Review, 74*(3), 433-444.
Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1-2), 305-353.
Stiglitz, J. (1976). The Efficiency Wage Hypothesis, Surplus Labour, and the Distribution of Income in LDCs. *Oxford Economic Papers, 28*(2), 185-207.
Strauss, B. (1997). Choking under pressure: Positive public expectations and performance in a motor task. *Zeitschrift für Experimentelle Psychologie, 44*(4), 636-655.
Szymanski, S. (2003). The Economic Design of Sporting Contests. *Journal of Economic Literature, XLI*, 1137-1187.
Tena, J., & Forrest, D. (2007). Within-season dismissal of football coaches: Statistical analysis of causes and consequences. *European Journal of Operational Research, 181*(1), 362-373.
Tian, L., Alizadeh, A., Gentles, A., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association, 109*(508), 1517-1532.
Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B, 58*(1), 267-288.
Tibshirani, R., Price, A., & Taylor, J. (2011). A statistician plays darts. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 174*(1), 213-226.
Toma, M. (2017). Missed Shots at the Free-Throw Line: Analyzing the Determinants of Choking Under Pressure. *Journal of Sports Economics, 18*(6), 539-559.
Tullock, G. (1980). Efficient rent-seeking. *In: J.M. Buchanan, R.D. Tollison and G. Tullock (Eds.), Toward a theory of the rent-seeking society. College Station: Texas A&M University Press*, 97-112.

Van Bulck, D., Goossens, D., Schönberger, J., & Guajardo, M. (2020). RobinX: A three-field classification and unified data format for round-robin sports timetabling. *European Journal of Operational Research, 280*(2), 568-580.
van der Laan, M., Polley, E., & Hubbard, A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1), 1-21.
Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association, 113*(523), 1228-1242.
Wager, S., & Walther, G. (2015). Adaptive Concentration of Regression Trees, with Application to Random Forests. *arXiv:1503.06388*.
Wang, C., Goossens, D., & Vandebroek, M. (2018). The Impact of the Soccer Schedule on TV Viewership and Stadium Attendance: Evidence From the Belgian Pro League. *Journal of Sports Economics, 19*(1), 82-112.
Wergin, V., Beckmann, J., Gröpel, P., & Mesagno, C. (2020). Investigating cumulative effects of preperformance routine interventions in beach volleyball serving. *PLoS ONE, 15*(1), e0228012.
Wright, M. (2014). OR analysis of sporting rules - A survey. *European Journal of Operational Research, 232*(1), 1-8.
Wunsch, C., & Lechner, M. (2008). What did all the money do? On the general ineffectiveness of recent west German labour market programmes. *Kyklos, 61*(1), 134-174.
Yi, X., Goossens, D., & Nobibon, F. (2020, 4). Proactive and reactive strategies for football league timetabling. *European Journal of Operational Research, 282*(2), 772-785.
Zajonc, R. (1965). Social Facilitation. *Science, 149*(3681), 269-274.
Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779*.

Curriculum Vitae

Education

2016 – 2021	PhD in Economics and Finance, University of St.Gallen, Main Adviser: Prof. Dr. Michael Lechner
2013 – 2016	M.Sc. in Quantitative Economics, Kiel University
2014	Exchange term, Stockholm University
2010 – 2013	B.Sc. in Economics, University of Mannheim

Professional experience

Since 2020	Postdoctoral Researcher, University of Bern
2016 – 2021	Research Assistant, Swiss Institute for Empirical Economic Research, University of St.Gallen
2014 – 2016	Teaching Assistant, University of Kiel
2007 – 2010	Apprenticeship Banking Professional, Kreissparkasse Tübingen