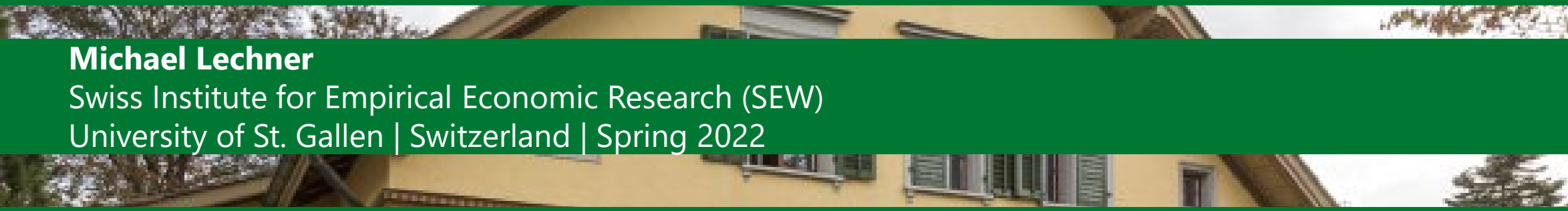


The Effects of Sports on Online Dating: Evidence from Causal ML



Michael Lechner
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen | Switzerland | Spring 2022

Joint work with
Daniel Boller (Asian Development Bank)
Gabriel Okasa (University of Lausanne)







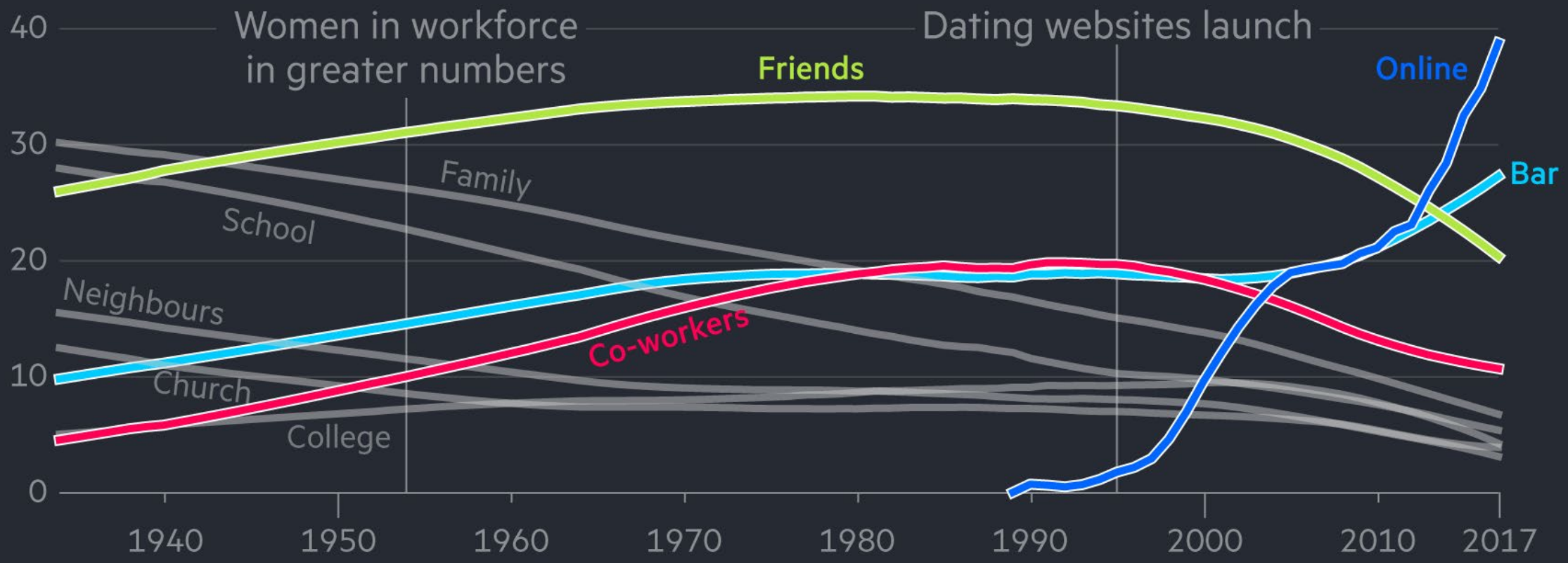
Motivation | 1

Individual sports activity has many well-documented positive effects

- Health
- Education
- Labour market outcomes
- …
- Dating?
  - **Online dating?**

Crazy in love: how heterosexual couples meet

% of US couples



Source: Rosenfeld, Michael J., Reuben J. Thomas, Sonia Hausen (2019)

FINANCIAL TIMES





Research questions & results

Research question 1: Does doing sports lead to a larger set of potential partners?

- Answer: Yes, it may substantially increase the # of potential partners

Research question 2: For whom?

- Only for men
- Effect increases with men's income





What's cool about the paper?

Practical relevance of research questions

- At least for some cohorts
  - Note: Average age on the platform is 40
- An additional potential motivation to do sports

Informative data makes identification credible

- Large German online dating platform
- Econometrician has same information as potential partners

  ➔ *Unconfoundedness* is credible identifying assumption

Large data makes estimation (more) precise & allows to detect heterogeneities

Nice & new estimation methods

- Causal machine learning ➔ *Modified Causal Forest (MCF)*





Why causal machine learning?

Various aggregation levels are of interest

- Population: **A**verage **T**reatment **E**ffects (ATE)
- Larger subgroups of … : **G**roup **ATE** (GATE)
  - Gender, income, …
- Almost individual level: **I**ndividualized **ATE** (IATE)
  - Understand variation & how it depends on *X* (detect heterogeneity)

Large set of covariates

Flexible estimation with few (statistical) assumptions





Literature | Online dating

Relevance

- Rosenfeld, Thomas, Hausen (2019): Online dating is most common way for heterosexual couples to meet (US)
- Cacioppo, Gonzaga, Ogburn, and VanderWeele (2013): 1/3 of marriages in the US (2005-2012) due to initial contact via online dating

Individual traits that improve chances on such platforms

- Many papers (see paper for details)
- Sports?





Literature | Sports | 1

Positive effects of sports activity on health

- Physical
  - Warburton, Nicol, Bredin (2006); Humphreys, McLeod, Ruseski (2014); Penedo, Dahn (2005); Eime, Young, Harvey, Charity, Payne (2013); many more …
- Mental
  - Hillman, Erickson, Kramer (2008); Strong et al. (2005); Janssen, LeBlanc (2010); many more

Positive effects on social & socio-economic variables

- Poverty & crime: Caruso (2011); …
- Labour market performance: Lechner (2009); Rooth (2011); …
- Performance in education & human capital accumulation: Lechner, Steckenleiter (2019); Fox, Barr-Anderson, Neumark-Sztainer, Wall (2010); Pfeifer, Cornelissen (2010); Felfe, Lechner, Steinmayr (2016); Lechner (2017); Fricke, Lechner, Steinmayr (2018); …





Literature | Sports & (online) dating

Very small literature

- Schulte-Hostedde, Eys, Johnson (2008, RCT): Effect of males' practiced sport discipline on females' willingness to engage in a relationship
  - "team sport athletes were perceived as being more desirable as potential mates than individual sport athletes and non-athletes"
- Farthing (2008): Risky (vs. non-risky) sports improve chances

Indirect effects: Literature on how sport improve traits (i.e. income, health, …) that have also positive effects on mating

This paper: Try to fill that gap by providing direct, up-to-date evidence





Literature | CML under CIA | Methods

Very many important developments recently

- Statistics, informatics, epidemiology, econometrics, political science, …

Survey by Knaus, Lechner, Strittmatter (2021, *Econometrics Journal*)

MCF (**M**odified **C**ausal **F**orest, Lechner 2018) ➔ **used here**

- Causal forest-based implementation of causal machine learning
- It builds on Wager & Athey (2018, JASA), Athey, Tibshirani, Wager (2019, Annals of Stat)
- MCF is a made for such applications
- MCF not included KLS 2021 → developed later





What we do?

Take large & very informative data from Germany

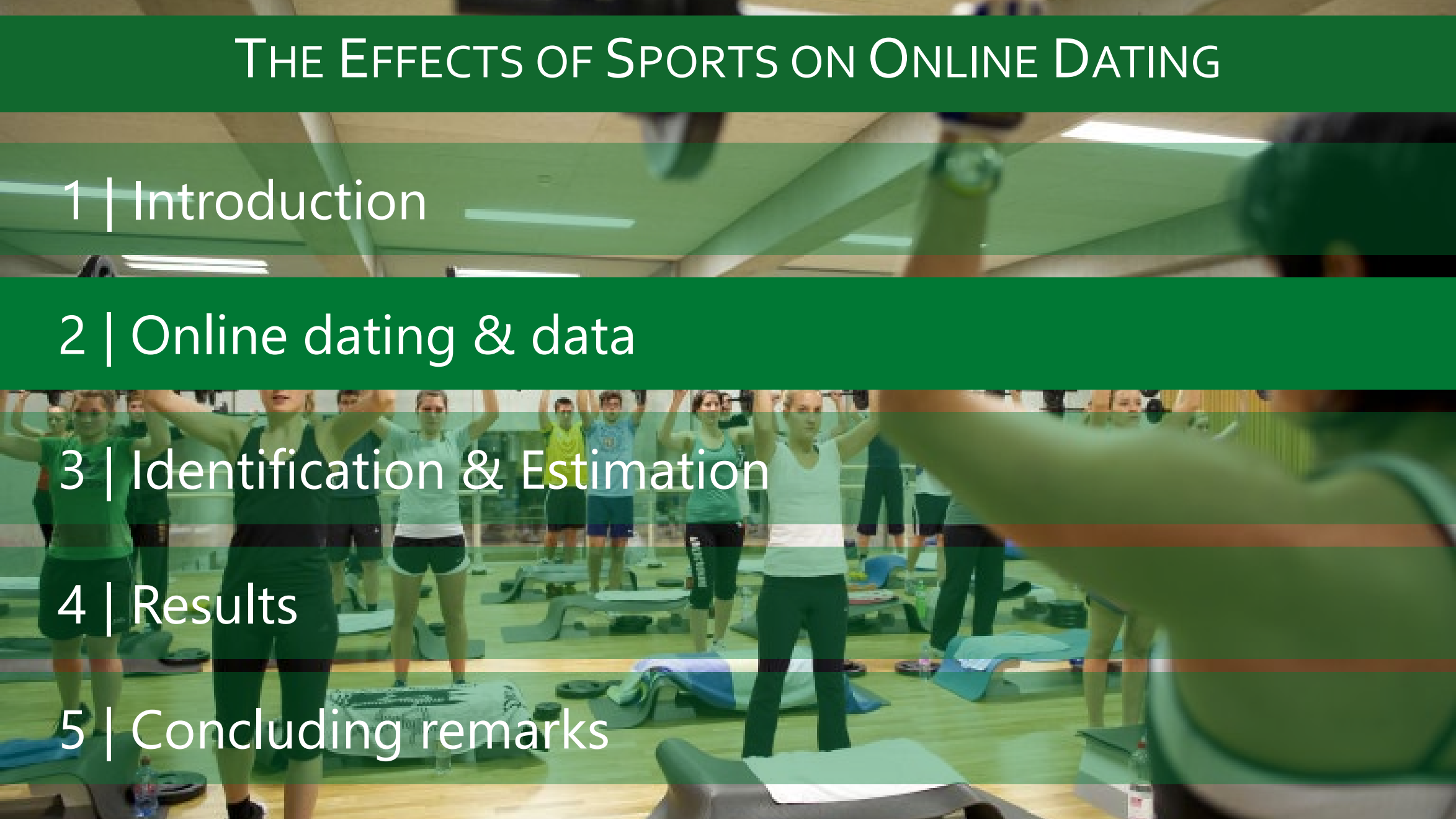
- Collaboration with a large German online dating platform

Justify selection-on-observables setting

- Main argument: Econometrician knows everything

Use MCF to estimate mean effects & heterogeneities

The Effects of Sports on Online Dating



1 | Introduction

2 | Online dating & data

3 | Identification & Estimation

4 | Results

5 | Concluding remarks





Finding a partner on the platform | 1

Registration

- Put information into profile

Define a query for a partner

- Preferred sex, age, geographic location
- Platform computes matching score
  - Computed by platform based on information from both individuals
  - All variables that enter the score are available to us
  - Potential partners are selected based on that matching score

Receive short-list of potential partners

- Username, age, blurred photo, matching score





Finding a partner on the platform | 2

User may browse on profile pages of these suggestions

- Will see the information

Multiple actions possible

- Send private message
  - Smile icon
  - Smile-back icon (if s/he received a smile-icon before)
- Leave a 'like' or 'text note' on his/her website
- Initiate a 'friendship'
  - Release profile (photo)
  - Or accepting a profile release from potential partner
- Ask for a structured game-of-questions ('Applet')
  - Find similarities & differences of user & potential partner
- User can always prevent specific other users to contact him/her





Available information

Socio-demographic information

Search criteria for potential partners

Personality questionnaire

- Life style, personality traits, attitudes, opinions

Photo

- We have no access
- But: Remains fully blurred until user decides to release photo for potential partner

Sports information

- Own activities & preference for partner
- Intensity & disciplines





Structure of data | 1

User sample = our population

- 18’036 newly registered users (2016/01/01-2016/04/30)
  - After some data cleaning: 16’864 users remained
- 667 variables for each user

Interaction sample

- 1’415’645 interactions between the new users that involve in any sense these newly registered users (2016/01/01-2016/04/30)
  - Type of action, time stamp, both ID’s
- Each interaction begins with a visit (invisible by the one visited)
  - Followed possibly by other actions
- Sender: User who initiated interaction (i.e. who visited profile)
- Recipient: User who was (passively) visited





Structure of data | 2

Outcome

- Whether visit on profile page is followed by action of sender (binary)
  - Do senders react to what they see?
  - Reaction of recipient is irrelevant

Data analyzed

- 178’372 valid unique interaction
  - Always different individuals
  - Each user received about 10 visits from a user of the inflow cohort in this period

Gender split

- 48% women in the initial user sample
- Female recipients (i.e. approached by men): 108’456 *observations*
- Male recipients (i.e. approached by women): 69’916 *observations*
  - Men are more actively approaching women than vice versa





Sports activity | 1

Sport activity is measured in the initial questionnaire

- Type of sports (27 categories): This information is ignored
  - Would lead to fairly small groups
- Intensity ('How often do you do sports?'
  - Never, rarely, monthly, weekly, daily
  - Daily is very thinly populated (3%): Not considered

Perhaps future research

- Team sports vs individual sports x intensity (8 categories)

Table 1: Shares of Sport Frequency for Male and Female Sample

	Never	*Rarely*	*Monthly*	*Weekly*	Observations
Males	0.07	0.08	0.29	0.56	69'916
Females	0.12	0.09	0.29	0.49	108'456
SOEP-IS	0.33	0.23	0.10	0.34	25'544

Note: Color intensity represents the corresponding share sizes for males and females.

Sport activity is measured in the initial questionnaire

- What type of sports (27 categories): This information is ignored

Our sample is more active than population in general

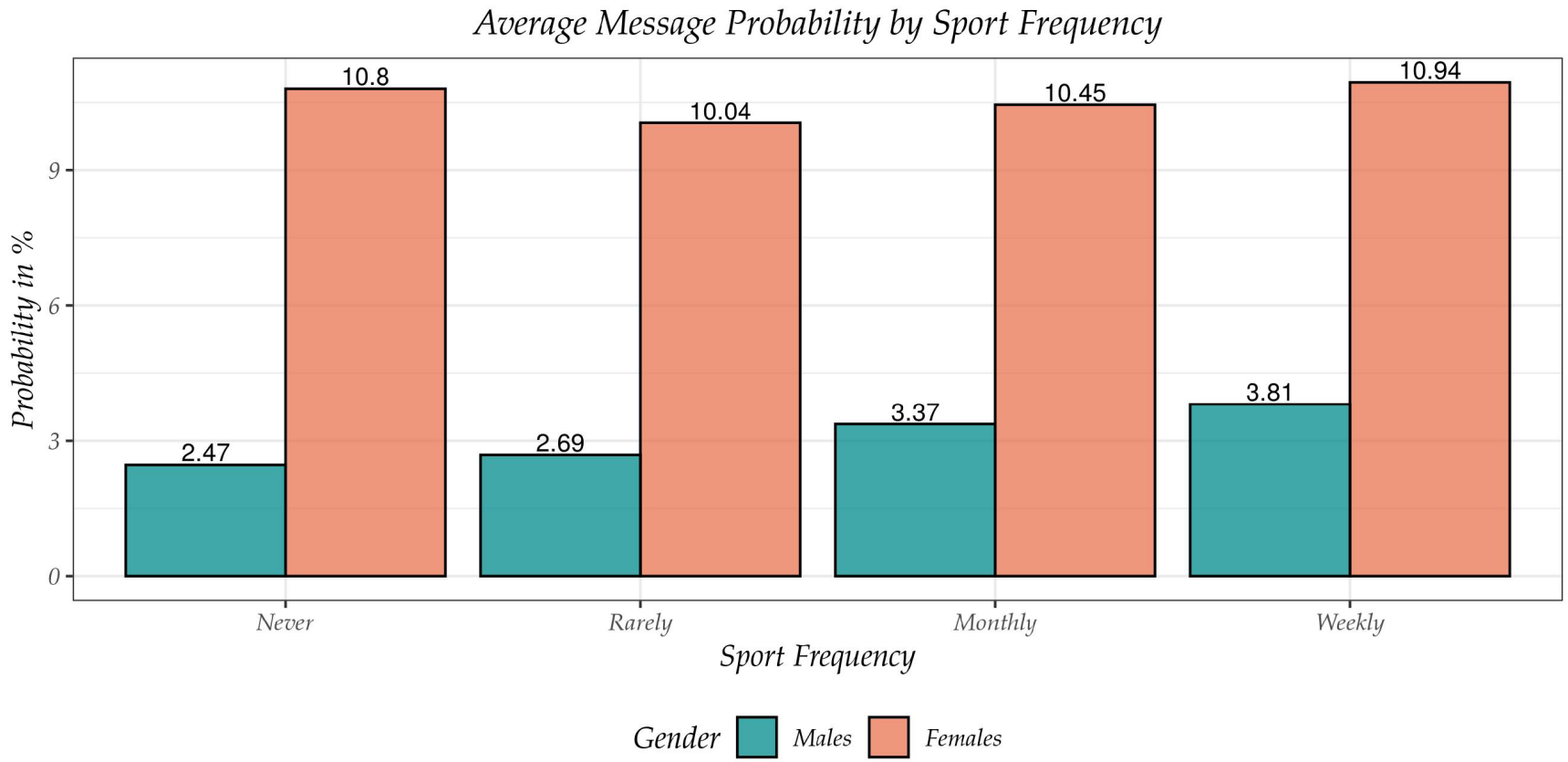
- Selective sample
  - Our sample is younger (average age 40 compared to 44) & better educated
  - Our sample is looking for a partner *(does this matter?)*
- Lying





Descriptive stat's of treatment variable

Figure 1: Average Contact Chances according to Sport Activity for Males and Females







Summary | Data & variables

Our sample

- 18k newly registered users forming over 170k interactions
  - Observation: When profile has been visited by somebody (in 4 month window)

Outcome

- Decision to contact a user by sending the first message after profile visit

Treatment

- Sport frequency (weekly, monthly, rarely, never)

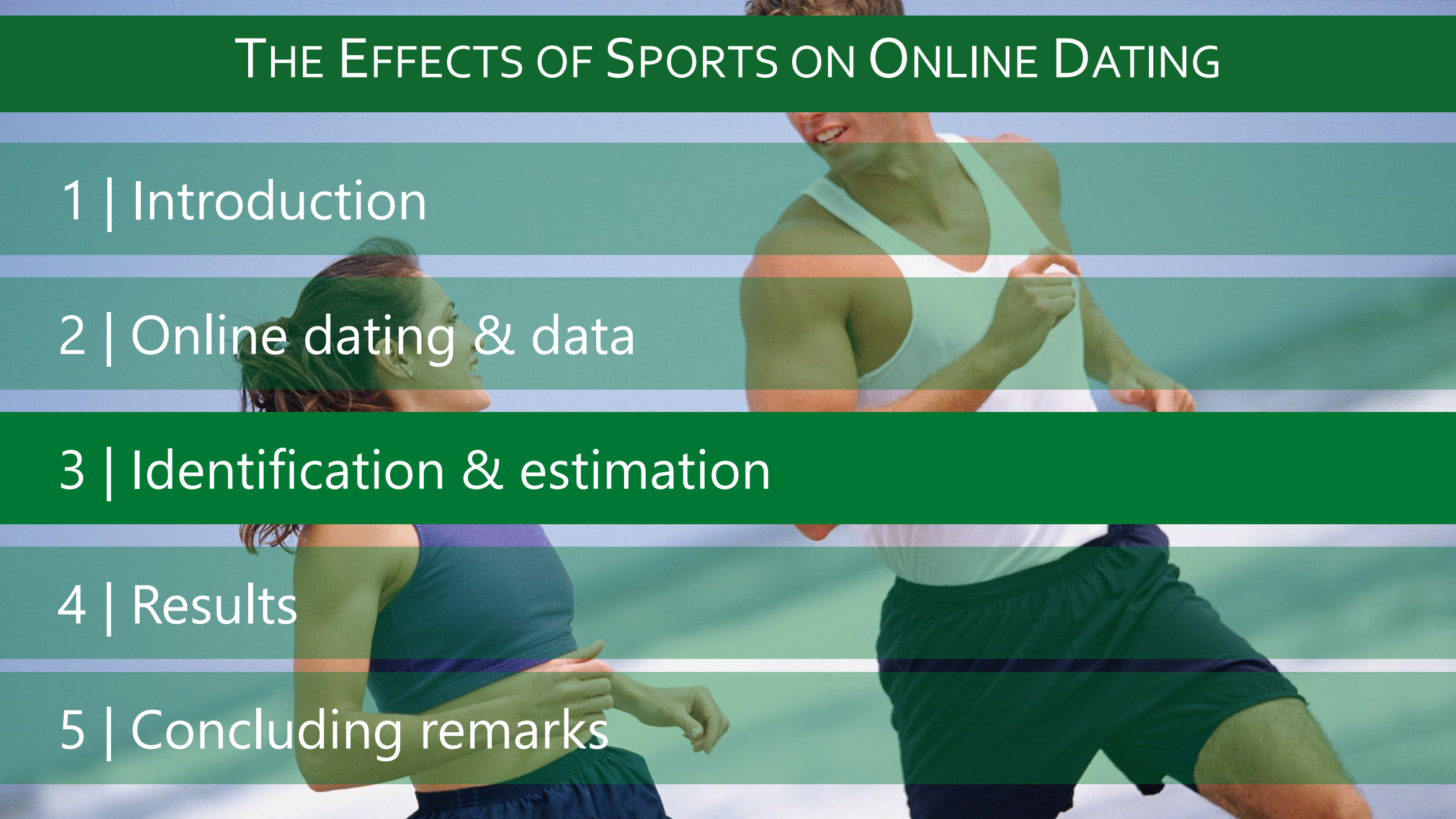
Confounders

- More than 600 user specific characteristics

Heterogeneity

- Age, income, education

The Effects of Sports on Online Dating



1 | Introduction

2 | Online dating & data

3 | Identification & estimation

4 | Results

5 | Concluding remarks





Mean effects – some examples

Individualized Average Treatment Effects

$$IATE(m,l;x) = E(Y^m - Y^l \mid X = x); \quad \forall m \neq l \in \{0,\ldots,M-1\}$$

Group Average Treatment Effects

$$GATE(m,l;z) = E(Y^m - Y^l \mid Z = z) = \int IATE(m,l;x) f_{X|Z=z}(x)dx$$

Average Treatment Effects

$$ATE(m,l) \quad = E(Y^m - Y^l) \quad = \int IATE(m,l;x) f_X(x)dx$$

IATEs and GATEs are specific Conditional ATEs (CATEs)

D: M different treatments (M -1 programmes $\oplus$ non-participation)

Y^m: Potential outcome for state m

Y: Observable outcome

X: Observable features (confounders & heterogeneity variables of interest)

$Z \subseteq X$: Heterogeneity features





Identification

Selection on observables (unconfoundedness, CIA, ...):

$$Y^m, Y^l \coprod D \mid X = x; \quad \forall m \neq l \in \{0, ..., M-1\}$$

IATEs are identified as a difference of conditional means

$$IATE(m, l; x) = E(Y \mid X = x, D = m) - E(Y \mid X = x, D = l)$$

IATE is identified ➔ GATE & ATE are identified





Identification

Why is *unconfoundedness* plausible in our setting?

- Our information set is the same as for the potential sender
- All variables that appear in the profile (& are thus visible to any sender) are available to us
  - BUT: Remove variables that are influenced by sports
    - All sports or sports-equipment related, holiday related, leisure time preferences, body & clothing related





Comprehensive vs. parameter specific approaches

Huge literature on how to best estimate each parameter by some specific causal machine learning method

- For IATEs see e.g. Knaus, Lechner, Strittmatter (2021): "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence," *The Econometrics Journal*.
- For GATE's, e.g. Zimmert, Lechner (2020), …
- For ATE's …

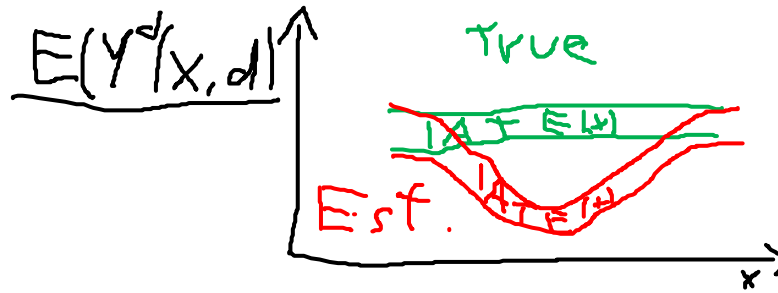
Not attractive to use different estimation concepts for different parameters

- Computationally intensive
  - Many estimators, lot's of different tuning parameters
- Lot's of intellectual effort to understand specifics of estimator & monitor problems
- There may be a lack of internal consistency (GATEs may not add up to ATE etc.)
- …





Comprehensive methodologies | 2



Modified **C**ausal **F**orest (MCF)

- Main idea
  - Random forest estimators can be expressed as weighted sum of outcomes
  - This is also true for causal forests
- Structure of estimators
  - Estimate IATE using a (modified) causal forest
  - Aggregate the weights to get GATEs & ATEs
- General theory for IATEs
  - Wager, S., and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests", *JASA*
- MCF
  - Lechner (2018): Modified Causal Forest, arXiv (currently revised & theoretical analysis added)





Modified Causal Forest (MCF)

CF of Wager & Athey (2018)

+ Improved splitting rule

+ Aggregation scheme to obtain other parameters

+ Weight based inference

The Effects of Sports on Online Dating



1 | Introduction

2 | Online dating & data

3 | Identification & estimation

4 | Results

5 | Concluding remarks





Split samples w.r.t. sex

We expect considerable heterogeneity of male & women

- Male sample (women are senders; men are recipients)
- Female sample (men are senders; women are recipients)





Results | ATE | 1

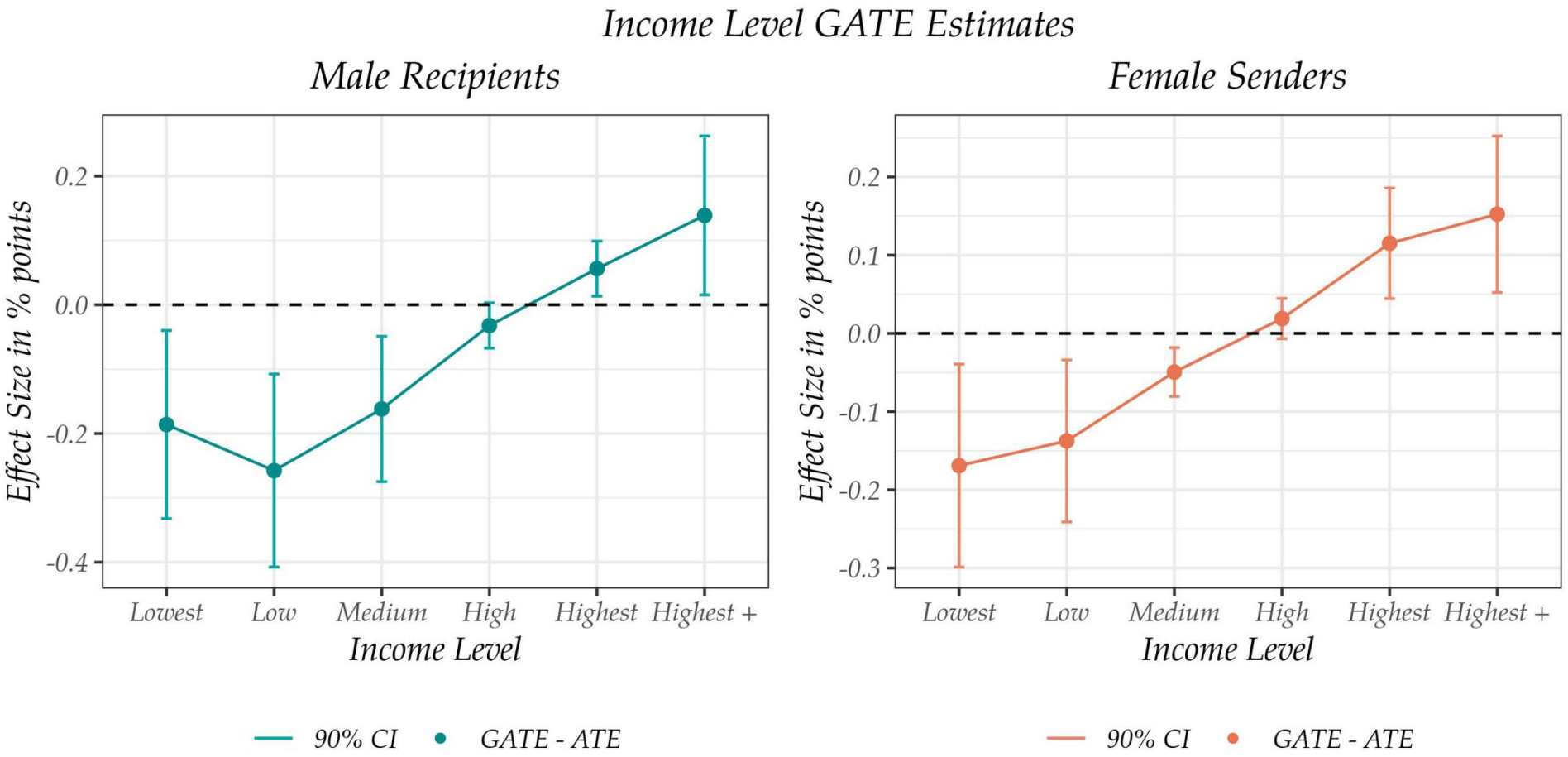
Table 2: Average Effects of Sport Activity on the Contact Chances for Males and Females

	Males				Females			
	Never	*Rarely*	*Monthly*	*Weekly*	*Never*	*Rarely*	*Monthly*	*Weekly*
Never	2.50 (0.46)				10.67 (0.60)			
Rarely	0.12 (0.63)	2.62 (0.43)			-1.30 (0.88)	9.37 (0.63)		
Monthly	0.86 (0.52)	0.74 (0.50)	3.36 (0.26)		-0.29 (0.71)	1.01 (0.72)	10.38 (0.40)	
Weekly	**1.32***** (0.50)	**1.20**** (0.47)	0.46 (0.32)	3.82 (0.19)	0.31 (0.67)	**1.61**** (0.68)	0.60 (0.45)	10.98 (0.30)

Note: Effects in % points. Potential outcomes on the diagonal. Standard errors in parentheses. Significance levels refer to: $^{***} < 0.01$, $^{**} < 0.05$, $^{*} < 0.1$. Color intensity represents the corresponding level sizes.

- Clear-cut results for men approached by women (increase by 50%)
- Much less clear-cut for women …

Figure 2: Heterogeneous Effects of Sport Activity based on Income for Males



Effect of sports is higher for men with higher income

- Probably due to increased interest of women with higher income





Results | GATEs | Summary

We investigated GATEs for

- **Income**
- Education
- **Sport frequency of sender**
- Distance between sender & recipient

Effect of sports is higher for male recipients with higher income

- Probably due to increased interest of women with higher income





Results | IATEs | 1

Investigating the IATEs for heterogeneity

- Reveals the already seen heterogeneity w.r.t. income
  - & variables related to income (education etc.)
- No additional clear-cut results

The Effects of Sports on Online Dating



1 | Introduction

2 | Online dating & data

3 | Identification & estimation

4 | Results

5 | Concluding remarks





Conclusions | 1

Investigate effects of sports on contact chances in online dating

Sport activity of women do not matter (much) for men

Sport activity of men increase their potential chances of being contacted by women

- Even more so when they are richer





Conclusions | 2

Doing more sports is good for many reasons

- Physical & mental health, earnings, life satisfaction, …

For men, it even increases the chances to find love ♥

University of St.Gallen
Swiss Institute for
Empirical Economic Research

Thank you for your attention!

Michael Lechner
University of St. Gallen | SEW
Michael.Lechner@unisg.ch
www.michael-lechner.eu

Paper of this talk & **MCF** paper is on arXiv.

MCF: **PYTHON** code available on **PyPI (*maintained*).**

MCF: R code available on request.







Additional Econometrics





Comprehensive methodologies | 1

Double / debiased machine learning (DML)

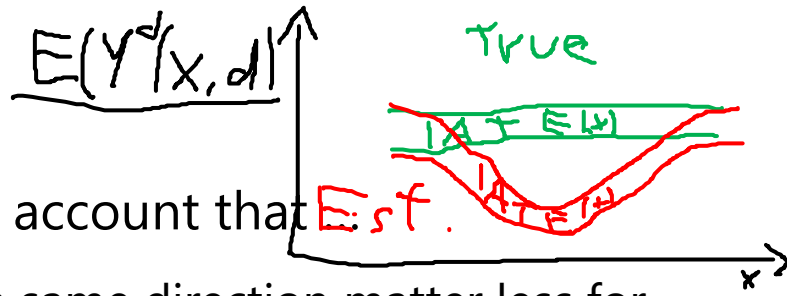
» Example of wide applicability of DML: prev. paper by Martin

- Main idea
  - Different estimators for different parameters, but all use same *doubly robust score*
    - DRS is based on some machine learning estimation of (high-dim.) *'nuisance functions'*
    - Still some issues about different tuning parameters (smoothing), sample splitting, etc. for different para's
- General theory
  - Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins (2018): "Double/debiased machine learning for treatment and structural parameters", *Econometrics Journal*
- How to use this for 'our' parameters
  - Knaus (2020): Double Machine Learning Based Program Evaluation under Unconfoundedness, arXiv.





Modifications of Causal Forest by MCF



Change objective function used to build trees, by taking into account that …

- … errors in estimating $E(Y^1|X=x, D=1)$ & $E(Y^0|X=x, D=0)$ that go in same direction matter less for MSE of treatment effect
- … propensity score plays an important role in reducing selection bias

Exploit that every Random & Causal Forest estimator is a weighted mean of *Y*

- Estimate effects at finest granularity & subsequently aggregate weights to obtain effects at higher aggregation levels

  ➔ same estimation method used for all aggregation levels
- Use these weights for inference at all aggregation levels





Additional descriptive stat's

University of St.Gallen

Swiss Institute for
Empirical Economic Research

Additional materia

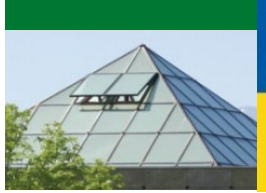




Table 5: Descriptive Statistics for the User Sample

	Mean	SD	Min	Max
Sport Frequency				
Never	0.13	0.34	0.00	1.00
Rarely	0.10	0.31	0.00	1.00
Monthly	0.29	0.45	0.00	1.00
Weekly	0.48	0.50	0.00	1.00
Demographic Features				
Gender (=1 if female)	0.48	0.50	0.00	1.00
Age (in years)	40.05	11.41	18.00	82.00
Income Level				
Lowest	0.12	0.32	0.00	1.00
Low	0.16	0.37	0.00	1.00
Medium	0.20	0.40	0.00	1.00
High	0.24	0.43	0.00	1.00
Highest	0.22	0.42	0.00	1.00
Highest+	0.06	0.24	0.00	1.00
Education Level				
Lowest	0.00	0.06	0.00	1.00
Low	0.08	0.27	0.00	1.00
Medium	0.37	0.48	0.00	1.00
High	0.16	0.37	0.00	1.00
Highest	0.39	0.49	0.00	1.00

Note: Main variables describing the population displayed.



Additional material

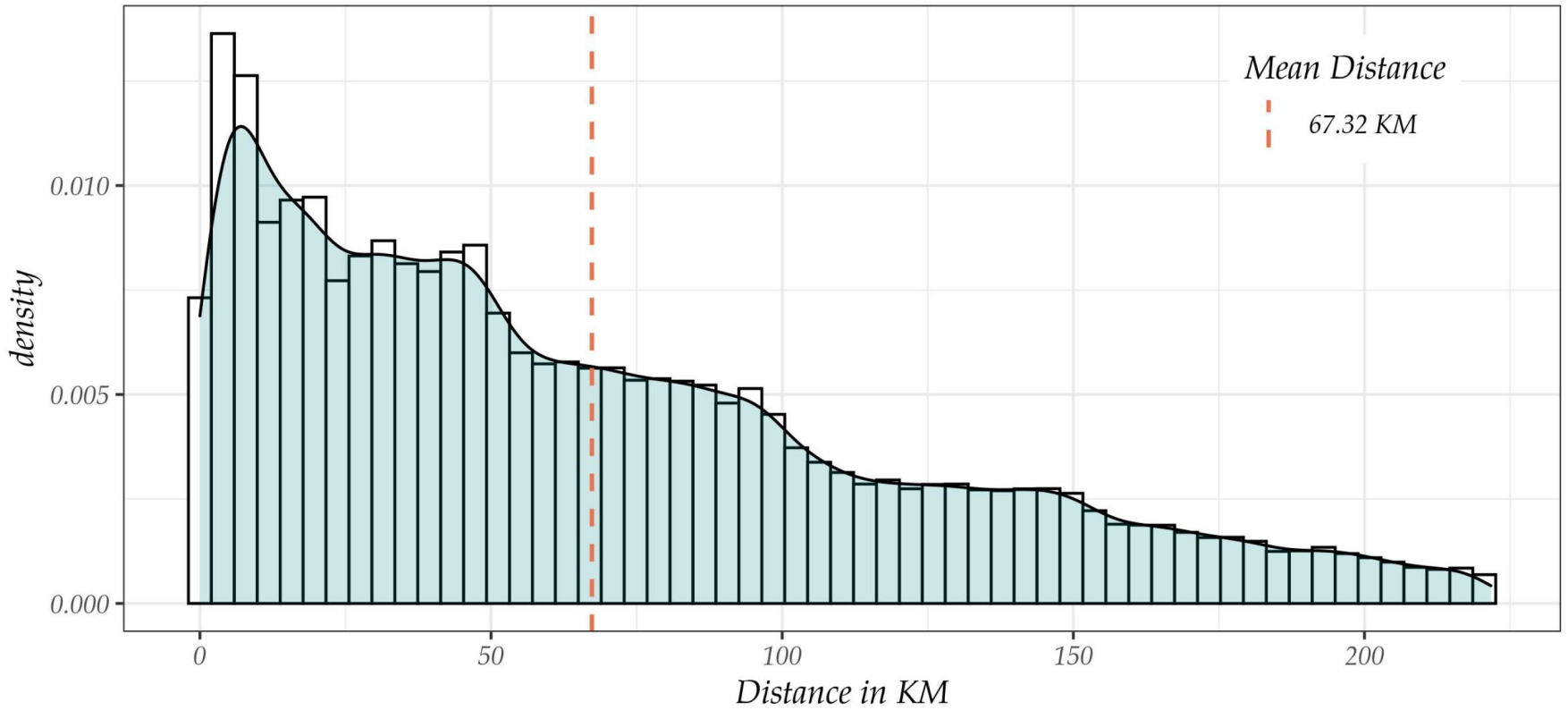


Figure 5: Distribution of the Distance between Users





Additional material: *Valid* user interactions

One-way user interactions

- Visit to profile by potential sender (invisible to potential recipient)
- Next step
  - Action by sender which is visible by recipient (treatment!)
  - No action by sender: **End**

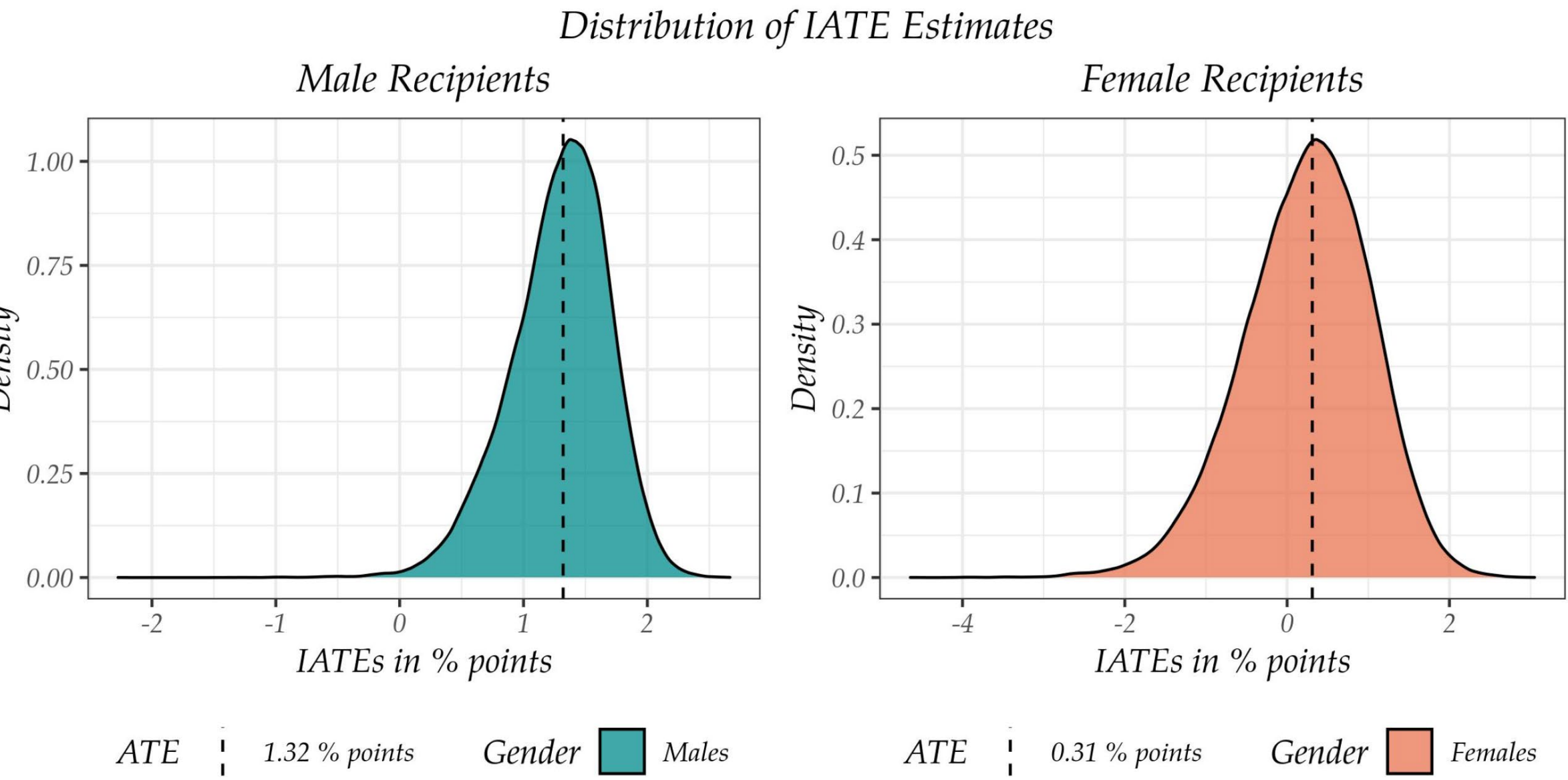
Valid interactions

- Sender visit followed by any visible action of sender
- Sender & recipient visit followed by any visible action of sender
- Only first time interactions are valid





Additional results for IATEs: Distribution & Cluster analysis


Distribution of IATE Estimates
Male Recipients
Female Recipients
Density
1.00
0.75
0.50
0.25
0.00
0.5
0.4
0.3
0.2
0.1
0.0
-4
-2
-1
0
1
2
IATEs in % points
ATE
1.32 % points
Gender
Males
ATE
0.31 % points
Gender
Females

Resu

Table 3: Clusters of the Individualized Effects of Sport Activity for Males and Females

	Males					Females				
Clusters	1	2	3	4	5	1	2	3	4	5
IATEs: Weekly vs. Never	0.41	0.88	1.22	1.52	1.85	-1.41	-0.52	0.12	0.71	1.38
Recipient Features										
Age	45.77	45.19	44.57	43.80	43.71	33.53	37.07	38.39	38.15	36.90
Education Level	3.51	3.86	4.24	4.58	4.76	3.65	3.74	3.80	3.93	4.05
Income Level	3.69	3.98	4.26	4.57	4.83	2.98	3.16	3.32	3.46	3.53
Sport Frequency	1.85	2.16	2.37	2.46	2.49	1.55	1.92	2.15	2.34	2.43
Sender Features										
Age	43.94	43.09	42.37	41.53	41.43	36.33	40.15	41.68	41.46	40.23
Education Level	3.46	3.69	3.96	4.16	4.34	3.74	3.82	3.93	4.06	4.25
Income Level	2.94	3.25	3.52	3.79	4.07	3.51	3.82	3.99	4.10	4.14
Sport Frequency	1.86	2.03	2.16	2.24	2.32	1.91	2.06	2.15	2.27	2.44
Shared Features										
Distance	70.39	71.65	70.23	70.16	64.91	64.43	66.22	67.10	66.36	62.97
Observations										
Share	0.07	0.18	0.29	0.31	0.15	0.07	0.20	0.29	0.29	0.15
Total	2288	6439	10153	10826	5252	3753	11048	15896	15484	8047

Note: Means of clustered effects sorted in an increasing order, matched with the heterogeneity variables. Color intensity represents the corresponding effect sizes and highlights the relevant GATEs.

Table 10: Descriptive Clusters of IATEs based on the k-means++ Clustering

Code		Males					Females				
	Clusters	1	2	3	4	5	1	2	3	4	5
	IATEs: Weekly vs. Never	0.41	0.88	1.22	1.52	1.85	-1.41	-0.52	0.12	0.71	1.38
	Recipient Features										
5	Smoking Frequency	1.30	0.85	0.32	0.10	0.03	0.50	0.47	0.45	0.37	0.29
14	Relevance of Sexuality	0.38	0.44	0.47	0.49	0.52	0.21	0.27	0.31	0.32	0.31
108	TV in Leisure Time	0.27	0.26	0.25	0.25	0.28	0.32	0.27	0.21	0.15	0.11
223	Radio/TV at Home	0.62	0.62	0.60	0.59	0.60	0.73	0.68	0.65	0.64	0.67
284	Appearance Satisfaction	0.18	0.19	0.20	0.21	0.22	0.11	0.16	0.18	0.20	0.21
292	Importance of Sexuality	0.29	0.33	0.36	0.38	0.40	0.21	0.24	0.28	0.29	0.30
303	Comfortable Dining	0.67	0.63	0.57	0.55	0.55	0.73	0.66	0.60	0.57	0.58
324	Wish Significant Other	0.27	0.28	0.29	0.32	0.33	0.26	0.26	0.26	0.27	0.30
	Sender Features										
5	Smoking Frequency	0.69	0.52	0.36	0.31	0.27	0.48	0.50	0.49	0.42	0.32
14	Relevance of Sexuality	0.24	0.28	0.29	0.33	0.40	0.35	0.40	0.44	0.45	0.46
108	TV in Leisure Time	0.23	0.21	0.20	0.19	0.17	0.29	0.30	0.29	0.27	0.24
223	Radio/TV at Home	0.65	0.61	0.60	0.58	0.55	0.67	0.66	0.64	0.63	0.61
284	Appearance Satisfaction	0.18	0.17	0.17	0.19	0.22	0.12	0.16	0.17	0.19	0.23
292	Importance of Sexuality	0.21	0.24	0.23	0.25	0.28	0.29	0.32	0.33	0.35	0.38
303	Comfortable Dining	0.68	0.65	0.60	0.56	0.54	0.63	0.62	0.60	0.58	0.54
324	Wish Significant Other	0.23	0.23	0.25	0.25	0.27	0.28	0.29	0.31	0.32	0.35
	Observations										
	Share	0.07	0.18	0.29	0.31	0.15	0.07	0.20	0.29	0.29	0.15
	Total	2288	6439	10153	10826	5252	3753	11048	15896	15484	8047

Note: Means of clustered effects sorted in an increasing order, matched with selected user characteristics. Variable codes





Additional Placebo





Placebo test | 1

Some concerns because …

- … sample is based on profile visits
  - which are influenced by value of matching score
    - which is unobservable & may (!) be influenced by sports activity

One way to check ➔ placebo study

- Observations: All potential interactions ➔ 38 million
- **Outcome: Profile visit** (instead of interaction after visit) or not
- Redo analysis with male & female samples
  - Random samples of sizes similar to actual analysis





Results | 1 | Placebo test

Table 4: Average Effects of Sport Activity on the Visit Chances for Males and Females

	Males				Females			
	Never	*Rarely*	*Monthly*	*Weekly*	*Never*	*Rarely*	*Monthly*	*Weekly*
Never	0.23 (0.14)				0.66 (0.15)			
Rarely	0.29 (0.24)	0.52 (0.20)			-0.07 (0.22)	0.59 (0.16)		
Monthly	0.08 (0.16)	-0.22 (0.22)	0.31 (0.08)		-0.20 (0.17)	-0.13 (0.17)	0.46 (0.08)	
Weekly	0.18 (0.15)	-0.12 (0.21)	0.10 (0.10)	0.41 (0.06)	-0.08 (0.16)	-0.01 (0.17)	0.12 (0.10)	0.58 (0.07)

Note: Effects in % points. Potential outcomes on the diagonal. Standard errors in parentheses. Significance levels refer to: *** < 0.01, ** < 0.05, * < 0.1. Color intensity represents the corresponding level sizes.

- No effect of sports on initial visit detectable