Essays on Causal Inference

DOCTORAL THESIS

presented to the Faculty of Management, Economics and Social Sciences at the University of Fribourg, in fulfillment of the requirements for the degree of Doctor of Philosophy in Economics

submitted by

Henrika LANGEN

from Germany

Accepted by the Faculty of Management, Economics and Social Sciences on 26.09.2022 at the proposal of

Prof. Martin Huber, Ph.D. (first supervisor) and

Prof. Robert Lieli, Ph.D. (second supervisor)

Fribourg, 2022

Essays on Causal Inference

DOCTORAL THESIS

presented to the Faculty of Management, Economics and Social Sciences at the University of Fribourg, in fulfillment of the requirements for the degree of Doctor of Philosophy in Economics

submitted by

Henrika LANGEN

from Germany

Accepted by the Faculty of Management, Economics and Social Sciences on 26.09.2022 at the proposal of

Prof. Martin Huber, Ph.D. (first supervisor) and

Prof. Robert Lieli, Ph.D. (second supervisor)

Fribourg, 2022

https://doi.org/10.51363/unifr.eth.2022.004

© Henrika Langen, 2022

The Faculty of Management, Economics and Social Sciences at the University of Fribourg neither approves nor disapproves the opinions expressed in a doctoral thesis. They are to be considered those of the author. (Decision of the Faculty Council of 23 January 1990)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The term “causal inference” refers to the process of drawing conclusions about causal mechanisms from data. Approaches to causal inference are manifold and include not only experimental but also a variety of non-experimental methods for evaluating policies and events based on observational data. By assessing observational data in an appropriate framework, such as the potential outcomes framework (see for instance Neyman (1923) and Rubin (1974)), experimental conditions can be mimicked in order to infer the impact of (policy or business) interventions or events on outcomes of interest. These non-experimental methods of causal inference can provide the basis for effective policy and business decision making even in contexts where experiments are infeasible, e.g., for ethical reasons, due to financial constraints, or simply because the matter under study is too urgent for decision makers to wait for experimental results. In addition to estimating the overall causal impact of an intervention or event, the field of causal inference also comprises deeper analyses of the underlying causal mechanisms, as well as the assessment of interpersonal differences in effects, which in turn allows for the identification of optimal policy targeting strategies.

This dissertation contributes, on the one hand, to the advancement of causal inference methods by developing a machine learning-based framework for causal mediation analysis and by illustrating approaches to inferring causal effects in the context of text data. On the other hand, it applies methods of causal inference to problems in a variety of disciplines, namely epidemic policy making, jurisprudence, and marketing, and demonstrates new approaches to addressing research questions in these areas.

The studies are ordered chronologically, that is, according to the time when the corresponding projects were initiated. Chapter 2 proposes a framework for conducting causal mediation analysis based on double machine learning (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018)). The proposed framework allows for decomposing the causal effect of a treatment on an outcome of interest into a direct effect as well as an indirect effect operating via an intermediate outcome, while controlling for observed confounders in a data-driven way. The effect estimates can be shown to be asymptotically normal and $n^{-1/2}$-consistent under certain regularity conditions. For illustrative purposes, the framework is applied to decompose the total effect of health insurance coverage on general health among young adults in the United States into a direct effect and an indirect effect through routine checkups.

In Chapter 3, we assess the effect of the timing of COVID-19 response measures on COVID-19-related hospitalization and death rates in Germany and Switzerland. We do so by exploiting the fact that the epidemic was more advanced in some regions than in others when certain lockdown measures came into force. We compare hospitalization and death rates across regions with earlier and later epidemic start dates, finding for both countries that a relatively later imposition of lockdown measures entails higher cumulative hospitalization and death rates. An assessment of curfews, as introduced in some German states, provides no evidence that, under the other lockdown measures already in place, curfews are more effective than the federally imposed ban on gatherings. This chapter constitutes a contribution to the COVID-19 literature, which at the early point in the COVID-19 pandemic when our study was published was dominated by studies simulating and predicting the future evolution of infection curves.

In Chapter 4, I assess the impact of the #MeToo movement on language in court by quantifying judicial opinions from U.S. appeal courts and then applying the Difference-in-Difference method as well as an event study approach. The opinions are quantified by means of text vectorization methods as well as by constructing indicators that measure the extent of victim blaming in each opinion. While not revealing significant effects of the movement on the quantifiers under study, this chapter does present approaches to analyzing the arguably most comprehensive data available: text.

Chapter 5, finally, points out the advantages of causal machine learning over predictive machine learning in the context of business decision making. Using a retailer's coupon campaign data as an example, we demonstrate how causal machine learning can leverage observational data from earlier campaigns to evaluate the effectiveness of coupon campaigns and determine the optimal coupon distribution strategy in terms of expected overall revenues. For the coupon campaigns under study, we find that only two out of five coupon categories, namely coupons applicable to the product categories of drugstore items and other food, show a significant purchase-increasing effect. The assessment of group average treatment effects reveals substantial differences in the impact of coupon provision across customer groups, particularly across customer groups as defined by pre-campaign spending.

Chapter 2

Causal Mediation Analysis with Double Machine Learning

with Helmut Farbmacher, Martin Huber, Lukáš Laffers and Martin Spindler

Abstract: This paper combines causal mediation analysis with double machine learning for a data-driven control of observed confounders in a high-dimensional setting. The average indirect effect of a binary treatment and the unmediated direct effect are estimated based on efficient score functions, which are robust w.r.t. misspecifications of the outcome, mediator, and treatment models. This property is key for selecting these models by double machine learning, which is combined with data splitting to prevent overfitting. We demonstrate that the effect estimators are asymptotically normal and $n^{-1/2}$-consistent under specific regularity conditions and investigate the finite sample properties of the suggested methods in a simulation study when considering lasso as machine learner. We also provide an empirical application to the U.S. National Longitudinal Survey of Youth, assessing the indirect effect of health insurance coverage on general health operating via routine checkups as mediator, as well as the direct effect.

Keywords: mediation, direct and indirect effects, causal mechanisms, double machine learning, efficient score.

JEL classification: C21.

This article is published in The Econometrics Journal as Farbmacher, Huber, Laffers, Langen, and Spindler (2022).
Lukáš Laffers acknowledges support provided by the Slovak Research and Development Agency under contract no. APVV-17-0329 and VEGA-1/0692/20. Helmut Farbmacher and Martin Spindler acknowledge support from the German Research Foundation (DFG), Germany - 431701914.

2.1 Introduction

Causal mediation analysis aims at decomposing the causal effect of a treatment on an outcome of interest into an indirect effect operating through a mediator (or intermediate outcome) and a direct effect comprising any causal mechanisms not operating through that mediator. Even if the treatment is random, direct and indirect effects are generally not identified by naively controlling for the mediator without accounting for its likely endogeneity, see Robins and Greenland (1992). While much of the earlier literature either neglected endogeneity issues or relied on restrictive linear models, see for instance Cochran (1957), Judd and Kenny (1981), and Baron and Kenny (1986), more recent contributions consider more general identification approaches using the potential outcome framework. Some of the numerous examples are Robins and Greenland (1992), Pearl (2001), Robins (2003), Petersen, Sinisi, and van der Laan (2006), VanderWeele (2009), Imai, Keele, and Yamamoto (2010), Hong (2010), Albert and Nelson (2011), Imai and Yamamoto (2013), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt, Bekaert, and Lange (2012), and Huber (2014). Using the denomination of Pearl (2001), the literature distinguishes between natural direct and indirect effects, where mediators are set to their potential values 'naturally' occurring under a specific treatment assignment, and the controlled direct effect, where the mediator is set to a 'prescribed' value.

The vast majority of identification strategies relies on selection-on-observable-type assumptions implying that the treatment and the mediator are conditionally exogenous when controlling for observed covariates. Empirical examples in economics and policy evaluation include Flores and Flores-Lagunes (2009), Heckman, Pinto, and Savelyev (2013), Keele, Tingley, and Yamamoto (2015), Conti, Heckman, and Pinto (2016), Huber (2015), Huber, Lechner, and Mellace (2017), Bellani and Bia (2018), Bijwaard and Jones (2018), and Huber, Lechner, and Strittmatter (2018). Such studies typically rely on the (implicit) assumption that the covariates to be controlled for can be unambiguously preselected by the researcher, for instance based on institutional knowledge or theoretical considerations. This assumes away uncertainty related to model selection w.r.t. covariates to be included and entails incorrect inference under the common practice of choosing and refining the choice of covariates based on their predictive power.

To improve upon this practice, this paper combines causal mediation analysis based on efficient score functions, see Tchetgen Tchetgen and Shpitser (2012), with double machine learning as outlined in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for a data-driven control of observed confounders to obtain valid inference under specific regularity conditions. In particular, one important condition is that the number of important confounders (that make the selection-on-observables assumptions to hold approximately) is not too large relative to the sample size. However, the set of these important confounders need not be known a priori and the set of potential confounders can be even larger than the sample size.[1] This is particularly useful in high dimensional data with a vast number of covariates that could potentially serve as control variables, which can render researcher-based covariate selection complicated if not infeasible. We demonstrate $n^{-1/2}$-consistency and asymptotic normality of the proposed effect estimators under specific regularity conditions by verifying that the general framework of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) for well-behaved double machine learning is satisfied in our context.

Tchetgen Tchetgen and Shpitser (2012) suggest estimating natural direct and indirect effects based on the efficient score functions of the potential outcomes, which requires plug-in estimates for the conditional mean outcome, mediator density, and treatment probability. Analogous to doubly robust estimation of average treatment effects, see Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995), the resulting estimators are semiparametrically efficient if all models of the plug-in estimates are correctly specified and remain consistent even if one model is misspecified. We show that the efficient score function of Tchetgen Tchetgen and Shpitser (2012) satisfies the so-called Neyman (1959) orthogonality discussed in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), which makes the estimation of direct and indirect effects rather insensitive to (local) estimation errors in the plug-in estimates. We transform the score function of Tchetgen Tchetgen and Shpitser (2012) by an application of Bayes' Law in a way that it avoids the estimation of the conditional mediator density, as discussed in Zheng and van der Laan (2012) and also adopted by Díaz and Hejazi (2020), and show it to be Neyman orthogonal. This appears particularly useful when the mediator is a vector of variables and/or continuous, making conditional mediator density estimation cumbersome. Further, we establish the score function required for estimating the controlled direct effect along with Neyman orthgonality.

[1] Different from conventional semiparametric methods, the double machine learning framework does not require the set of potential confounders to be restricted by Donsker conditions, but permits the set to be unbounded and to grow with the sample size.

Neyman orthgonality is key for the fruitful application of double machine learning, ensuring robustness in the estimation of the nuisance parameters which is crucial when applying modern machine learning methods. Random sample splitting – to estimate the parameters of the plug-in models in one part of the data, while predicting the score function and estimating the direct and indirect effects in the other part – avoids overfitting the plug-in models (e.g. by controlling for too many covariates). It increases the variance by only using part of the data for effect estimation. This is avoided by cross-fitting which consists of swapping the roles of the data parts for estimating the plug-in models and the treatment effects to ultimately average over the effect estimates in either part. When combining efficient score-based effect estimation with sample splitting, $n^{-1/2}$-convergence of treatment effect estimation can be obtained under a substantially slower convergence of $n^{-1/4}$ for the plug-in estimates, see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). Under specific regularity conditions, this convergence rate can attained by various machine learning algorithms including lasso regression, see Tibshirani (1996).

We investigate the estimators' finite sample behaviour based on the score function of Tchetgen Tchetgen and Shpitser (2012) and the alternative score suggested in this paper when using post-lasso regression as machine learner for the plug-in estimates. Furthermore, we apply our method to data from the National Longitudinal Survey of Youth 1997 (NLSY97) conducted by the Bureau of Labor Statistics at the U.S. Department of Labor (2019), where a large set of potential control variables is available. We disentangle the short-term effect of health insurance coverage on general health into an indirect effect which operates via the incidence of a routine checkup in the last year and a direct effect covering any other causal mechanisms. While we find a moderate, though statistically insignificant health-improving direct effect, the indirect effect is very close to zero. We therefore do not find evidence that health insurance coverage affects general health through routine checkups in the short run.

We note that basing estimation on efficient score functions is not the only framework satisfying the previously mentioned robustness w.r.t. estimation errors in plug-in parameters. This property is also satisfied by the targeted maximum likelihood estimation (TMLE) framework by van der Laan and Rubin (2006), see the discussion in Díaz (2020). TMLE relies on iteratively updating (or robustifying) an initial estimate of the parameter of interest based on regression steps that involve models for the plug-in parameters. Zheng and van der Laan (2012) have developed an estimation approach for natural direct and indirect effects using TMLE, where the plug-in parameters might by estimated by machine learners, e.g. the super learner, an ensemble method suggested by van der Laan, Polley, and Hubbard (2007). This iterative estimation approach is therefore an alternative to the double machine learning-based approach suggested in this paper, for which we demonstrate $n^{-1/2}$-consistency under specific conditions.

This paper proceeds as follows. Section 2.2 introduces the concepts of direct and indirect effect identification in the potential outcome framework. In Section 2.3, we present the identifying assumptions and discuss identification based on efficient score functions. Section 2.4 proposes an estimation procedure based on double machine learning and shows $n^{-1/2}$-consistency and asymptotic normality under specific conditions. Section 2.5 provides a simulation study. Section 2.6 presents an empirical application to data from the NLSY97. Section 2.7 concludes.

2.2 Definition of direct and indirect effects

We aim at decomposing the average treatment effect (ATE) of a binary treatment, denoted by D, on an outcome of interest, Y, into an indirect effect operating through a discrete mediator, M, and a direct effect that comprises any causal mechanisms other than through M. We use the potential outcome framework, see for instance Rubin (1974), to define the direct and indirect effects of interest, see also Ten Have, Joffe, Lynch, Brown, Maisto, and Beck (2007) and Albert (2008) for further examples in the context of mediation. $M(d)$ denotes the potential mediator under treatment value $d \in \{0, 1\}$, while $Y(d, m)$ denotes the potential outcome as a function of both the treatment and some value m of the mediator M.[2] The observed outcome and mediator correspond to the respective potential variables associated with the actual treatment assignment, i.e. $Y = D\cdot Y(1, M(1))+(1-D)\cdot Y(0, M(0))$ and $M = D\cdot M(1)+(1-D)\cdot M(0)$, implying that any other potential outcomes or mediators are a priori (i.e. without further statistical assumptions) unknown.

We denote the ATE by $\Delta = E[Y(1, M(1)) - Y(0, M(0))]$, which comprises both direct and indirect effects. To decompose the latter, note that the average direct effect, denoted by $\theta(d)$, equals the difference in mean potential outcomes when switching the treatment while keeping

[2] Throughout this paper, capital letters denote random variables and small letters specific values of random variables.

the potential mediator fixed, which blocks the causal mechanism via M:

$$\theta(d) \quad = \quad E[Y(1,M(d)) - Y(0,M(d))], \quad d \in \{0,1\}. \tag{2.1}$$

The (average) indirect effect, $\delta(d)$, equals the difference in mean potential outcomes when switching the potential mediator values while keeping the treatment fixed to block the direct effect.

$$\delta(d) \quad = \quad E[Y(d,M(1)) - Y(d,M(0))], \quad d \in \{0,1\}. \tag{2.2}$$

Robins and Greenland (1992) and Robins (2003) referred to these parameters as pure/total direct and indirect effects, Flores and Flores-Lagunes (2009) as net and mechanism average treatment effects, and Pearl (2001) as natural direct and indirect effects, which is the denomination used in the remainder of this paper.

The ATE is the sum of the natural direct and indirect effects defined upon opposite treatment states d, which can be easily seen from adding and subtracting the counterfactual outcomes $E[Y(0,M(1))]$ and $E[Y(1,M(0))]$:

$$\begin{aligned}\Delta &= E[Y(1,M(1)) - Y(0,M(0))] \\ &= E[Y(1,M(1)) - Y(0,M(1))] + E[Y(0,M(1)) - Y(0,M(0))] = \theta(1) + \delta(0) \\ &= E[Y(1,M(0)) - Y(0,M(0))] + E[Y(1,M(1)) - Y(1,M(0))] = \theta(0) + \delta(1). \end{aligned} \tag{2.3}$$

The distinction between $\theta(1)$ and $\theta(0)$ as well as $\delta(1)$ and $\delta(0)$ hints to the possibility of heterogeneous effects across treatment states d due to interaction effects between D and M. For instance, the direct effect of health insurance coverage (D) on general health (Y) might depend on whether or not a person underwent routine check-ups (M). We note that a different approach to dealing with the interaction effects between D and M is a three-way decomposition of the ATE into the pure direct effect ($\theta(0)$), the pure indirect effect ($\delta(0)$) and the mediated interaction effect, see VanderWeele (2013).

The so-called controlled direct effect, denoted by $\gamma(m)$, is a further parameter that received much attention in the mediation literature. It corresponds to the difference in mean potential outcomes when switching the treatment and fixing the mediator at some value m:

$$\gamma(m) = E[Y(1,m) - Y(0,m)], \qquad \text{for } m \text{ in the support of } M. \tag{2.4}$$

In contrast to $\theta(d)$, which is conditional on the potential mediator value 'naturally' realized for treatment d which may differ across subjects, $\gamma(m)$ is conditional on enforcing the same mediator state in the entire population. The two parameters are only equivalent in the absence of an interaction between D and M. Whether the natural or controlled direct effect is more relevant depends on the feasibility and desirability to intervene on or prescribe the mediator, see Pearl (2001) for a discussion of the 'descriptive' and 'prescriptive' natures of natural and controlled effects. There is no indirect effect parameter matching the controlled direct effect, implying that the difference between the total effect and the controlled direct effect does in general not correspond to the indirect effect, unless there is no interaction between D and M, see e.g. Kaufman, MacLehose, and Kaufman (2004).

2.3 Assumptions and identification

Our identification strategy is based on the assumption that confounding of the treatment-outcome, treatment-mediator, and mediator-outcome relations can be controlled for by conditioning on observed covariates, denoted by X. The latter must not contain variables that are influenced by the treatment, such that X is typically evaluated prior to treatment assignment. Figure 2.1 provides a graphical illustration using a directed acyclic graph, with arrows representing causal effects. Each of D, M, and Y might be causally affected by distinct and statistically independent sets of unobservables not displayed in Figure 2.1, but none of these unobservables may jointly affect two or all three elements (D, M, Y) conditional on X.

Formally, the first assumption invokes conditional independence of the treatment and potential mediators or outcomes given X. This restriction has been referred to as conditional independence, selection on observables, or exogeneity in the treatment evaluation literature, see e.g. Imbens (2004). This rules out confounders jointly affecting the treatment on the one hand and the mediator and/or the outcome on the other hand conditional on X. In non-experimental data, the plausibility of this assumption critically hinges on the richness of X.

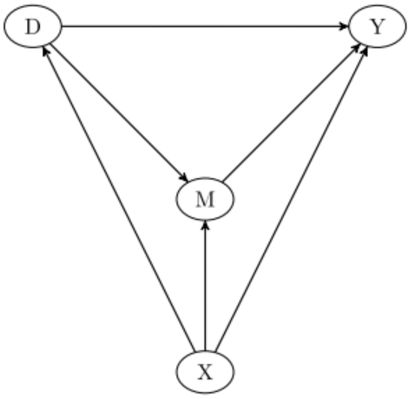


Figure 2.1: Causal paths under conditional exogeneity given pre-treatment covariates

Assumption 1 (conditional independence of the treatment):

$\{Y(d',m),M(d)\} \perp D|X=x$ for all $d',d \in \{0,1\}$ and m,x in the support of M,X, where '$\perp$' denotes statistical independence.

The second assumption requires the mediator to be conditionally independent of the potential outcomes given the treatment and the covariates.

Assumption 2 (conditional independence of the mediator):

$Y(d',m) \perp M|D=d,X=x$ for all $d',d \in \{0,1\}$ and m,x in the support of M,X.

Assumption 2 rules out confounders jointly affecting the mediator and the outcome conditional on D and X. If X is pre-treatment (as is common to avoid controlling for variables potentially affected by the treatment), this implies the absence of post-treatment confounders of the mediator-outcome relation. Such a restriction needs to be rigorously scrutinized and appears for instance less plausible if the time window between the measurement of the treatment and the mediator is large in a world of time-varying variables.

The third assumption imposes common support on the conditional treatment probability across treatment states.

Assumption 3 (common support):

$\Pr(D=d|M=m,X=x)>0$ for all $d \in \{0,1\}$ and m,x in the support of M,X.

The common support assumption, also known as positivity or covariate overlap assumption, restricts the conditional probability to be or not be treated given M,X, henceforth referred to as propensity score, to be larger than zero. It implies the weaker condition that $\Pr(D=d|X=x)>0$ such that the treatment must not be deterministic in X, otherwise no comparable units in terms of X are available across treatment states. By Bayes' Law, Assumption 3 also implies that $\Pr(M = m|D = d, X = x) > 0$ if M is discrete or that the conditional density of M given D, X is larger than zero if M is continuous. Conditional on X, the mediator state must not be deterministic in the treatment, otherwise no comparable units in terms of the treatment are available across mediator states. Assumptions 1 to 3 are standard in the causal mediation literature, see for instance Imai, Keele, and Yamamoto (2010), Tchetgen Tchetgen and Shpitser (2012), Vansteelandt, Bekaert, and Lange (2012), and Huber (2014), or also Pearl (2001), Petersen, Sinisi, and van der Laan (2006), and Hong (2010), for closely related restrictions.

We identify the counterfactual $E[Y(d, M(1-d))]$ based on the following lemma proven by Tchetgen Tchetgen and Shpitser (2012).

Lemma 1:
Under Assumptions 1-3, the counterfactual $E[Y(d, M(1-d))]$ is identified by the following efficient score function:

$$\begin{aligned} E[Y(d, M(1-d))] &= E[\psi_d], \\ \text{with } \psi_d &= \frac{I\{D=d\} \cdot f(M|1-d, X)}{p_d(X) \cdot f(M|d, X)} \cdot [Y - \mu(d, M, X)] \\ &\quad + \frac{I\{D=1-d\}}{1-p_d(X)} \cdot \Big[\mu(d, M, X) - \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\, dm\Big] \\ &\quad + \int_{m \in \mathcal{M}} \mu(d, m, X) \cdot f(m|1-d, X)\, dm \end{aligned} \tag{2.5}$$

where $f(M|D, X)$ denotes the conditional density of M given D and X (if M is discrete, this is a conditional probability and integrals need to be replaced by sums), $p_d(X) = \Pr(D = d|X)$ the probability of treatment $D = d$ given X, and $\mu(D, M, X) = E(Y|D, M, X)$ the conditional expectation of outcome Y given D, M, and X.

(2.5) satisfies a multiple robustness property in the sense that estimation remains consistent even if one out of the three models for the plug-in parameters $f(M|D, X)$, $p_d(X)$, and $\mu(D, M, X)$ is misspecified.

To derive an alternative expression for identification, note that by Bayes' Law,

$$\begin{aligned} \frac{f(M|1-d, X)}{p_d(X) \cdot f(M|d, X)} &= \frac{\big(1 - p_d(M, X)\big) \cdot f(M|X)}{1 - p_d(X)} \cdot \frac{p_d(X)}{p_d(M, X) \cdot f(M|X) \cdot p_d(X)} \\ &= \frac{1 - p_d(M, X)}{p_d(M, X) \cdot \big(1 - p_d(X)\big)} \end{aligned}$$

where $f(M|X)$ is the conditional distribution of M given X and $p_d(X, M) = \Pr(D = d|X, M)$. Furthermore,

$$\int \mu(d, m, X) \cdot f(m|1-d, X)dm = E\Big[\mu(d, M, X)\Big|D = 1-d, X\Big].$$

As also noticed in Zheng and van der Laan (2012), the counterfactual can as well be identified based on an alternative multiply robust representation of (2.5) as provided in the following lemma.

Lemma 2:

Under Assumptions 1-3, the counterfactual $E[Y(d, M(1-d))]$ is identified by the following alternative efficient score function:

$$\begin{aligned}
E[Y(d, M(1-d))] &= E[\psi_d^*], \\
\text{with } \psi_d^* &= \frac{I\{D = d\} \cdot (1 - p_d(M, X))}{p_d(M, X) \cdot (1 - p_d(X))} \cdot [Y - \mu(d, M, X)] \\
&+ \frac{I\{D = 1-d\}}{1 - p_d(X)} \cdot \Big[\mu(d, M, X) - E\Big[\mu(d, M, X)\Big|D = 1-d, X\Big]\Big] \\
&+ E\Big[\mu(d, M, X)\Big|D = 1-d, X\Big]. \qquad (2.6)
\end{aligned}$$

Similarly as the approaches based on inverse probability weighting (rather than efficient scores) in Huber (2014) and Tchetgen Tchetgen (2013), (2.6) avoids conditional mediator densities, which appears attractive if M is continuous and/or multidimensional. On the other hand, it requires the estimation of an additional parameter, namely the nested conditional mean $E[\mu(d, M, X)|D = 1-d, X]$, as similarly found in Miles, Shpitser, Kanki, Meloni, and Tchetgen Tchetgen (2020), who suggest a multiply robust score function for assessing path-specific effects. Alternatively to rearranging the score function by Tchetgen Tchetgen and Shpitser (2012) as outlined above, ratios of conditional densities as for instance appearing in the first component of (2.5) might be treated as additional nuisance parameter and estimated directly via density-ratio estimation, see e.g. Sugiyama, Kawanabe, and Chui (2010) for density-ratio estimation in high-dimensional settings. Such methods based on directly estimating the density ratio without going through estimating the densities in numerator and denominator separately are shown in several studies to compare favourably with estimating the densities separately, see e.g. Kanamori, Suzuki, and Sugiyama (2012).

Efficient score-based identification of $E[Y(d, M(d))]$ under $Y(d,m) \perp \{D, M\}|X = x$ (see Assumptions 1 and 2) has been established in the literature on doubly robust ATE estimation, see for instance Robins, Rotnitzky, and Zhao (1994) and Hahn (1998):

Lemma 3:

Under Assumptions 1-3, the potential outcome $E[Y(d, M(d))]$ is identified by the following efficient score function:

$$E[Y(d, M(d))] = E[\alpha_d] \text{ with } \alpha_d = \frac{I\{D = d\} \cdot [Y - \mu(d, X)]}{p_d(X)} + \mu(d, X) \tag{2.7}$$

where $\mu(D, X) = E(Y|D, M(D), X) = E(Y|D, X)$ is the conditional expectation of outcome Y given D and X.

For identifying the controlled direct effect, we now assume that M is discrete (while this need not be the case in the context of natural direct and indirect effects) such that for all m in the support of M, it must hold that $\Pr(M = m) > 0$. As Assumptions 1 and 2 imply $Y(d,m) \perp \{D, M\}|X = x$, doubly robust identification of the potential outcome $E[Y(d,m)]$, which is required for the controlled direct effect, follows from replacing $I\{D = d\}$ and $p_d(X)$ in (2.7) by $I\{D = d, M = m\} = I\{M = m\} \cdot I\{D = d\}$ and $\Pr(D = d, M = m|X) = f(m|d, X) \cdot p_d(X)$:

Lemma 4:

Under Assumptions 1-3, the potential outcome $E[Y(d,m)]$ is identified by the following efficient score function:

$$\begin{aligned} E[Y(d,m)] &= E[\psi_{dm}] \\ \text{with } \psi_{dm} &= \frac{I\{D = d\} \cdot I\{M = m\} \cdot [Y - \mu(d, m, X)]}{f(m|d, X) \cdot p_d(X)} + \mu(d, m, X). \end{aligned} \tag{2.8}$$

2.4 Estimation of the counterfactual with K-fold Cross-Fitting

We subsequently propose an estimation strategy for the counterfactual $E[Y(d, M(1 - d))]$ with $d \in \{0, 1\}$ based on the efficient score function by Tchetgen Tchetgen and Shpitser (2012) provided in (2.5) and show its $n^{-1/2}$-consistency under specific regularity conditions. To this end, let $\mathcal{W} = \{W_i | 1 \leq i \leq N\}$ with $W_i = (Y_i, M_i, D_i, X_i)$ for $i = 1, \ldots, n$ denote the set of observations in an i.i.d. sample of size n. η denotes the plug-in (or nuisance) parameters, i.e. the conditional mean outcome, mediator density and treatment probability. Their respective estimates are referred to by $\hat{\eta} = \{\hat{\mu}(D, M, X), \hat{f}(M|D, X), \hat{p_d}(X)\}$ and the true nuisance parameters by $\eta_0 = \{\mu_0(D, M, X), f_0(M|D, X), p_{d0}(X)\}$. Finally, $\psi_{d0} = E[Y(d, M(1-d))]$ denotes the true counterfactual.

We suggest estimating ψ_{d0} using the following algorithm that combines orthogonal score estimation with sample splitting and is $n^{-1/2}$-consistent under conditions outlined further below.

Algorithm 1 (Estimation of $E[Y(d, M(1-d))]$ based on equation (2.5)):

1. Split $\mathcal{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}_k^C$ the complement set of all observations not in $\mathcal{W}_k$.

2. For each k, use $\mathcal{W}_k^C$ to estimate the model parameters of $p_d(X)$, $f(M|D, X)$, and $\mu(D, M, X)$ in order to predict these models in $\mathcal{W}_k$, where the predictions are denoted by $\hat{p_d}^k(X)$, $\hat{f}^k(M|D, X)$, and $\hat{\mu}^k(D, M, X)$.

3. For each k, obtain an estimate of the efficient score function (see ψ_d in (2.5)) for each observation i in $\mathcal{W}_k$, denoted by $\hat{\psi}_{d,i}^k$:

$$\begin{aligned}\hat{\psi}_{d,i}^k \;\; = \;\; & \frac{I\{D_i = d\} \cdot \hat{f}^k(M_i|1-d, X_i)}{\hat{p}_d^k(X_i) \cdot \hat{f}^k(M_i|d, X_i)} \cdot [Y_i - \hat{\mu}^k(d, M_i, X_i)] \\ & + \frac{I\{D_i = 1-d\}}{1 - \hat{p}_d^k(X_i)} \cdot \left[\hat{\mu}^k(d, M_i, X_i) - \int_{m \in \mathcal{M}} \hat{\mu}^k(d, m, X_i) \cdot \hat{f}^k(m|1-d, X_i) dm\right] \\ & + \int_{m \in \mathcal{M}} \hat{\mu}^k(d, m, X_i) \cdot \hat{f}^k(m|1-d, X_i) dm. \qquad (2.9)\end{aligned}$$

4. Average the estimated scores $\hat{\psi}_{d,i}^k$ over all observations across all K subsamples to obtain an estimate of $\psi_{d0} = E[Y(d, M(1-d))]$ in the total sample, denoted by $\hat{\psi}_d = 1/n \sum_{k=1}^K \sum_{i=1}^{n_k} \hat{\psi}_{d,i}^k$.

Algorithm 1 can be adapted to estimate the counterfactuals required for the controlled direct effect, see (2.8). To this end, denote by $\psi_{dm0} = E[Y(d, m)]$ the true counterfactual of interest, which is estimated by replacing ψ_d and ψ_{d0} by ψ_{dm} and ψ_{dm0}, respectively, everywhere in Algorithm 1.

In order to achieve $n^{-1/2}$-consistency for counterfactual estimation, we make specific assumptions about the prediction qualities of the machine learners for our plug-in estimates of the nuisance parameters. Closely following Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), we to this end introduce some further notation. Let $(\delta_n)_{n=1}^{\infty}$ and $(\Delta_n)_{n=1}^{\infty}$ denote sequences of positive constants with $\lim_{n\to\infty}\delta_n = 0$ and $\lim_{n\to\infty}\Delta_n = 0$. Also, let $c, \epsilon, C, \underline{f}, \overline{f}$ and q be positive constants such that $q > 2$, and let $K \geq 2$ be a fixed integer. Furthermore, for any random vector $Z = (Z_1, ..., Z_l)$, let $\|Z\|_q = \max_{1\leq j\leq l}\|Z_l\|_q$, where $\|Z_l\|_q = (E[|Z_l|^q])^{1/q}$. For the sake of easing notation, we assume that n/K is an integer. For brevity, we omit the dependence of probability $\Pr_P(\cdot)$, expectation $E_P(\cdot)$, and norm $\|\cdot\|_{P,q}$ on the probability measure P.

Assumption 4 (regularity conditions and quality of plug-in parameter estimates): For all probability laws $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws, the following conditions hold for the random vector (Y, D, M, X) for $d \in \{0, 1\}$:

(a) $\|Y\|_q \leq C$ and $\left\|E[Y^2|d, M, X]\right\|_\infty \leq C^2$,

(b) $\Pr(\epsilon \leq p_{d0}(X) \leq 1 - \epsilon) = 1$,

(c) $\Pr(\underline{f} \leq f(M|D, X) \leq \overline{f}) = 1$,

(d) $\|Y - \mu_0(d, M, X)\|_2 = E\Big[(Y - \mu_0(d, M, X)))^2\Big]^{1/2} \geq c$

(e) Given a random subset $\mathcal{W}_k$ of size n/K, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0(\mathcal{W}_k^C)$ satisfies the following conditions. With P-probability no less than $1 - \Delta_n$:

$$\begin{aligned}
\|\hat{\eta}_0 - \eta_0\|_q &\leq C,\\
\|\hat{\eta}_0 - \eta_0\|_2 &\leq \delta_n,\\
\|\hat{p}_{d0}(X) - 1/2\|_\infty &\leq 1/2 - \epsilon,\\
\left\|\hat{f}_0(M|D, X) - (\underline{f} + \overline{f})/2\right\|_\infty &\leq (\overline{f} - \underline{f})/2,\\
\|\hat{\mu}_0(D, M, X) - \mu_0(D, M, X)\|_2 \times \|\hat{p}_{d0}(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2},\\
\|\hat{\mu}_0(D, M, X) - \mu_0(D, M, X)\|_2 \times \left\|\hat{f}_0(M|1 - D, X) - f_0(M|1 - D, X)\right\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}$$

For demonstrating $n^{-1/2}$-consistency of the proposed estimation strategy for the counterfactual, we heavily draw from Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) by showing that our estimation strategy satisfies the requirements for their double machine learning framework.

Lemma 5 (Neyman Orthogonality and Linearity):

The following conditions are satisfied: (a) the moment condition $E\Big[\psi_d(W,\eta_0,\psi_{d0})\Big]=0$ holds, (b) the score $\psi_d(W,\eta_0,\psi_{d0})$ is linear in ψ_{d0}, (c) the second Gateaux derivative of $\eta \mapsto E\Big[\psi_d(W,\hat{\eta},\psi_{d0})\Big]$ is continuous, (d) the score function is Neyman orthogonal and (e) singular values of $E[\psi_d^a(W;\eta_0)]$ are bounded.

The proof is provided in Appendix 2.B.1.

Then, as e.g. $\psi_d(W,\eta,\psi_{d0})$ is smooth in (η,ψ_{d0}), the plug-in estimators must converge with rate $n^{-1/4}$ in order to achieve $n^{-1/2}$-convergence for the estimation of $\hat{\psi}_d$. This convergence rate of $n^{-1/4}$ is achievable for many commonly used machine learners such as lasso, random forest, boosting and neural nets. The rates for L_2-boosting were, for instance, derived in Luo and Spindler (2016).

Theorem 1:

Under Assumptions 1-4, it holds for estimating $E[Y(d,M(1-d))]$, $E[Y(d,m)]$ based on Algorithm 1:

$\sqrt{n}\Big(\hat{\psi}_d-\psi_{d0}\Big)\rightarrow N(0,\sigma^2_{\psi_d})$, where $\sigma^2_{\psi_d}=E[(\psi_d-\psi_{d0})^2]$.

$\sqrt{n}\Big(\hat{\psi}_{dm}-\psi_{dm0}\Big)\rightarrow N(0,\sigma^2_{\psi_{dm}})$, where $\sigma^2_{\psi_d}=E[(\psi_d-\psi_{dm0})^2]$.

The proof is provided in Appendix 2.B.1.

Analogous results follow for the estimation of $\Lambda=E[Y(d,M(d))]$ when replacing $\hat{\psi}_d$ in the algorithm above by an estimate of score function α_d from (2.7),

$$\hat{\alpha}_d=\frac{I\{D=d\}\cdot(Y_i-\hat{\mu}^k(d,X_i))}{\hat{p_d}^k(X_i)}+\hat{\mu}^k(d,X_i), \tag{2.10}$$

where $\hat{\mu}^k(d,x)$ is an estimate of $\mu(d,x)$. This approach has been discussed in literature on ATE estimation based on double machine learning, see for instance Belloni, Chernozhukov, Fernández-Val, and Hansen (2017) and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018). Denoting by $\hat{\Lambda}$ the estimate of Λ, it follows under Assumptions 1-4 that $\sqrt{n}\Big(\hat{\Lambda}_d-\Lambda_d\Big)\rightarrow N(0,\sigma^2_{\alpha_d})$, where $\sigma^2_{\alpha_d}=E[(\alpha_d-\Lambda_d)^2]$. Therefore, $n^{-1/2}$-consistent estimates of the total as well as the direct and indirect effects are obtained as difference of the estimated potential outcomes, which we denote by $\hat{\Delta}$, $\hat{\theta}(d)$, and $\hat{\delta}(d)$. That is, $\hat{\Delta}=\hat{\Lambda}_1-\hat{\Lambda}_0$, $\hat{\theta}(1)=\hat{\Lambda}_1-\hat{\psi}_0$, $\hat{\theta}(0)=\hat{\psi}_1-\hat{\Lambda}_0$, $\hat{\delta}(1)=\hat{\Lambda}_1-\hat{\psi}_1$, and $\hat{\delta}(0)=\hat{\psi}_0-\hat{\Lambda}_0$.

Naturally, the asymptotic variance of any effect is obtained based on the variance of the difference in the score functions of the potential outcomes required for the respective effect. For instance, the asymptotic variance of $\hat{\theta}(1)$ is given by $Var(\hat{\theta}(1)) = Var(\alpha_1 - \psi_0)/n = (\sigma^2_{\alpha_1} + \sigma^2_{\psi_0} - 2Cov(\alpha_1, \psi_0))/n$.

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) show that under Assumptions 1-4, $\hat{\sigma}^2_{\psi_d}$ can be estimated as:

$$\hat{\sigma}^2_{\psi_d} = \frac{1}{K}\sum_{k=1}^{K}\left[1/n_k \sum_{i=1}^{n_k} \psi_d(W_i, \hat{\eta}_0^k, \hat{\psi}_d)^2\right] \tag{2.11}$$

The asymptotic variance of α_d can be estimated accordingly, with ψ_d and $\hat{\psi}_{d0}$ substituted by α_d and $\hat{\Lambda}_{d0}$.

We subsequently discuss estimation based on the score function ψ_d^* in expression (2.6). We note that in this case, we have to estimate the nested nuisance parameter $E\Big[\mu(d, M, X)\Big|D = 1 - d, X\Big]$, which we henceforth denote by $\omega(1 - d, X)$. To avoid overfitting, the models for $\mu(d, M, X)$ and $\omega(1 - d, X)$ are estimated in different subsamples. The plug-in estimates for the conditional mean outcome, the nested conditional mean outcome, mediator density and treatment probability are referred to by $\hat{\eta}^* = \{\hat{\mu}(D, M, X), \hat{\omega}(D, X), \hat{p}_d(M, X), \hat{p}_d(X)\}$ and the true nuisance parameters by $\eta_0^* = \{\mu_0(D, M, X), \omega_0(D, X), p_{d0}(M, X), p_{d0}(X)\}$.

Algorithm 2 (Estimation of $E[Y(d, M(1 - d))]$ based on equation (2.6)):

1. Split $\mathcal{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathcal{W}_k$ the set of observations in the sample and $\mathcal{W}_k^C$ the complement set of all observations not in $\mathcal{W}_k$.

2. For each k, use $\mathcal{W}_k^C$ to estimate the model parameters of $p_d(X)$ and $p_d(M, X)$. Split $\mathcal{W}_k^C$ into 2 nonoverlapping subsamples, estimate the model parameters of the conditional mean $\mu(d, M, X)$ in one subsample and use it for estimating the nested conditional mean $\omega(1 - d, X) = E\Big[\mu(d, M, X)\Big|D = 1 - d, X\Big]$ in the other subsample. Predict the nuisance parameters in $\mathcal{W}_k$, where the predictions are denoted by $\hat{p_d}^k(X)$, $\hat{p}_d^k(M, X)$, $\hat{\mu}^k(D, M, X)$ and $\hat{\omega}(D, X)^k$.

3. For each k, obtain an estimate of the efficient score function (see ψ_d^* in (2.6)) for each

observation i in $\mathcal{W}_k$, denoted by $\hat{\psi}^{*k}_{d,i}$:

$$\begin{aligned}\hat{\psi}^{*k}_{d,i} \;=\;& \frac{I\{D_i = d\}\big(1-\hat{p}^k_d(M_i,X_i)\big)}{\hat{p}^k_d(M_i,X_i)\,\big(1-\hat{p}^k_d(X_i)\big)}\cdot[Y-\hat{\mu}^k(d,M_i,X_i)]\\ &+\frac{I\{D_i=1-d\}}{1-\hat{p}^k_d(X_i)}\cdot\Big[\hat{\mu}^k(d,M_i,X_i)-\hat{\omega}(1-d,X_i)^k\Big]+\hat{\omega}(1-d,X_i)^k. \quad (2.12)\end{aligned}$$

4. Average the estimated scores $\hat{\psi}^{*k}_{d,i}$ over all observations across all K subsamples to obtain an estimate of $\psi_{d0} = E[Y(d, M(1-d))]$ in the total sample, denoted by $\hat{\psi}^*_d = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\hat{\psi}^{*k}_{d,i}$.

Also this approach can be shown to be $n^{-1/2}$-consistent under specific regularity conditions outlined below.

Assumption 5 (regularity conditions and quality of plug-in parameter estimates)

For all probability laws $P \in \mathcal{P}$ the following conditions hold for the random vector (Y, D, M, X) for all $d \in \{0, 1\}$:

(a) $\|Y\|_q \le C$ and $\big\|E[Y^2|d,M,X]\big\|_\infty \le C^2$,

(b) $\Pr(\epsilon \le p_{d0}(X) \le 1-\epsilon) = 1$,

(c) $\Pr(\epsilon \le p_{d0}(M,X) \le 1-\epsilon) = 1$,

(d) $\|Y-\mu_0(d,M,X)\|_2 = E\Big[(Y-\mu_0(d,M,X)))^2\Big]^{1/2} \ge c$

(e) Given a random subset $\mathcal{W}_k$ of size n/K, the nuisance parameter estimator $\hat{\eta}^*_0 = \hat{\eta}^*_0(\mathcal{W}^C_k)$ satisfies the following conditions. With P-probability no less than $1-\Delta_n$:

$$\begin{aligned}\|\hat{\eta}^*_0-\eta^*_0\|_q &\le C,\\ \|\hat{\eta}^*_0-\eta^*_0\|_2 &\le \delta_n,\\ \|\hat{p}_{d0}(X)-1/2\|_\infty &\le 1/2-\epsilon,\\ \|\hat{p}_{d0}(M,X)-1/2\|_\infty &\le 1/2-\epsilon,\\ \|\hat{\mu}_0(D,M,X)-\mu_0(D,M,X)\|_2\times\|\hat{p}_{d0}(X)-p_{d0}(X)\|_2 &\le \delta_n n^{-1/2},\\ \|\hat{\mu}_0(D,M,X)-\mu_0(D,M,X)\|_2\times\|\hat{p}_{d0}(M,X)-p_{d0}(M,X)\|_2 &\le \delta_n n^{-1/2},\\ \|\hat{\omega}_0(D,X)-\omega_0(D,X)\|_2\times\|\hat{p}_{d0}(X)-p_{d0}(X)\|_2 &\le \delta_n n^{-1/2}.\end{aligned}$$

Theorem 2:

Under Assumptions 1-3 and 5, it holds for estimating $E[Y(d, M(1-d))]$ based on Algorithm 2: $\sqrt{n}\left(\hat{\psi}_d^* - \psi_{d0}^*\right) \rightarrow N(0, \sigma^2_{\psi_d^*})$, where $\sigma^2_{\psi_d^*} = E[(\psi_d^* - \psi_{d0}^*)^2]$.
The proof is provided in Appendix 2.B.2.

2.5 Simulation study

This section provides a simulation study to investigate the finite sample behaviour of the proposed methods based on the following data generating process:

$$\begin{aligned} Y &= 0.5D + 0.5M + 0.5DM + X'\beta + U, \\ M &= I\{0.5D + X'\beta + V > 0\}, \quad D = I\{X'\beta + W > 0\}, \\ X &\sim N(0, \Sigma), \quad U, V, W \sim N(0,1) \text{ independently of each other and } X. \end{aligned}$$

Outcome Y is a function of the observed variables D, M, X, including an interaction between the mediator and the treatment, and an unobserved term U. The binary mediator M is a function of D, X and the unobservable V, while the binary treatment D is determined by X and the unobservable W. X is a vector of covariates of dimension p, which is drawn from a multivariate normal distribution with zero mean and covariance matrix Σ. The latter is defined based on setting the covariance of the ith and jth covariate in X to $\Sigma_{ij} = 0.5^{|i-j|}$.[3] Coefficients β gauge the impact of X on Y, M, and D, respectively, and thus, the strength of confounding. U, V, W are random and standard normally distributed scalar unobservables. We consider two sample sizes of $n = 1000, 4000$ and run 1000 simulations per data generating process.

We investigate the performance of effect estimation based on (i) Theorem 1 using the identification result in expression (2.5) derived by Tchetgen Tchetgen and Shpitser (2012) as well as (ii) Theorem 2 using the modified score function in expression (2.6) which avoids conditional mediator densities. The nuisance parameters are estimated by post-lasso regression based on the 'causalweight' package by Bodory and Huber (2018) for the statistical software 'R' (R Core Team (2020)), in which our estimation procedure is made available, using logit specifications for $p_d(X)$, $p_d(M, X)$, and $f(M|D, X)$ and linear specifications for $\mu(D, M, X)$ and $\omega(1-d, X)$.

[3] The results presented below are hardly affected when setting Σ to the identity matrix (zero correlation across X).

The estimation of direct and indirect effects is based on 4-fold cross-fitting. For all methods investigated, we drop observations whose (products of) estimated conditional probabilities in the denominator of any potential outcome expression are close to zero, namely smaller than a trimming threshold of 0.05 (or 5%). Furthermore, we normalize the weights related to the inverse propensity scores in our estimators such that they sum up to one within treatment groups, as for instance advocated in Busso, DiNardo, and McCrary (2009).

In our first simulation design, we set $p = 200$ and the ith element in the coefficient vector β to $0.3/i^2$ for $i = 1, ..., p$, meaning a quadratic decay of covariate importance in terms of confounding. This specification implies that the R^2 of X when predicting Y amounts to 0.22 in large samples, while the Nagelkerke (1991) pseudo-R^2 of X when predicting D and M by probit models amounts to 0.10 and 0.13, respectively. The left panel of Table 2.1 reports the results for either sample size. For $n = 1000$, double machine learning based on Theorem 2 on average exhibits a slightly lower absolute bias ('abias') and standard deviation ('sd') than estimation based on Theorem 1. The behaviour of both approaches improves when increasing sample size to $n = 4000$, as the absolute bias is very close to zero for any effect estimate and standard deviation is roughly cut by half. Under the larger sample size, differences in terms of root mean squared error ('rmse') between estimation based on Theorems 1 and 2 are very close to zero. By and large, the results suggest that the estimators converge to the true effects at rate $n^{-1/2}$.

In our second simulation, confounding is increased by setting β to $0.5/i^2$ for $i = 1, ..., p$. This specification implies that the R^2 of X when predicting Y amounts to 0.42, while the Nagelkerke (1991) pseudo-R^2 of X when predicting D and M amounts to 0.23 and 0.28, respectively. The results are displayed in the right panel of Table 2.1. Again, estimation based on Theorem 2 slightly dominates in terms of having a smaller absolute bias and standard deviation, in particular for $n = 1000$. However, in other settings, the two methods might compare differently in terms of finite sample performance. Both methods based on Theorems 1 and 2, respectively, appear to converge to the true effects at rate $n^{-1/2}$, and differences in terms of root mean squared errors are minor for $n = 4000$.

Appendix 2.A reports the simulation results (namely the absolute bias, standard deviation, and root mean squared error) for the standard errors obtained by an asymptotic approximation based on the estimated variance of the score functions. The results suggest that the asymptotic standard errors decently estimate the actual standard deviation of the point estimators.

	Coefficients given by $0.3/i^2$ for $i = 1, ..., p$							Coefficients given by $0.5/i^2$ for $i = 1, ..., p$						
	abias	sd	rmse	abias	sd	rmse	true	abias	sd	rmse	abias	sd	rmse	true
		n=1000			n=4000				n=1000			n=4000		
	Double machine learning based on Theorem 1													
$\hat{\Delta}$	0.01	0.08	0.08	0.00	0.04	0.04	1.02	0.00	0.10	0.10	0.01	0.04	0.05	1.00
$\hat{\theta}(1)$	0.01	0.08	0.08	0.00	0.04	0.04	0.84	0.01	0.09	0.09	0.01	0.04	0.04	0.83
$\hat{\theta}(0)$	0.00	0.08	0.08	0.00	0.04	0.04	0.75	0.00	0.10	0.10	0.01	0.04	0.04	0.75
$\hat{\delta}(1)$	0.00	0.06	0.06	0.00	0.03	0.03	0.27	0.00	0.07	0.07	0.00	0.03	0.03	0.25
$\hat{\delta}(0)$	0.01	0.05	0.05	0.00	0.02	0.02	0.18	0.01	0.05	0.05	0.00	0.02	0.02	0.17
trimmed		17.24			19.19				80.25			237.50		
	Double machine learning based on Theorem 2													
$\hat{\Delta}$	0.01	0.08	0.08	0.00	0.04	0.04	1.02	0.01	0.09	0.09	0.01	0.04	0.04	1.00
$\hat{\theta}(1)$	0.00	0.07	0.07	0.00	0.04	0.04	0.84	0.01	0.08	0.08	0.00	0.04	0.04	0.83
$\hat{\theta}(0)$	0.00	0.08	0.08	0.00	0.04	0.04	0.75	0.00	0.08	0.08	0.00	0.04	0.04	0.75
$\hat{\delta}(1)$	0.00	0.06	0.06	0.00	0.03	0.03	0.27	0.00	0.06	0.06	0.00	0.03	0.03	0.25
$\hat{\delta}(0)$	0.00	0.04	0.04	0.00	0.02	0.02	0.18	0.00	0.04	0.04	0.00	0.02	0.02	0.17
trimmed		1.20			0.11				16.76			25.45		

Table 2.1: Simulation results for effect estimates ($p = 200$). Note: 'abias', 'sd', and 'rmse' denote the absolute bias, standard deviation and root mean squared error of the respective effect estimate. 'true' provides the true effect. 'trimmed' is the average number of trimmed observations per simulation. The propensity score-based trimming threshold is set to 0.05.

2.6 Application

In this section, we apply our method to data from the National Longitudinal Survey of Youth 1997 (NLSY97), a survey conducted by the Bureau of Labor Statistics at the U.S. Department of Labor (2019) following a U.S. nationally representative sample of 8,984 individuals born in the years 1980-84. Since 1997, the participants have been interviewed on a wide range of demographic, socioeconomic, and health-related topics in a one- to two-year circle. We investigate the causal effect of health insurance coverage (D) on general health (Y) and decompose it into an indirect pathway via the incidence of a regular medical checkup (M) and a direct effect entailing any other causal mechanisms. Whether or not an individual undergoes routine checkups appears to be an interesting mediator, as it is likely to be affected by health insurance coverage and may itself have an impact on the individual's health, because checkups can help identifying medical conditions before they get serious to prevent them from affecting a person's general health state.

The effect of health insurance coverage on self-reported health has been investigated in different countries with no compulsory medical insurance and no publicly provided universal health coverage, see for example Simon, Soni, and Cawley (2017), Sommers, Maylone, Blendon, Orav, and Epstein (2017), Baicker, Taubman, Allen, Bernstein, Gruber, Newhouse, Schneider, Wright, Zaslavsky, and Finkelstein (2013), Yörük (2016) and Cardella and Depew (2014) for the U.S. and King, Gakidou, Imai, Lakin, Moore, Nall, Ravishankar, Vargas, Tellez-Rojo, Avila, et al. (2009) for Mexico). Most of these studies find a significant positive effect of insurance coverage on self-reported health. The impact of insurance coverage on the utilization of preventive care measures, particularly routine checkups like cancer, diabetes and cardiovascular screenings, is also extensively covered in public health literature. Most studies find that health insurance coverage increases the odds of attending routine checkups. While some contributions include selected demographic, socioeconomic and health-related control variables to account for the endogeneity of health insurance status (see e.g. Faulkner and Schauffler (1997), Press (2014), Burstin, Swartz, O'Neil, Orav, and Brennan (1998), Fowler-Brown, Corbie-Smith, Garrett, and Lurie (2007)), others exploit natural experiments: Simon, Soni, and Cawley (2017) estimate a difference-in-differences model comparing states which did and did not expand Medicaid to low-income adults in 2005, while Baicker, Taubman, Allen, Bernstein, Gruber, Newhouse, Schneider, Wright, Zaslavsky, and Finkelstein (2013) exploit that the state of Oregon expanded Medicaid based on lottery drawings from a waiting list. The results of both studies suggest that the Medicaid expansions increased use of certain forms of preventive care. In a study on Mexican adults, Pagán, Puig, and Soldo (2007) use self-employment and commission pay as instruments for insurance coverage and also find a more frequent use of some types of preventive care by individuals with health insurance coverage.

While the bulk of studies investigating checkups focus on one particular type of screening (rather than general health checkups), see Maciosek, Coffield, Flottemesch, Edwards, and Solberg (2010) for a literature review, several experimental contributions also assess general health checkups. For instance, Rasmussen, Thomsen, Kilsmark, Hvenegaard, Engberg, Lauritzen, and Sogaard (2007) conduct an experiment with individuals aged 30 to 49 in Denmark by randomly offering a set of health screenings, including advice on healthy living and find a significant positive effect on life expectation. In a study on Japan's elderly population, Nakanishi, Tatara, and Fujiwara (1996) find a significantly negative correlation between the rate of attendance at health check-ups and hospital admission rates. Despite the effects of health insurance coverage and routine checkups being extensively covered in the public health literature, the indirect effect of insurance on general health operating via routine checkups as mediator has to the best of our knowledge not yet been investigated. A further distinction to most previous studies is that we consider comparably young individuals with an average age below 30. For this population, the relative importance of different health screenings might differ from that for other age groups. We also point out that our application focuses on short-term health effects.

We consider a binary indicator for health insurance coverage, equal to one if an individual reports to have any kind of health insurance when interviewed in 2006 and zero otherwise. The outcome, self-reported general health, is obtained from the 2008 interview and measured with an ordinal variable, taking on the values 'excellent', 'very good', 'good', 'fair' and 'poor'. In the 2007 interview, participants were asked whether they have gone for routine checkups since the 2006 interview. This information serves as binary mediator, measured post-treatment but pre-outcome.

To ensure that the control variables (X) are not influenced by the treatment, they come from the pre-treatment 2005 and earlier interview rounds. They cover demographic characteristics, family background and quality of the home environment during youth, education and training, labour market status, income and work experience, marital status and fertility, household characteristics, received monetary transfers, attitudes and expectations, state of physical and mental health as well as health-related behaviour regarding e.g. nutrition and physical activity. For some variables, we only consider measurements from 2005 or from the initial interview round covering demographics and family related topics. For other variables we include measurements from both the indiviuals' youth and 2005 in order to capture their social, emotional and physical development. Treatment and mediator state in the pre-treatment period (2005) are also considered as potential control variables. Item non-response in control variables is dealt with by including missing dummies for each control variable and setting the respective missing values to zero. In total, we end up with a set of 755 control variables, 593 of which are dummy variables (incl. 251 dummies for missing values).

	overall	$D = 1$	$D = 0$	diff	p-val	$M = 1$	$M = 0$	diff	p-val
n	7,061	2,335	4,726			3,612	3,449		
Female	0.5	0.54	0.41	0.13	0	0.66	0.35	0.31	0
Age	22.5	22.54	22.44	0.1	0	22.54	22.46	0.08	0.02
Ethnicity									
Black	0.27	0.25	0.3	-0.04	0	0.32	0.22	0.1	0
Hispanic	0.21	0.19	0.25	-0.06	0	0.21	0.22	-0.01	0.58
Mixed	0.01	0.01	0.01	0	0.35	0.01	0.01	0	0.3
White or Other	0.51	0.55	0.44	0.11	0	0.46	0.55	-0.1	0
Relationship/Marriage									
Not Cohabiting	0.62	0.61	0.65	-0.03	0	0.61	0.64	-0.03	0.01
Cohabiting	0.17	0.16	0.18	-0.02	0.01	0.16	0.17	0	0.61
Married	0.18	0.21	0.14	0.07	0	0.2	0.17	0.03	0
Separated/ Widowed	0.02	0.02	0.03	-0.01	0.02	0.02	0.02	0	0.55
Missing	0	0	0	0	0.42	0	0	0	0.92
Urban	0.72	0.75	0.67	0.08	0	0.75	0.7	0.05	0
Missing	0.11	0.08	0.16	-0.08	0	0.09	0.14	-0.05	0
HH Income [4]	41,851	47,908	31,433	16475	0	43,338	40,460	2878	0.03
Missing	0.24	0.2	0.31	-0.11	0	0.21	0.26	-0.05	0
HH Size	2.99	3.05	2.89	0.16	0	3.1	2.89	0.21	0
Missing	0.09	0.06	0.14	-0.09	0	0.06	0.11	-0.05	0
HH Members under 18	0.67	0.65	0.69	-0.04	0.13	0.76	0.58	0.18	0
Missing	0.09	0.06	0.14	-0.09	0	0.07	0.11	-0.05	0
Biological Children	0.48	0.47	0.5	-0.02	0.24	0.55	0.42	0.13	0
Highest Grade	11.78	12.61	10.36	2.25	0	12.26	11.33	0.93	0
Missing	0.09	0.06	0.15	-0.09	0	0.07	0.11	-0.05	0
Employment									
Employed	0.71	0.73	0.68	0.05	0	0.7	0.72	-0.02	0.11
Unemployed	0.05	0.04	0.07	-0.03	0	0.05	0.06	-0.01	0.24
Out of Labour Force	0.2	0.19	0.23	-0.04	0	0.21	0.2	0	0.7
Military	0.03	0.04	0.01	0.02	0	0.04	0.02	0.02	0
Missing	0	0	0.01	0	0.01	0	0	0	0.59
Working Hours (per week)	24.04	25.35	21.79	3.57	0	24.11	23.98	0.13	0.78
Missing	0.09	0.06	0.14	-0.09	0	0.06	0.11	-0.05	0
Weight (pounds)	152	156	145	11	0	152	152	1	0.72
Missing	0.11	0.08	0.17	-0.09	0	0.09	0.14	-0.05	0
Height (feet)	4.97	5.16	4.64	0.52	0	5.03	4.91	0.12	0
Missing	0.12	0.08	0.18	-0.1	0	0.09	0.14	-0.05	0
Days 5+ drinks (per month)	1.57	1.55	1.62	-0.07	0.44	1.22	1.9	-0.68	0
Missing	0.11	0.08	0.17	-0.09	0	0.09	0.14	-0.05	0
Days of Exercise (per week)	2.37	2.42	2.3	0.11	0.05	2.33	2.41	-0.08	0.15
Missing	0.06	0.05	0.09	-0.05	0	0.05	0.08	-0.03	0
Depressed/ Down									
Never	0.3	0.31	0.28	0.03	0	0.29	0.31	-0.02	0.05
Sometimes	0.49	0.52	0.45	0.07	0	0.51	0.47	0.04	0
Mostly	0.09	0.09	0.09	0	0.68	0.1	0.08	0.02	0
Always	0.02	0.02	0.02	-0.01	0.03	0.02	0.02	0	0.41
Missing	0.1	0.07	0.16	-0.09	0	0.08	0.12	-0.04	0

Table 2.2: Descriptive Statistics. Note 'overall', '$D = 1$', '$D = 0$', '$M = 1$', '$M = 0$' report the mean of the respective variable in the total sample, among treated, among non-treated, among mediated, and among non-mediated, respectively. 'diff' and 'p-val' provide the mean difference (across treatment or mediator states) and the p-value of a two-sample t-test, respectively.

[4] The HH income variable is the sum of several variables measuring HH income components (different sources & receivers).

After excluding 1,498 observations with either mediator or treatment status missing, we remain with 7,486 observations. Table 2.2 presents some descriptive statistics for a selection of control variables. It shows that the group of individuals with and without health insurance coverage differ substantially. There are significant differences with respect to most of the control variables listed in the table. Females are significantly more likely to have health insurance coverage. Education and household income also show a significant positive correlation with health insurance coverage while the number of household members for example is negatively correlated with insurance coverage. Regarding the mediator, we find a similar pattern as for the treatment. With respect to many of the considered variables, the group of individuals who went for medical checkup differs substantially from those who did not. Further, we see that the correlation between many control variables and the treatment appear to have the same sign as that with the mediator.

In order to assess the direct and indirect effect of health insurance coverage on general health, we consider estimation based on Theorem 1 and expression (2.5) derived by Tchetgen Tchetgen and Shpitser (2012) as well as (ii) Theorem 2 and expression (2.6). We estimate the nuisance parameters and treatment effects in the same way as outlined in Section 2.5 (i.e. post-lasso regression for modelling the nuisance parameters and 3-fold cross fitting for effect estimation) after augmenting the set of covariates with 380 selected interaction and higher order terms of covariates measuring demographic characteristics, health status, and health-related behaviour. The trimming threshold for discarding observations with too extreme propensity scores is set to 0.02 (2%), such that 777 and 54 observations are dropped when basing estimation on Theorems 1 and 2, respectively. As for the simulations, the propensity score-based weights in our estimators are normalized such that they sum up to one within treatment groups.

Table 2.3 provides the estimated effects along with the standard error ('se') and p-value ('p-val') and also provides the estimated mean potential outcome under non-treatment for comparison ('$\hat{E}[Y(0, M(0))]$'). The ATEs of health insurance coverage on general health in the year 2008 (columns 2 and 8), estimated based on Theorems 1 or 2, are statistically significant at the 10% and 5% levels, respectively. As the outcome is measured on an ordinal scale ranging from 'excellent' to 'poor', the negative ATEs suggest a short-term health improving effect of health coverage. The direct effects under treatment (columns 3 and 9) and under non-treatment

These variables are capped but only a total of 11 observations are in critical cap categories (columns 4 and 10) mostly have a similar magnitude as the ATEs, even though they are not statistically significant in 3 out of 4 cases. The indirect effects under treatment (columns 5 and 11) and non-treatment (columns 6 and 12) are generally close to zero and not statistically significant in 3 out of 4 cases either. We therefore conclude that in the short run, health insurance coverage does not seem to importantly affect general health of young adults in the U.S. through routine checkups.

	Estimations based on Theorem 1						Estimations based on Theorem 2					
	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{E}[Y(0,M(0))]$	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$	$\hat{E}[Y(0,M(0))]$
effect	-0.05	-0.04	-0.04	-0.01	-0.01	2.27	-0.05	-0.03	-0.05	-0.00	-0.02	2.28
se	0.03	0.03	0.03	0.01	0.01	0.03	0.03	0.03	0.03	0.01	0.01	0.02
p-val	0.10	0.23	0.23	0.49	0.17	0.00	0.04	0.22	0.05	0.88	0.04	0.00

Table 2.3: Total, direct, and indirect effects on general health in 2008. Note: 'effect', 'se', and 'p-val' report the respective effect estimate, standard error and p-value. Lasso regression is used for the estimation of nuisance parameters. The propensity score-based trimming threshold is set to 0.02.

2.7 Conclusion

In this paper, we combined causal mediation analysis with double machine learning under selection-on-observables assumptions which avoids ad hoc pre-selection of control variables. Thus, this approach appears particularly fruitful in high-dimensional data with many potential control variables. We proposed estimators for natural direct and indirect effects as well as the controlled direct effect exploiting efficient score functions, sample splitting, and machine learning-based plug-in estimates for conditional outcome means, mediator densities, and/or treatment propensity scores. We demonstrated the $n^{-1/2}$-consistency and asymptotic normality of the effect estimators under specific regularity conditions. Furthermore, we investigated the finite sample behaviour of the proposed estimators in a simulation study and found the performance to be decent in samples with several thousand observations. Finally, we applied our method to data from the U.S. National Longitudinal Survey of Youth 1997 and found a moderate short-term effect of health insurance coverage on general health, which was, however, not importantly mediated by routine checkups. The estimators considered in the simulation study and the application are available in the 'causalweight' package for the statistical software 'R'.

Appendix

2.A Simulation results for standard errors

	Coefficients given by $0.3/i^2$ for $i = 1, ..., p$								Coefficients given by $0.5/i^2$ for $i = 1, ..., p$							
	abias	sd	rmse	true	abias	sd	rmse	true	abias	sd	rmse	true	abias	sd	rmse	true
	n=1000				n=4000				n=1000				n=4000			
Double machine learning based on Theorem 1																
$se(\hat{\Delta})$	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.04	0.01	0.01	0.01	0.10	0.00	0.00	0.00	0.04
$se(\hat{\theta}(1))$	0.01	0.01	0.01	0.08	0.00	0.00	0.00	0.04	0.01	0.01	0.01	0.09	0.00	0.00	0.00	0.04
$se(\hat{\theta}(0))$	0.00	0.01	0.01	0.08	0.00	0.00	0.00	0.04	0.01	0.02	0.02	0.10	0.00	0.00	0.00	0.04
$se(\hat{\delta}(1))$	0.00	0.01	0.01	0.06	0.00	0.00	0.00	0.03	0.01	0.01	0.02	0.07	0.00	0.00	0.00	0.03
$se(\hat{\delta}(0))$	0.00	0.01	0.01	0.05	0.00	0.00	0.00	0.02	0.01	0.01	0.01	0.05	0.00	0.00	0.00	0.02
Double machine learning based on Theorem 2																
$se(\hat{\Delta})$	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.09	0.00	0.00	0.00	0.04
$se(\hat{\theta}(1))$	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.08	0.00	0.00	0.00	0.04
$se(\hat{\theta}(0))$	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.04	0.00	0.01	0.01	0.08	0.00	0.00	0.00	0.04
$se(\hat{\delta}(1))$	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.03	0.00	0.01	0.01	0.06	0.00	0.00	0.00	0.03
$se(\hat{\delta}(0))$	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.02

Table 2.A.1: Simulation results for standard errors ($p = 200$). Note: 'abias', 'sd', and 'rmse' denote the absolute bias, standard deviation and root mean squared error of the respective standard error ('se'). 'true' provides the true standard deviation.

2.B Proofs

For the proofs of Theorems 1 and 2, it suffices verifying the conditions of Assumptions 3.1 and 3.2 underlying Theorem 3.1 and 3.2 in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

2.B.1 Proof of Theorem 1

We first show that Assumptions 3.1 and 3.2 in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) are satisfied for $\psi_{d0} = E[Y(d, M(1-d))]$ based on (2.5).

Then, we show that Assumption 3.1 holds for $\psi_{dm0} = E[Y(d,m)]$ based on (2.8), but omit the proof of the validity of Assumption 3.2, as it follows in a very similar manner as for ψ_{d0}. All bounds hold uniformly over all probability laws $P \in \mathcal{P}$, where $\mathcal{P}$ is the set of all possible probability laws, and we omit P for brevity.

Let $\eta = (\mu(D,M,X), f(M|D,X), p_d(X))$ be the vector of nuisance parameters. Also, let $\mathcal{T}_n$ be the set of all $\eta = (\mu, f, p_d)$ in a neighbourhood of η_0 that is shrinking with increasing n, consisting of P-square integrable functions μ, f, and p_d such that

$$\begin{aligned}
\|\eta - \eta_0\|_q &\leq C, \qquad (2.13)\\
\|\eta - \eta_0\|_2 &\leq \delta_n,\\
\|p_d(X) - 1/2\|_\infty &\leq 1/2 - \epsilon,\\
\left\|f(M|D,X) - (\underline{f} + \overline{f})/2\right\|_\infty &\leq (\overline{f} - \underline{f})/2,\\
\|\mu(D,M,X) - \mu_0(D,M,X)\|_2 \times \|p_d(X) - p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2},\\
\|\mu(D,M,X) - \mu_0(D,M,X)\|_2 \times \|f(M|1-D,X) - f_0(M|1-D,X)\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}$$

We furthermore replace the sequence $(\delta_n)_{n\geq 1}$ by $(\delta_n')_{n\geq 1}$, where $\delta_n' = C_\epsilon \max(\delta_n, n^{-1/2})$, where C_ϵ is sufficiently large constant that only depends on C and ϵ. Let $R \equiv \overline{f}/\underline{f}$ stands for the maximal ratio of densities $f(m|d,X)$.

Counterfactual $E[Y(d,M(1-d))]$

The score function for the counterfactual $\psi_{d0} = E[Y(d,M(1-d))]$ proposed by Tchetgen Tchetgen and Shpitser (2012) is given by the following expression, with $W = (Y,M,D,X)$:

$$\begin{aligned}
\psi_d(W,\eta,\psi_{d0}) \;=\; & \frac{I\{D=d\}\cdot f(M|1-d,X)}{p_d(X)\cdot f(M|d,X)}\cdot[Y - \mu(d,M,X)]\\
&+\frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X) - \overbrace{\int_{m\in\mathcal{M}}\mu(d,m,X)\cdot f(m|1-d,X)dm}^{=:\nu(1-d,X)}\Big]\\
&+\underbrace{\int_{m\in\mathcal{M}}\mu(d,m,X)\cdot f(m|1-d,X)dm}_{=:\nu(1-d,X)} \;-\; \psi_{d0}.
\end{aligned}$$

Assumption 3.1: Moment Condition, Linear scores and Neyman orthogonality

Assumption 3.1(a): Moment Condition:

The moment condition $E\Big[\psi_d(W,\eta_0,\psi_{d0})\Big] = 0$ is satisfied:

$$\begin{aligned}
E\Big[\psi_d(W,\eta_0,\psi_{d0})\Big] \;&=\; E\left[\overbrace{E\left[\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot[Y-\mu_0(d,M,X)]\middle|X\right]}^{=E[E[Y-\mu_0(d,M,X)|D=d,M,X]|D=1-d,X]=0}\right]\\
&\quad + E\left[\overbrace{E\left[\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot[\mu_0(d,M,X)-\nu_0(1-d,X)]\middle|X\right]}^{=E[\mu_0(d,M,X)-\nu_0(1-d,X)|D=1-d,X]=0}\right]\\
&\quad + E[\nu_0(1-d,X)] \;-\; \psi_{d0}\\
&= \;\psi_{d0} \;-\; \psi_{d0} \;= 0,
\end{aligned}$$

where the first equality follows from the law of iterated expectations. To better see this result, note that

$$\begin{aligned}
& E\left[\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot[Y-\mu_0(d,M,X)]\middle|X\right]\\
=\;& E\left[\frac{I\{D=d\}\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))}\cdot[Y-\mu_0(d,M,X)]\middle|X\right]\\
=\;& E\left[E\left[\frac{I\{D=d\}}{p_{d0}(M,X)}\cdot[Y-\mu_0(d,M,X)]\middle|M,X\right]\cdot\frac{(1-p_{d0}(M,X))}{(1-p_{d0}(X))}\middle|X\right]\\
=\;& E\left[E[Y-\mu_0(d,M,X)|D=d,M,X]\cdot\frac{(1-p_{d0}(M,X))}{(1-p_{d0}(X))}\middle|X\right]\\
=\;& E[E[Y-\mu_0(d,M,X)|D=d,M,X]|D=1-d,X]\\
=\;& E[\mu_0(d,M,X)-\mu_0(d,M,X)|D=1-d,X]=0,
\end{aligned}$$

where the first equality follows from Bayes' Law, the second from the law of iterated expectations, the third from basic probability theory, and the fourth from Bayes' Law. Furthermore,

$$\begin{aligned}
& E\left[\frac{I\{D=1-d\}}{1-p_{d0}(X)} \cdot [\mu_0(d,M,X) - \nu_0(1-d,X)]\middle| X\right] \\
= \;& E\left[E\left[\frac{I\{D=1-d\}}{1-p_{d0}(X)} \cdot [\mu_0(d,M,X) - \nu_0(1-d,X)]\middle| M,X\right]\middle| X\right] \\
= \;& E\left[[\mu_0(d,M,X) - \nu_0(1-d,X)] \cdot \frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}\middle| X\right] \\
= \;& E[\mu_0(d,M,X) - \nu_0(1-d,X)|D=1-d,X] = E[\mu_0(d,M,X)|D=1-d,X] - \nu_0(1-d,X) \\
= \;& \nu_0(1-d,X) - \nu_0(1-d,X) = 0,
\end{aligned}$$

where the first equality follows from the law of iterated expectations and the third from Bayes' Law.

Assumption 3.1(b): Linearity:

The score $\psi_d(W,\eta_0,\psi_{d0})$ is linear in ψ_{d0} as it can be written as: $\psi_d(W,\eta_0,\psi_{d0}) = \psi_d^a(W,\eta_0) \cdot \psi_{d0} + \psi_d^b(W,\eta_0)$ with $\psi_d^a(W,\eta_0) = -1$ and

$$\begin{aligned}
\psi_d^b(W,\eta_0) &= \frac{I\{D=d\} \cdot f_0(M|1-d,X)}{p_{d0}(X) \cdot f_0(M|d,X)}[Y - \mu_0(d,M,X)] \\
&+ \frac{I\{D=1-d\}}{1-p_{d0}(X)}\Big[\mu_0(d,M,X) - \nu_0(1-d,X)\Big] + \nu_0(1-d,X).
\end{aligned}$$

Assumption 3.1(c): Continuity:

The expression for the second Gateaux derivative of a map $\eta \mapsto E\Big[\psi_d(W,\hat{\eta},\psi_{d0})\Big]$, given in (2.5), is continuous.

Assumption 3.1(d): Neyman Orthogonality:

For any $\eta \in \mathcal{T}_n$, the Gateaux derivative in the direction $\eta - \eta_0 = (\mu(d,M,X) - \mu_0(d,M,X), f(M|D,X) - f_0(M|D,X), p_d(X) - p_{d0}(X))$ is given by:

$$
\begin{aligned}
&\partial E\big[\psi_d(W,\eta,\psi_{d0})\big]\big[\eta-\eta_0\big]\\
&= E\Bigg[\frac{\big[f(M|1-d,X)-f_0(M|1-d,X)\big]\cdot f_0(M|d,X)-\big[f(M|d,X)-f_0(M|d,X)\big]\cdot f_0(M|1-d,X)}{f_0^2(M|d,X)}\cdot\overbrace{\frac{I\{D=d\}}{p_{d0}(X)}\cdot\Big(Y-\mu_0(d,M,X)\Big)}^{E[\cdot|X]=E[Y-\mu_0(d,M,X)|D=d,X]=0}\ \Bigg]\\
&\underbrace{-\ E\Bigg[\underbrace{\frac{I\{D=1-d\}}{1-p_{d0}(X)}}_{E[\cdot|X]=1}\cdot\partial E[\nu_0(1-d,X)][f(M|1-d,X)-f_0(M|1-d,X)]\Bigg]+\partial E[\nu_0(1-d,X)][f(M|1-d,X)-f_0(M|1-d,X)]}_{=0}\\
&-\ E\Bigg[\underbrace{\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot\Big(Y-\mu_0(d,M,X)\Big)}_{E[\cdot|X]=E[E[Y-\mu_0(d,M,X)|D=d,M,X]|D=1-d,X]=0}\cdot\frac{p_d(X)-p_{d0}(X)}{p_{d0}(X)}\Bigg]\\
&+\ E\Bigg[\underbrace{\frac{I\{D=1-d\}}{(1-p_{d0}(X))}\cdot\Big(\mu_0(d,M,X)-\nu_0(1-d,X)\Big)}_{E[\cdot|X]=E[\mu_0(d,M,X)-\nu_0(1-d,X)|D=1-d,X]=0}\cdot\frac{p_d(X)-p_{d0}(X)}{(1-p_{d0}(X))}\Bigg]\\
&-\quad\underbrace{E\Bigg[\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot\Big[\mu(d,M,X)-\mu_0(d,M,X)\Big]\Bigg]}_{E[\cdot]=E[E[\cdot|M,X]]=E\left[\frac{p_{d0}(M,X)\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}\cdot[\mu(d,M,X)-\mu_0(d,M,X)]\right]} \qquad (*)\\
&+\quad\underbrace{E\Bigg[\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot\Big[\mu(d,M,X)-\mu_0(d,M,X)\Big]\Bigg]}_{E[\cdot]=E[E[\cdot|M,X]]=E\left[\frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}\cdot[\mu(d,M,X)-\mu_0(d,M,X)]\right]} \qquad (**)\\
&\underbrace{-\ E\Bigg[\underbrace{\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot\partial E[\nu_0(1-d,X)][\mu(d,M,X)-\mu_0(d,M,X)]}_{E[\cdot|X]=\frac{1-p_{d0}(X)}{1-p_{d0}(X)}\cdot\partial E[\nu_0(1-d,X)][\mu(d,M,X)-\mu_0(d,M,X)]}\Bigg]+\partial E[\nu_0(1-d,X)][\mu(d,M,X)-\mu_0(d,M,X)],}_{=0}
\end{aligned}
$$

where terms $(*)$ and $(**)$ cancel out by Bayes' Law, $\frac{p_{d0}(M,X)\cdot f_0(M|1-d,X)}{p_{d0}(X)\cdot f_0(M|d,X)}=\frac{p_{d0}(M,X)\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))}=\frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}$. Thus, it follows that:

$$\partial E\big[\psi_d(W,\eta_0,\psi_{d0})\big]\big[\eta-\eta_0\big]=0$$

proving that the score function is orthogonal.

Assumption 3.1(e): Singular values of $E[\psi_d^a(W;\eta_0)]$ are bounded:

This holds trivially, because $\psi_d^a(W,\eta_0)=-1$.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

Assumption 3.2(a):

This assumption follows directly from the regularity conditions (Assumption 4) and the definition of $\mathcal{T}_n$ given in (2.13).

Assumption 3.2(b):

Bounds for m_n:

We have

$$\begin{aligned}
\|\mu_0(D,M,X)\|_q &= \left(E\left[|\mu_0(D,M,X)|^q\right]\right)^{1/q} = \left(\sum_{d\in\{0,1\}} E\left[|\mu_0(d,M,X)|^q\, Pr(D=d|M,X)\right]\right)^{1/q} \\
&\geq \epsilon^{1/q}\left(\sum_{d\in\{0,1\}} E\left[|\mu_0(d,M,X)|^q\right]\right)^{1/q} \\
&\geq \epsilon^{1/q}\left(\max_{d\in\{0,1\}} E\left[|\mu_0(d,M,X)|^q\right]\right)^{1/q} \\
&= \epsilon^{1/q}\max_{d\in\{0,1\}}\left(E\left[|\mu_0(d,M,X)|^q\right]\right)^{1/q} = \epsilon^{1/q}\max_{d\in\{0,1\}}\|\mu_0(d,M,X)\|_q\,.
\end{aligned}$$

The first equality follows from definition, the second from the law of total probability, the first inequality from $Pr(D=d|M,X)\geq\epsilon$. Using the same line of arguments we get that

$$\|f_0(M|D,X)\|_q \geq \epsilon^{1/q}\max_{d\in\{0,1\}}\|f_0(M|d,X)\|_q$$

Also, by Jensen's inequality $\|\mu_0(D,M,X)\|_q \leq \|Y\|_q$, such that for any $d\in\{0,1\}$:

$$\begin{aligned}
\|\mu_0(d,M,X)\|_q &\leq C/\epsilon^{1/q}, \qquad (2.14)\\
\|f_0(M|d,X)\|_q &\leq C/\epsilon^{1/q},
\end{aligned}$$

because of $\|Y\|_q \leq C$ by Assumption 4(a).

Similarly, for any $\eta\in\mathcal{T}_n$ we obtain:

$$\begin{aligned}
\|\mu(d,M,X)-\mu_0(d,M,X)\|_q &\leq C/\epsilon^{1/q},\\
\|f(M|d,X)-f_0(M|d,X)\|_q &\leq C/\epsilon^{1/q},
\end{aligned}$$

due to the definition of $\mathcal{T}_n$ given in (2.13).

Also,

$$\begin{aligned}
\|\nu_0(1-d,X)\|_q &= \left(E\left[|\nu_0(1-d,X)|^q\right]\right)^{1/q} \qquad (2.15)\\
&= \left(E\left[\left|\int_{m\in\mathcal{M}} \mu_0(d,m,X)\cdot f_0(m|1-d,X)dm\right|^q\right]\right)^{1/q}\\
&\leq \left(E\left[\int_{m\in\mathcal{M}} |\mu_0(d,m,X)|^q\cdot f_0(m|1-d,X)dm\right]\right)^{1/q}\\
&= \left(E\left[\int_{m\in\mathcal{M}} |\mu_0(d,m,X)|^q\cdot f(m|d,X)\cdot \frac{f_0(m|1-d,X)}{f(m|d,X)}dm\right]\right)^{1/q}\\
&\leq R^{1/q}\left(E\left[\int_{m\in\mathcal{M}} |\mu_0(d,m,X)|^q\cdot f_0(m|d,X)dm\right]\right)^{1/q}\\
&= R^{1/q}\|\mu_0(d,M,X)\|_q \leq \epsilon^{1/q}R^{1/q}\|\mu_0(D,M,X)\|_q
\end{aligned}$$

where we make use of the definition of ν_0, Jensen's inequality, and the boundedness of the ratio of densities. We therefore obtain $\|\nu_0(1-d,X)\|_q \leq C/(\epsilon^{1/q}R^{1/q})$ by inequality (2.14).

This permits bounding the following quantities:

$$\begin{aligned}
\|\mu(d,M,X)\|_q &\leq \|\mu(d,M,X)-\mu_0(d,M,X)\|_q + \|\mu_0(d,M,X)\|_q \leq 2C/\epsilon^{1/q}, \qquad (2.16)\\
\|\nu(1-d,X)\|_q &\leq \|\nu(1-d,X)-\nu_0(1-d,X)\|_q + \|\nu_0(1-d,X)\|_q \leq 2C/(\epsilon^{1/q}R^{1/q}),\\
|\psi_{d0}| &= |E[\nu_0(1-d,X)]| \leq E\Big[|\nu_0(1-d,X)|^1\Big]^{1/1} = \|\nu_0(1-d,X)\|_1\\
&\leq \|\nu_0(1-d,X)\|_2 \leq \|Y_2\|_2/(\epsilon^{1/2}R^{1/2}) \overbrace{\leq}^{q>2} \|Y_2\|_q/(\epsilon^{1/2}R^{1/2}) \leq C/(\epsilon^{1/2}R^{1/2}),
\end{aligned}$$

using the triangular inequality, Jensen's inequality, and properties of statistical l_q norms.

Finally, rearranging $\psi_d(W,\eta,\psi_{d0})$

$$\begin{aligned}
\psi_d(W,\eta,\psi_{d0}) &= \underbrace{\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_d(X)\cdot f_0(M|d,X)}\cdot Y}_{=I_1}\\
&+ \underbrace{\left(\frac{I\{D=1-d\}}{1-p_d(X)} - \frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_d(X)\cdot f_0(M|d,X)}\right)\cdot \mu(d,M,X)}_{=I_2}\\
&+ \underbrace{\left(1-\frac{I\{D=1-d\}}{1-p_d(X)}\right)\nu(1-d,X)}_{=I_3} - \psi_{d0},
\end{aligned}$$

provides

$$\begin{aligned}\|\psi_d(W,\eta,\psi_{d0})\|_q &\leq \|I_1\|_q + \|I_2\|_q + \|I_3\|_q + \|\psi_{d0}\|_q \\ &\leq \frac{R}{\epsilon}\|Y\|_q + \frac{1+R}{\epsilon}\|\mu(d,M,X)\|_q + \\ &+ \frac{1-\epsilon}{\epsilon}\|\nu(1-d,X)\|_q + |\psi_{d0}| \\ &\leq C\left(\frac{R}{\epsilon} + \frac{2}{\epsilon^{1+1/q}}\left(1+R+\frac{1-\epsilon}{R^{1/q}}\right) + \frac{1}{\epsilon^{1/2}R^{1/2}}\right),\end{aligned}$$

making use of the triangular inequality and inequalities (2.16). This provides the upper bound on m_n in Assumption 3.2(b).

Bound for m'_n:

We note that

$$\Big(E[|\psi_d^a(W,\eta)|^q]\Big)^{1/q} = 1,$$

which provides the upper bound on m'_n in Assumption 3.2(b).

Assumption 3.2(c)

Bound for r_n: For any $\eta = (\mu, f, p_d)$ we have

$$\Big|E\Big(\psi_d^a(W,\eta) - \psi_d^a(W,\eta_0)\Big)\Big| = |1-1| = 0 \leq \delta'_n,$$

providing the bound on r_n in Assumption 3.2(c).

Bound for r'_n:

Using the triangular inequality

$$\begin{aligned}\|\psi_d(W,\eta,\psi_{d0}) - \psi_d(W,\eta_0,\psi_{d0})\|_2 &\leq \left\|I\{D=d\}\cdot Y\cdot\left(\frac{f(M|1-d,X)}{p_d(X)f(M|d,X)} - \frac{f_0(M|1-d,X)}{p_{d0}(X)f_0(M|d,X)}\right)\right\|_2 \\ &+ \left\|I\{D=d\}\cdot\left(\frac{\mu(d,M,X)f(M|1-d,X)}{p_d(X)f(M|d,X)} - \frac{\mu_0(d,M,X)f_0(M|1-d,X)}{p_{d0}(X)f_0(M|d,X)}\right)\right\|_2 \\ &+ \left\|I\{D=1-d\}\cdot\left(\frac{\mu(d,M,X)}{1-p_d(X)} - \frac{\mu_0(d,M,X)}{1-p_{d0}(X)}\right)\right\|_2 + \left\|I\{D=1-d\}\cdot\left(\frac{\nu(1-d,X)}{1-p_d(X)} - \frac{\nu_0(1-d,X)}{1-p_{d0}(X)}\right)\right\|_2 \\ &+ \|\nu(1-d,X) - \nu_0(1-d,X)\|_2 \\ &\leq \delta_n\left(\frac{C\cdot R^2}{\epsilon^2} + \frac{C\cdot R^2}{\epsilon^2}\left(\frac{1}{\epsilon^{1/2}} + \frac{C}{\epsilon^{1/2}}\right) + \frac{1}{\epsilon^2}\left(\frac{1}{\epsilon^{1/2}} + \frac{C}{\epsilon^{1/2}}\right) + \frac{1}{\epsilon^2}\left(\frac{1}{\epsilon^{1/2}R^{1/2}} + \frac{C}{R^{1/2}}\right) + \frac{1}{\epsilon^{1/2}R^{1/2}}\right) \leq \delta'_n,\end{aligned}$$

as long as C_ϵ in the definition of δ'_n is sufficiently large. This gives the bound on r'_n in Assumption 3.2(c). In order to show the second to the last inequalities, we provide bounds for the terms below, where we made use of the facts that $\|\mu(d,M,X)-\mu_0(d,M,X)\|_2 \leq \delta_n/\epsilon^{1/2}$, and $\|\nu(1-d,X)-\nu_0(1-d,X)\|_2 \leq \delta_n/(\epsilon^{1/2}R^{1/2})$ using similar steps as in Assumption 3.1(b) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

For the first term:

$$
\begin{aligned}
& \left\| I\{D=d\}\cdot Y \cdot \left(\frac{f(M|1-d,X)}{p_d(X)f(M|d,X)} - \frac{f_0(M|1-d,X)}{p_{d0}(X)f_0(M|d,X)} \right) \right\|_2 \\
\leq \;& C \cdot \left\| \frac{f(M|1-d,X)}{p_d(X)f(M|d,X)} - \frac{f_0(M|1-d,X)}{p_{d0}(X)f_0(M|d,X)} \right\|_2 \\
\leq \;& \frac{C}{\epsilon^2 \underline{f}^2} \left\| f(M|1-d,X)f_0(M|1-d,X)p_{d0}(X) - f(M|1-d,X)f_0(M|1-d,X)p_d(X) \right\|_2 \\
\leq \;& \frac{C\cdot \overline{f}^2}{\epsilon^2 \underline{f}^2} \left\| p_{d0}(X) - p_d(X) \right\|_2 \leq \delta_n \frac{C\cdot R^2}{\epsilon^2},
\end{aligned}
$$

where we use $\left\|E[Y^2|d,M,X]\right\|_\infty \leq C^2$ (see our Assumption 4(a)) in the first inequality.

For the second term:

$$
\begin{aligned}
& \left\| I\{D=d\}\cdot \left(\frac{\mu(d,M,X)f(M|1-d,X)}{p_d(X)f(M|d,X)} - \frac{\mu_0(d,M,X)f_0(M|1-d,X)}{p_{d0}(X)f_0(M|d,X)} \right) \right\|_2 \\
\leq \;& \left\| \frac{\mu(d,M,X)f(M|1-d,X)}{p_d(X)f(M|d,X)} - \frac{\mu_0(d,M,X)f_0(M|1-d,X)}{p_{d0}(X)f_0(M|d,X)} \right\|_2 \\
\leq \;& \frac{C}{\epsilon^2 \underline{f}^2} \left\| \mu(d,M,X)f(M|1-d,X)f_0(M|1-d,X)p_{d0}(X) - \mu_0(d,M,X)f(M|1-d,X)f_0(M|1-d,X)p_d(X) \right. \\
\leq \;& \frac{C\overline{f}^2}{\epsilon^2 \underline{f}^2} \left\| \mu(d,M,X)p_{d0}(X) - \mu_0(d,M,X)p_d(X) \right\|_2 \\
= \;& \frac{C\cdot R^2}{\epsilon^2} \left\| \mu(d,M,X)p_{d0}(X) - \mu_0(d,M,X)p_d(X) + \mu_0(d,M,X)p_{d0}(X) - \mu_0(d,M,X)p_{d0}(X) \right\|_2 \\
\leq \;& \frac{C\cdot R^2}{\epsilon^2} \big(\left\| p_{d0}(X)(\mu(d,M,X)-\mu_0(d,M,X)) \right\|_2 + \left\| \mu_0(d,M,X)(p_{d0}(X)-p_d(X)) \right\|_2 \big) \\
\leq \;& \frac{C\cdot R^2}{\epsilon^2} \big(\left\| \mu(d,M,X)-\mu_0(d,M,X) \right\|_2 + C\left\| (p_{d0}(X)-p_d(X)) \right\|_2 \big) \\
\leq \;& \frac{C\cdot R^2}{\epsilon^2} \left(\frac{\delta_n}{\epsilon^{1/2}} + C\delta_n \right) = \delta_n \frac{C\cdot R^2}{\epsilon^2} \left(\frac{1}{\epsilon^{1/2}} + C \right)
\end{aligned}
$$

where the fifth inequality follows from $E[Y^2|D=d,M,X] \geq (E[Y|D=d,M,X])^2 = \mu_0^2(d,M,X)$ by conditional Jensen's inequality and therefore $\|\mu_0(d,M,X)\|_\infty \leq C^2$.

For the third term:

$$\begin{aligned}
&\left\| I\{D=1-d\}\cdot\left(\frac{\mu(d,M,X)}{1-p_d(X)}-\frac{\mu_0(d,M,X)}{1-p_{d0}(X)}\right)\right\|_2 \leq \left\|\frac{\mu(d,M,X)}{1-p_d(X)}-\frac{\mu_0(d,M,X)}{1-p_{d0}(X)}\right\|_2\\
\leq\ & \frac{1}{\epsilon^2}\left\|\mu(d,M,X)p_{1-d,0}-\mu_0(d,M,X)p_{1-d}\right\|_2\\
=\ & \frac{1}{\epsilon^2}\left\|\mu(d,M,X)p_{1-d,0}-\mu_0(d,M,X)p_{1-d}+\mu_0(d,M,X)p_{1-d,0}-\mu_0(d,M,X)p_{1-d,0}\right\|_2\\
\leq\ & \frac{1}{\epsilon^2}\big(\left\|p_{1-d,0}(\mu(d,M,X)-\mu_0(d,M,X))\right\|_2+\left\|\mu_0(d,M,X)(p_{1-d,0}-p_{1-d})\right\|_2\big)\\
\leq\ & \frac{1}{\epsilon^2}\big(\left\|\mu(d,M,X)-\mu_0(d,M,X)\right\|_2+C\left\|p_{1-d,0}-p_{1-d}\right\|_2\big)\\
\leq\ & \frac{1}{\epsilon^2}\left(\frac{\delta_n}{\epsilon^{1/2}}+C\delta_n\right)=\delta_n\frac{1}{\epsilon^2}\left(\frac{1}{\epsilon^{1/2}}+C\right).
\end{aligned}$$

For the fourth term:

$$\begin{aligned}
&\left\| I\{D=1-d\}\cdot\left(\frac{\nu(1-d,X)}{1-p_d(X)}-\frac{\nu_0(1-d,X)}{1-p_{d0}(X)}\right)\right\|_2\\
\leq\ & \frac{1}{\epsilon^2}\big(\left\|p_{1-d,0}(\nu(1-d,X)-\nu_0(1-d,X))\right\|_2+\left\|\nu_0(1-d,X)(p_{1-d,0}-p_{1-d})\right\|_2\big)\\
\leq\ & \frac{1}{\epsilon^2}\big(\left\|\nu(1-d,X)-\nu_0(1-d,X)\right\|_2+\frac{C}{R^{1/2}}\left\|p_{1-d,0}-p_{1-d}\right\|_2\big)\\
\leq\ & \frac{1}{\epsilon^2}\left(\frac{\delta_n}{\epsilon^{1/2}R^{1/2}}+\frac{C}{R^{1/2}}\delta_n\right)=\delta_n\frac{1}{\epsilon^2}\left(\frac{1}{\epsilon^{1/2}R^{1/2}}+\frac{C}{R^{1/2}}\right),
\end{aligned}$$

where we used Jensen's inequality similarly to 2.15 in order to get $E[\nu_0^2(1-d,X)] \leq R\cdot E[\mu_0^2(d,M,X)]$ and hence$\|\nu_0(1-d,X)\|_\infty \leq C^2/R$.

Bound on λ'_n*:*

Consider

$$f(r) := E[\psi(W,\eta_0+r(\eta-\eta_0),\psi_{d0})]$$

We subsequently omit arguments for the sake of brevity and use $p_d = p_d(X), f_d = f_d(M|d,X), \mu = \mu(d,M,X), \nu = \nu(1-d,X)$ and similarly $p_{d0}, f_{0d}, \mu_0, \nu_0$.

For any $r \in (0,1)$:

$$\begin{aligned}
\frac{\partial^2 f(r)}{\partial r^2} &= E\left[2\cdot I\{D=1-d\}\frac{(\mu-\mu_0)(p_d-p_{d0})}{(1-p_{d0}+r(p_{d0}-p_d))^2}\right] + E\left[2\cdot I\{D=1-d\}\frac{(\nu-\nu_0)(p_d-p_{d0})}{(1-p_{d0}+r(p_{d0}-p_d))^2}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(f_d-f_{d0})(f_{1-d}-f_{1-d,0})\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))\,(f_{d0}+r(f_d-f_{d0}))^2}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(p_d-p_{d0})(f_{1-d}-f_{1-d,0})\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))^2\,(f_{d0}+r(f_d-f_{d0}))}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(f_d-f_{d0})\,(f_{1-d,0}+r(f_{1-d}-f_{1-d,0}))\,(-(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))\,(f_{d0}+r(f_d-f_{d0}))^2}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(p_d-p_{d0})\,(f_{1-d,0}+r(f_{1-d}-f_{1-d,0}))\,(-(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))^2\,(f_{d0}+r(f_d-f_{d0}))}\right] \\
&+ E\left[(-2)\cdot I\{D=d\}\frac{(f_{1-d}-f_{1-d,0})\,(\mu-\mu_0)}{(p_{d0}+r(p_d-p_{d0}))\,(f_{d0}+r(f_d-f_{d0}))}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(f_d-f_{d0})^2\,(f_{1-d,0}+r(f_{1-d}-f_{1-d,0}))\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))\,(f_{d0}+r(f_d-f_{d0}))^3}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(f_d-f_{d0})(p_d-p_{d0})\,(f_{1-d,0}+r(f_{1-d}-f_{1-d,0}))\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))^2\,(f_{d0}+r(f_d-f_{d0}))^2}\right] \\
&+ E\left[2\cdot I\{D=d\}\frac{(p_d-p_{d0})^2\,(f_{1-d,0}+r(f_{1-d}-f_{1-d,0}))\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{d0}+r(p_d-p_{d0}))^3\,(f_{d0}+r(f_d-f_{d0}))}\right] \\
&+ E\left[2\cdot I\{D=1-d\}\frac{(\mu_0-\nu_0)\,(p_d-p_{d0})^2}{(1-p_{d0}+r(p_{d0}-p_d))^3}\right] \\
&+ E\left[2\cdot I\{D=1-d\}\frac{(r(\mu-\mu_0)-r(\nu-\nu_0))\,(p_d-p_{d0})^2}{(1-p_{d0}+r(p_{d0}-p_d))^3}\right]
\end{aligned}$$

Note that the following inequalities can be shown to hold using similar steps as in Assumption 3.1(b) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018):

$$\begin{aligned}
\|\mu-\mu_0\|_2 &= \|\mu(d,M,X)-\mu_0(d,M,X)\|_2 \le \|\mu(D,M,X)-\mu_0(D,M,X)\|_2/\epsilon^{1/2} \le \delta_n/\epsilon^{1/2}, \\
\|\nu-\nu_0\|_2 &= \|\nu(1-d,X)-\nu_0(1-d,X)\|_2 \le \|\mu(D,M,X)-\mu_0(D,M,X)\|_2/\epsilon^{1/2}R^{1/2} \le \delta_n/(\epsilon^{1/2}R^{1/2}),
\end{aligned}$$

These inequalities together with our Assumption 4 imply

$$\begin{aligned} E[Y-\mu_0(d,M,X)|D=d,M,X] &= 0, \\ |p_d - p_{d0}| &\leq 2, \\ \|\mu_0\|_q \leq \|Y\|_q/\epsilon^{1/q} &\leq C/\epsilon^{1/q} \\ \|\mu-\mu_0\|_2 \times \|p_d - p_{d0}\|_2 &\leq \delta_n n^{-1/2}/\epsilon^{1/2}, \\ \|\mu-\mu_0\|_2 \times \|f_{1-d} - f_{1-d,0}\|_2 &\leq \delta_n n^{-1/2}/(\epsilon R^{1/2}), \end{aligned}$$

for all $d \in \{1,0\}$ and consequently

$$\|\nu - \nu_0\|_2 \times \|p_d - p_{d0}\|_2 \leq \delta_n n^{-1/2}/(\epsilon^{1/2}R^{1/2}).$$

Putting everything together, we get that for some value C''_ϵ that only depends on C and ϵ

$$\left|\frac{\partial^2 f(r)}{\partial r^2}\right| \leq C''_\epsilon \delta_n n^{-1/2} \leq \delta'_n n^{-1/2}.$$

This gives the upper bound on λ'_n in Assumption 3.2(c) as long as C_ϵ in the definition of δ'_n satisfies $C_\epsilon \geq C''_\epsilon$. In order to verify that this inequality holds we consider all the terms in $\frac{\partial^2 f(r)}{\partial r^2}$ separately. For the first term we obtain

$$\left|E\left[2\cdot I\{D=1-d\}\frac{(\mu-\mu_0)(p_d-p_{d0})}{(1-p_{d0}+r(p_{d0}-p_d))^2}\right]\right| \leq \frac{2}{\epsilon^3}\left|E\left[(\mu-\mu_0)(p_d-p_{d0})\right]\right| \leq \frac{2}{\epsilon^3}\frac{\delta_n}{\epsilon^{1/2}R^{1/2}}n^{-1/2},$$

where we made use of the fact that $1 \geq p_{d0} + r(p_d - p_{d0}) = (1-r)p_{d0} + rp_d \geq (1-r)\epsilon + r\epsilon = \epsilon$, $\underline{f} \leq f_{d0} + r(f_d - f_{d0}) \leq \overline{f}$, and Holder's inequality. For the third term we obtain

$$\begin{aligned} &\left|E\left[2\cdot I\{D=d\}\frac{(f_d - f_{d0})(f_{1-d}-f_{1-d,0})\left(Y-\mu_0-r(\mu-\mu_0)\right)}{(p_{d0}+r(p_d-p_{d0}))\left(f_{d0}+r(f_d-f_{d0})\right)^2}\right]\right| \\ \leq\;& \frac{2}{\epsilon\underline{f}^2}(\overline{f}-\underline{f})^2\left|E\left[I\{D=d\}\left(Y-\mu_0\right)\right]\right| + \frac{2}{\epsilon\underline{f}^2}\left|E\left[I\{D=d\}(f_d-f_{d0})(f_{1-d}-f_{1-d,0})r(\mu-\mu_0)\right]\right| \\ \leq\;& \frac{2}{\epsilon\underline{f}^2}(\overline{f}-\underline{f})\left|E\left[1\cdot(f_{1-d}-f_{1-d,0})(\mu-\mu_0)\right]\right| \leq \frac{2}{\epsilon\underline{f}^2}(\overline{f}-\underline{f})\frac{\delta_n}{\epsilon^{1/2}}n^{-1/2}. \end{aligned}$$

And for the second to the last terms, we obtain

$$
\begin{aligned}
&\left|E\left[2\cdot I\{D=1-d\}\frac{(\mu_0-\nu_0)\,(p_d-p_{d0})^2}{(1-p_{d0}+r(p_{d0}-p_d))^3}\right]\right|\\
= &\left|E\left[\overbrace{I\{D=1-d\}\frac{(\mu_0-\nu_0)}{p_{1-d,0}}}^{E[E[\cdot|M,X]|X]=0}\cdot\frac{p_{1-d,0}(p_d-p_{d0})^2}{(1-p_{d0}+r(p_{d0}-p_d))^3}\right]\right|=0.
\end{aligned}
$$

All the remaining terms are bounded similarly.

Assumption 3.2(d):

Finally, we consider

$$
\begin{aligned}
E\Big[(\psi_d(W,\eta,\psi_{d0}))^2\Big] &= E\Bigg[\Bigg(\underbrace{\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_d(X)\cdot f_0(M|d,X)}\cdot(Y-\mu_0(d,M,X))}_{=I_1}\\
&+ \underbrace{\left(\frac{I\{D=1-d\}}{1-p_d(X)}\right)\cdot(\mu_0(d,M,X)-\nu_0(1-d,X))}_{=I_2}+\underbrace{\nu_0(1-d,X)-\psi_{d0}}_{=I_3}\Bigg)^2\Bigg]\\
&= E[I_1^2+I_2^2+I_3^2]\geq E[I_1^2]\\
&= E\left[\left(\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_d(X)\cdot f_0(M|d,X)}\right)^2(Y-\mu_0(d,M,X))^2\right]\\
&\geq \frac{\underline{f}^2}{(1-\epsilon)\overline{f}^2}E\left[(Y-\mu_0(d,M,X))^2\right]\geq\frac{c^2}{(1-\epsilon)R^2}>0,
\end{aligned}
$$

where the second equality follows from

$$
\begin{aligned}
E\Big[I_1\cdot I_2\Big] &= E\Bigg[\overbrace{\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_d(X)\cdot f_0(M|d,X)}\frac{I\{D=1-d\}}{1-p_d(X)}}^{I\{D=d\}\cdot I\{D=1-d\}=0}\cdot(Y-\mu_0(d,M,X))\cdot(\mu_0(d,M,X)-\nu_0(1-d,X))\\
E\Big[I_2\cdot I_3\Big] &= E\Bigg[\overbrace{\frac{I\{D=1-d\}}{1-p_d(X)}\cdot(\mu_0(d,M,X)-\nu_0(1-d,X))}^{E[\cdot|X]=0}\cdot(\nu_0(1-d,X)-\psi_{d0})\Bigg],\\
E\Big[I_1\cdot I_3\Big] &= E\Bigg[\overbrace{\frac{I\{D=d\}\cdot f_0(M|1-d,X)}{p_d(X)\cdot f_0(M|d,X)}\cdot(Y-\mu_0(d,M,X))}^{E[\cdot|X]=0}\cdot(\nu_0(1-d,X)-\psi_{d0})\Bigg].
\end{aligned}
$$

Counterfactual $E[Y(d,m)]$

The score for the estimation of $E[Y(d,m)]$ based on (2.8) is given by:

$$E\big[\psi_{dm}(W,\eta,\psi_{dm0})\big] = E\left[\frac{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu(d,m,X)]}{f(m|d,X)\cdot p_d(X)} + \mu(d,m,X) - \psi_{dm0}\right].$$

Assumption 3.1: Moment Condition, Linear scores and Neyman orthogonality

Assumption 3.1(a): Moment condition:

The moment condition $E\Big[\psi_{dm}(W,\eta_0,\psi_{dm0})\Big] = 0$ is satisfied:

$$\begin{aligned} E\Big[\psi_{dm}(W,\eta_0,\psi_{dm0})\Big] &= E\left[\frac{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu_0(d,m,X)]}{f_0(m|d,X)\cdot p_{d0}(X)} + \mu_0(d,m,X) - \psi_{dm0}\right] \\ &= E\left[\overbrace{E\left[\frac{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu_0(d,m,X)]}{f_0(m|d,X)\cdot p_{d0}(X)}\middle| X\right]}^{=E[Y-\mu_0(d,m,X)|d,m,X]=0}\right] + E\Big[\mu_0(d,m,X)\Big] - \psi_{dm0} \\ &= \psi_{dm0} - \psi_{dm0} = 0. \end{aligned}$$

Assumption 3.1(b): Linearity:

The score $\psi_{dm}(W,\eta_0,\psi_{dm0})$ is linear in ψ_{dm0} as it can be written as: $\psi_{dm}(W,\eta_0,\psi_{dm0}) = \psi_d^a(W,\eta_0)\cdot\psi_{dm0} + \psi_d^b(W,\eta_0)$ with $\psi_d^a(W,\eta_0) = -1$ and

$$\psi_d^b(W,\eta_0) = \frac{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu(d,m,X)]}{f(m|d,X)\cdot p_d(X)} + \mu(d,m,X)$$

Assumption 3.1(c): Continuity:

The expression for the second Gateaux derivative of a map $\eta \mapsto E\Big[\psi_{dm}(W,\eta,\psi_{dm0})\Big]$, is continuous.

Assumption 3.1(d): Neyman orthogonality:

The Gateaux derivative in the direction $\eta - \eta_0 = (\mu(d,M,X) - \mu_0(d,M,X), f(M|D,X) -$

$f_0(M|D,X), p_d(X)-p_{d0}(X))$ is given by:

$$\begin{aligned}
&\partial E\big[\psi_{dm}(W,\eta,\psi_{dm0})\big]\big[\eta-\eta_0\big]\\
&= \overbrace{-E\Bigg[\underbrace{\frac{I\{D=d\}\cdot I\{M=m\}}{f_0(m|d,X)\cdot p_{d0}(X)}}_{E[\cdot|X]=\frac{\Pr(D=d,M=m|X)}{\Pr(D=d,M=m|X)}=1}\cdot\Big[\mu(d,m,X)-\mu_0(d,m,X)\Big]\Bigg]+E\Big[\mu(d,m,X)-\mu_0(d,m,X)\Big]}^{=0}\\
&\quad -E\Bigg[\frac{\overbrace{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu_0(d,m,X)]}^{E[\cdot|X]=E[Y-\mu_0(d,m,X)|d,m,X]=0}}{f_0(m|d,X)\cdot p_{d0}(X)}\cdot\frac{f(m|d,X)-f_0(m|d,X)}{f_0(m|d,X)}\Bigg]\\
&\quad -E\Bigg[\frac{\overbrace{I\{D=d\}\cdot I\{M=m\}\cdot[Y-\mu_0(d,m,X)]}^{E[\cdot|X]=E[Y-\mu_0(d,m,X)|d,m,X]=0}}{f_0(m|d,X)\cdot p_{d0}(X)}\cdot\frac{p_d(X)-p_{d0}(X)}{p_{d0}(X)}\Bigg].
\end{aligned}$$

Thus, it follows that:

$$\partial E\big[\psi_{dm}(W,\eta_0,\psi_{dm0})\big]\big[\eta-\eta_0\big]=0$$

proving that the score function is orthogonal.

Assumption 3.1(e): Singular values of $E[\psi_d^a(W;\eta_0)]$ are bounded:

This holds trivially, because $\psi_d^a(W;\eta_0)=-1$.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

This proof is omitted for the sake of brevity. It follows along similar lines as the proof for $Y(d,M(1-d))$ presented in subsection 2.B.1.

This concludes the proof of Theorem 1. □

2.B.2 Proof of Theorem 2

The alternative score for the counterfactual based on (2.6) is given by:

$$
\begin{aligned}
\psi_d^*(W, \eta^*, \psi_{d0}) = \quad & E\Bigg[\frac{I\{D=d\}\cdot(1-p_d(M,X))}{p_d(M,X)\cdot(1-p_d(X))}\cdot\Big[Y-\mu(d,M,X)\Big] \\
+ \quad & \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Bigg[\mu(d,M,X)-\overbrace{E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]}^{=:\omega(1-d,X)}\Bigg] \\
+ \quad & \overbrace{E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]}^{=:\omega(1-d,X)}\Bigg]-\psi_{d0}
\end{aligned}
$$

with $\eta^* = (\mu(D,M,X), \omega(D,X), p_d(M,X), p_d(X))$.

Let $\mathcal{T}_n^*$ be the set of all η^* consisting of P-square integrable functions $\mu(D,M,X), \omega(D,X), p_d(M,X)$, and $p_d(X)$ such that

$$
\begin{aligned}
\|\eta^*-\eta_0^*\|_q &\leq C, \qquad (2.17)\\
\|\eta^*-\eta_0^*\|_2 &\leq \delta_n,\\
\|p_d(X)-1/2\|_\infty &\leq 1/2-\epsilon,\\
\|p_d(M,X)-1/2\|_\infty &\leq 1/2-\epsilon,\\
\|\mu(D,M,X)-\mu_0(D,M,X)\|_2\times\|p_d(X)-p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2},\\
\|\mu(D,M,X)-\mu_0(D,M,X)\|_2\times\|p_d(M,X)-p_{d0}(M,X)\|_2 &\leq \delta_n n^{-1/2},\\
\|\omega(D,X)-\omega_0(D,X)\|_2\times\|p_d(X)-p_{d0}(X)\|_2 &\leq \delta_n n^{-1/2}.
\end{aligned}
$$

We replace the sequence $(\delta_n)_{n\geq1}$ by $(\delta_n')_{n\geq1}$, where $\delta_n' = C_\epsilon \max(\delta_n, n^{-1/2})$, where C_ϵ is a sufficiently large value that only depends on C and ϵ.

Assumption 3.1: Moment Condition, Linear scores and Neyman orthogonality

Assumption 3.1(a): Moment condition:

The moment condition $E\Big[\psi_d^*(W,\eta_0^*,\psi_{d0})\Big]=0$ is satisfied:

$$
\begin{aligned}
E\Big[\psi_d^*(W,\eta_0^*,\psi_{d0})\Big] \;&=\; E\Bigg[\overbrace{E\Bigg[\frac{I\{D=d\}\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))}\cdot[Y-\mu_0(d,M,X)]\Bigg|X\Bigg]}^{=E[E[Y-\mu_0(d,M,X)|D=d,M,X]|D=1-d,X]=0}\Bigg]\\
&\quad + E\Bigg[\overbrace{E\Bigg[\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot[\mu_0(d,M,X)-\omega_0(1-d,X)]\Bigg|X\Bigg]}^{=E[\mu_0(d,M,X)-\omega_0(1-d,X)|D=1-d,X]=0}\Bigg]\\
&\quad + E[\omega_0(1-d,X)] \quad - \quad \psi_{d0}\\
&=\; \psi_{d0} \quad - \quad \psi_{d0} \quad = 0.
\end{aligned}
$$

To better see this result, note that

$$
\begin{aligned}
& E\Bigg[\frac{I\{D=d\}\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))}\cdot[Y-\mu_0(d,M,X)]\Bigg|X\Bigg]\\
=\;& E\Bigg[E\Bigg[\frac{I\{D=d\}}{p_{d0}(M,X)}\cdot[Y-\mu_0(d,M,X)]\Bigg|M,X\Bigg]\cdot\frac{(1-p_{d0}(M,X))}{(1-p_{d0}(X))}\Bigg|X\Bigg]\\
=\;& E\Bigg[E[Y-\mu_0(d,M,X)|D=d,M,X]\cdot\frac{(1-p_{d0}(M,X))}{(1-p_{d0}(X))}\Bigg|X\Bigg]\\
=\;& E[E[Y-\mu_0(d,M,X)|D=d,M,X]|D=1-d,X]\\
=\;& E[\mu_0(d,M,X)-\mu_0(d,M,X)|D=1-d,X]=0,
\end{aligned}
$$

where the first equality follows from the law of iterated expectations, the second from basic probability theory, and the third from Bayes' Law. Furthermore,

$$
\begin{aligned}
& E\Bigg[\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot[\mu_0(d,M,X)-\omega_0(1-d,X)]\Bigg|X\Bigg]\\
=\;& E\Bigg[E\Bigg[\frac{I\{D=1-d\}}{1-p_{d0}(X)}\cdot[\mu_0(d,M,X)-\omega_0(1-d,X)]\Bigg|M,X\Bigg]\Bigg|X\Bigg]\\
=\;& E\Bigg[[\mu_0(d,M,X)-\omega_0(1-d,X)]\cdot\frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}\Bigg|X\Bigg]\\
=\;& E[\mu_0(d,M,X)-\omega_0(1-d,X)|D=1-d,X]=E[\mu_0(d,M,X)|D=1-d,X]-\omega_0(1-d,X)\\
=\;& \omega_0(1-d,X)-\omega_0(1-d,X)=0,
\end{aligned}
$$

where the first equality follows from the law of iterated expectations and the third from Bayes' Law.

Assumption 3.1(b):Linearity:

The score $\psi_d^*(W,\eta_0^*,\psi_{d0})$ is linear in ψ_{d0} as it can be written as: $\psi_d^*(W,\eta_0^*,\psi_{d0}) = \psi_d^a(W,\psi_{d0}) \cdot \psi_{d0} + \psi_d^b(W,\eta_0^*)$ with $\psi_d^a(W,\eta_0^*) = -1$ and

$$\begin{aligned}
\psi_d^b(W,\eta_0^*) &= \frac{I\{D=d\}\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))} \cdot \Big[Y - \mu_0(d,M,X)\Big] \\
&+ \frac{I\{D=1-d\}}{1-p_{d0}(X)} \cdot \Big[\mu_0(d,M,X) - \omega_0(1-d,X)\Big] + \omega_0(1-d,X)
\end{aligned}$$

Assumption 3.1(c): Continuity:

The expression for the second Gateaux derivative of a map $\eta^* \mapsto E\Big[\psi_d^*(W,\eta^*,\psi_{d0})\Big]$ is continuous.

Assumption 3.1(d): Neyman orthogonality:

The Gateaux derivative in the direction
$\eta^* - \eta_0^* = (\mu_d(d,M,X) - \mu_0(d,M,X), \omega(1-d,X) - \omega_0(1-d,X), p_d(M.X) - p_{d0}(M,X), p_d(X) - p_{d0}(X))$ is given by:

$$\partial E\big[\psi_d^*(W,\eta^*,\psi_{d0})\big]\big[\eta^* - \eta_0^*\big]$$

$$\begin{aligned}
=&E\Bigg[\frac{-[p_d(M,X)-p_{d0}(M,X)]}{p_{d0}(M,X)^2} \cdot \overbrace{\frac{I\{D=d\}}{1-p_{d0}(X)} \cdot \big(Y - \mu(d,M,X)\big)}^{E[\cdot|X]=E[Y-\mu(d,M,X)|D=d,X]\cdot\frac{p_{d0}(X)}{1-p_{d0}(X)}=0}\Bigg] \\
+&E\Bigg[\overbrace{\frac{I\{D=d\}\cdot(1-p_{d0}(M,X))}{p_{d0}(M,X)\cdot(1-p_{d0}(X))} \cdot \big(Y - \mu_0(d,M,X)\big)}^{E[\cdot|X]=E[E[Y-\mu_0(d,M,X)|D=d,M,X]|D=1-d,X]=0} \cdot \frac{p_d(X)-p_{d0}(X)}{(1-p_{d0}(X))}\Bigg] \\
+&E\Bigg[\underbrace{\frac{I\{D=1-d\}}{(1-p_{d0}(X))} \cdot \big(\mu_0(d,M,X) - \omega_0(1-d,X)\big)}_{E[\cdot|X]=E[\mu_0(d,M,X)-\omega_0(1-d,X)|D=1-d,X]=0} \cdot \frac{p_d(X)-p_{d0}(X)}{(1-p_{d0}(X))}\Bigg] \\
&\underbrace{-E\Bigg[\underbrace{\frac{I\{D=d\}}{p_{d0}(M,X)}}_{E[\cdot|M,X]=1} \cdot \frac{(1-p_{d0}(M,X))}{(1-p_{d0}(X))} \cdot \Big[\mu(d,M,X) - \mu_0(d,M,X)\Big]\Bigg] + E\Bigg[\underbrace{\frac{I\{D=1-d\}}{1-p_{d0}(X)}}_{E[\cdot|M,X]=\frac{1-p_{d0}(M,X)}{1-p_{d0}(X)}} \cdot \Big[\mu(d,M,X) - \mu_0(d,M,X)\Big]\Bigg]}_{=0} \\
&\underbrace{-E\Bigg[\underbrace{\frac{I\{D=1-d\}}{1-p_{d0}(X)}}_{E[\cdot|X]=1} \cdot \Big[\omega(1-d,X) - \omega_0(1-d,X)\Big] + \Big[\omega(1-d,X) - \omega_0(1-d,X)\Big]\Bigg]}_{=0}.
\end{aligned}$$

Thus, it follows that:

$$\partial E\left[\psi_d^*(W, \eta^*, \psi_{d0})\right]\left[\eta^* - \eta_0^*\right] = 0$$

proving that the score function is orthogonal.

Assumption 3.1(e): Singular values of $E[\psi_d^a(W; \eta_0^*)]$ are bounded:

This holds trivially, because $\psi_d^a(W; \eta_0^*) = -1$.

Assumption 3.2: Score regularity and quality of nuisance parameter estimators

Bounds for m_n, m_n', r_n, r_n' are omitted for the sake of brevity, because their derivations follow similarly as in the proof for $Y(d, M(1-d))$ in subsection 2.B.1. However, the proof differs in establishing the bound on λ_n' in 3.2(c) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), as it is based on the regularity conditions in Assumption 5 that include $p_d(M, X)$ and $\omega(1-d, X)$.

Bound for λ_n': Consider

$$f(r) := E[\psi(W, \eta_0^* + r(\eta^* - \eta_0^*), \psi_{d0})]$$

We subsequently omit arguments for the sake of brevity and use

$\mu = \mu(d, M, X), \omega = \omega(1-d, X), p_d = p_d(X), p_{dm} = p_d(M, X)$ and similarly $\mu_0, \omega_0, p_{d0}, p_{dm0}$.

For any $r \in (0,1)$:

$$
\begin{aligned}
\frac{\partial^2 f(r)}{\partial r^2} \;=\;& E\left[(-2)\cdot I\{D=d\}\frac{(p_{dm}-p_{dm0})(p_d-p_{d0})\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{dm0}+r(p_{dm}-p_{dm0}))\,(1-p_{d0}+r(p_{d0}-p_d))^2}\right] \\
+\;& E\left[2\cdot I\{D=d\}\frac{(p_{dm}-p_{dm0})^2\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{dm0}+r(p_{dm}-p_{dm0}))^2\,(1-p_{d0}+r(p_{d0}-p_d))}\right] \\
+\;& E\left[2\cdot I\{D=d\}\frac{(1-p_{dm0}+r(p_{dm0}-p_{dm}))\,(p_d-p_{d0})(\mu-\mu_0)}{(p_{dm0}+r(p_{dm}-p_{dm0}))\,(1-p_{d0}+r(p_{d0}-p_d))^2}\right] \\
+\;& E\left[(-2)\cdot I\{D=d\}\frac{(1-p_{dm0}+r(p_{dm0}-p_{dm}))\,(p_{dm}-p_{dm0})\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{dm0}+r(p_{dm}-p_{dm0}))^2\,(1-p_{d0}+r(p_{d0}-p_d))}\right] \\
+\;& E\left[(-2)\cdot I\{D=d\}\frac{(p_d-p_{d0})^2\,(1-p_{dm0}+r(p_{dm0}-p_{dm}))\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{dm0}+r(p_{dm}-p_{dm0}))\,(1-p_{d0}+r(p_{d0}-p_d))^3}\right] \\
+\;& E\left[(-2)\cdot I\{D=d\}\frac{(p_{dm}-p_{dm0})(p_d-p_{d0})\,(1-p_{dm0}+r(p_{dm0}-p_{dm}))\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{dm0}+r(p_{dm}-p_{dm0}))^2\,(1-p_{d0}+r(p_{d0}-p_d))^2}\right] \\
+\;& E\left[(-2)\cdot I\{D=d\}\frac{(p_{dm}-p_{dm0})^2\,(1-p_{dm0}+r(p_{dm0}-p_{dm}))\,(Y-\mu_0-r(\mu-\mu_0))}{(p_{dm0}+r(p_{dm}-p_{dm0}))^3\,(1-p_{d0}+r(p_{d0}-p_d))}\right] \\
+\;& E\left[2\cdot I\{D=d\}\frac{(p_{dm}-p_{dm0})(\mu-\mu_0)}{(p_{dm0}+r(p_{dm}-p_{dm0}))\,(1-p_{d0}+r(p_{d0}-p_d))}\right] \\
+\;& E\left[2\cdot I\{D=1-d\}\frac{(p_d-p_{d0})^2\,(\mu_0-\omega_0)}{(1-p_{d0}+r(p_{d0}-p_d))^3}\right] \qquad (2.18) \\
+\;& E\left[2\cdot I\{D=1-d\}\frac{(p_d-p_{d0})^2 r\,((\mu-\mu_0)-(\omega-\omega_0))}{(1-p_{d0}+r(p_{d0}-p_d))^3}\right] \\
+\;& E\left[2\cdot I\{D=1-d\}\frac{(p_d-p_{d0})\,(\mu-\mu_0)}{(1-p_{d0}+r(p_{d0}-p_d))^2}\right] \qquad (2.19) \\
+\;& E\left[(-2)\cdot I\{D=1-d\}\frac{(p_d-p_{d0})\,(\omega-\omega_0)}{(1-p_{d0}+r(p_{d0}-p_d))^2}\right]
\end{aligned}
$$

Bounding these twelve terms proceeds similarly as in subsection 2.B.1. In order to bound the eighth term, we make use of the sixth inequality in 2.17. Similarly, for bounding the tenth and the twelfth terms we make use of the last inequality in 2.17. Thus, we get that for some C''_ϵ that only depends on C and ϵ

$$\left|\frac{\partial^2 f(r)}{\partial r^2}\right| \leq C''_\epsilon \delta_n n^{-1/2} \leq \delta'_n n^{-1/2}.$$

This provides the upper bound on λ'_n in Assumption 3.2(c) of Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) as long as $C_\epsilon \geq C''_\epsilon$.

This concludes the proof of Theorem 2. □

Chapter 3

Timing Matters: The Impact of Response Measures on COVID-19-Related Hospitalization and Death Rates in Germany and Switzerland

with Martin Huber

Abstract: We assess the impact of the timing of lockdown measures implemented in Germany and Switzerland on cumulative COVID-19-related hospitalization and death rates. Our analysis exploits the fact that the epidemic was more advanced in some regions than in others when certain lockdown measures came into force, based on measuring health outcomes relative to the region-specific start of the epidemic and comparing outcomes across regions with earlier and later start dates. When estimating the effect of the relative timing of measures, we control for regional characteristics and initial epidemic trends by linear regression (Germany and Switzerland), doubly robust estimation (Germany), or synthetic controls (Switzerland). We find for both countries that a relatively later exposure to the measures entails higher cumulative hospitalization and death rates on region-specific days after the outbreak of the epidemic, suggesting that an earlier imposition of measures is more effective than a later one. For Germany, we further evaluate curfews (as introduced in a subset of states) based on cross-regional variation. We do not find any effects of curfews on top of the federally imposed contact restriction that banned groups of more than 2 individuals.

Keywords: COVID-19, pandemic, social distancing, lockdown, treatment effect, synthetic control.

JEL classification: I18, I12, H12.

This article is published in the Swiss Journal of Economics and Statistics as Huber and Langen (2020)
We have benefited from comments by Cristian Carmeli, Arnaud Chiolero, and Reiner Eichenberger. We are grateful to the Swiss Federal Office of Public Health for providing access to their COVID-19 data, to the COVID-19 platform of the Swiss School of Public Health for their data management, and to Andrea Sommer-Gauch for her excellent research assistance.

3.1 Introduction

This paper assesses how the timing of the lockdown measures implemented in Switzerland and Germany affects the development of cumulative COVID-19-related hospitalization and death rates. In both countries, the federal governments implemented extensive lockdown measures, including the closure of non-essential shops, schools, childcare centers, cafes, bars and restaurants. In Germany, these measures were further enhanced with a ban on gatherings with more than two people decided at federal level and curfews implemented in several states. With the measures in place for some weeks, both countries report a flattening of the COVID-19 epidemic curve. This alone, however, does not necessarily exclusively reflect the impact of the measures, but likely also general time trends in the spread of the virus. For this reason, this study aims to provide evidence about the causal effects of the German and Swiss measures by exploiting variation (i) in their relative timing due the fact that the epidemic was more advanced in some regions than in others when certain measures came into force and (ii) across regions due to the fact that some measures were only introduced in a subset of regions.

A range of studies on the impact of COVID-19 response measures focus on predicting the development of the pandemic in terms of infections, hospitalizations, or death rates based on simulating the spread of the virus and calibrating the model as a function of the measures. For instance, Koo, Cook, Park, Sun, Sun, Lim, Tam, and Dickens (2020) provide a simulation study on the COVID-19 outbreak in Singapore and model the development of COVID-19 infections under four potential intervention scenarios. Likewise, Bicher, Rippinger, Urach, Brunmeir, and Popper (2020) developed an agent-based simulation model to predict the development of infections under different scenarios of lockdown timing and exit strategies out of the lockdown in Austria, finding that delaying the lockdown by 1 week would have translated into an increase of infections by 4 times. Donsimoni, Glawion, Plachter, Weiser, and Wälde (2020) simulate the effect of lockdown timing and duration on the rate of COVID-19 infections and the expected end date of the epidemic in Germany. The study suggests that a complete lift of measures on April 20th would have borne the risk of increasing infection rates. The authors further advise to adopt exit strategies and policies that differ across regions in order to learn about which measures are most effective for containing the epidemic while reducing social and economic costs.

In contrast to such simulations, in which empirical data serve for calibrating parameters in prediction models, a growing literature applies policy evaluation methods as outlined in Imbens and Wooldridge (2009) to assess the effectiveness of lockdown measures based on variation across regions and over time. Qiu, Chen, and Shi (2020) for instance investigate the influence of socioeconomic factors and COVID-19 response measures on transmission dynamics in China, finding that measures at a local level have a larger impact on the epidemic curve than restricting population flows between cities. Juranek and Zoutman (2020) use an event study approach to assess the effect of the lockdown measures of Denmark and Norway on hospitalizations based on a comparison with Sweden whose measures are comparably lenient. Results suggest that the peak number of hospitalizations would have more than doubled in Denmark and Norway had they followed Sweden's strategy.

Dave, Friedson, Matsuzawa, and Sabia (2020) use a difference-in-differences approach to evaluate lockdown measures (namely shelter in place orders) in the US by exploiting variation in responses across states and over time. As a consequence of the measures, they find an important increase (of 5 -10%) in the rate at which state residents remained in their homes full-time as well as substantial reductions in cumulative COVID-19 cases (44% after three weeks),[1] with early adopting states with a high population density benefiting most. See also Fowler, Hill, Obradovich, and Levin (2020) for a related difference-in-differences strategy for the US that suggests reductions in infections, too, as well as in fatalities. Results in Friedson, McNichols, Sabia, and Dave (2020), who use a synthetic control approach to analyze the measures' effectiveness in California, and Dave, Friedson, Matsuzawa, Sabia, and Safford (2020), who evaluate the impact of the measures implemented in Texas in an event study framework, point in the same direction. Weber (2020) exploits regional differences in the timing of measures in Germany finding that school closures, prohibition of mass events, as well as gathering bans and curfews played a major role in reducing the number of confirmed infections, while border closures and shut-downs of the service and retail sector did not show a significant effect. Studies on the impact of face mask requirements in public transport, retailers and public businesses find evidence for a reduction in the spread of the virus through such requirements, see e.g. Mitze, Kosfeld, Rode, and Wälde (2020) for a synthetic control study on German data and Chernozhukov, Kasaha, and Schrimpf (2020), who assess the impact of such requirements in the US within a causal framework that allows for both, direct effects of COVID-19 response measures and indirect effects

[1] The estimated effect on fatalities is also negative but less precise.

through behavioral changes.

Askitas, Tatsiramos, and Verheyden (2020) apply an event study design to assess a range of different response measures across 135 countries and find that canceling public events and restricting gatherings reduce new infections more effectively than mobility restrictions like international travel controls. This is in line with Bonardi, Gallea, Kalanoski, and Lalive (2020) who consider first difference and AR(1) models based on 184 countries and conclude that lockdown measures generally reduce confirmed infections and fatalities (and even more so if imposed rather earlier than later), while border closures do not show important effects. Findings in Banholzer, van Weenen, Kratzwald, Seeliger, Tschernutter, Bottrighi, Cenedese, Salles, Feuerriegel, and Vach (2020), a study on 20 Western countries in a Bayesian framework, suggest that venue closures and gathering bans are most effective in reducing infections but also attest a significant effect of border closures.

Our paper contributes to this growing literature by analyzing COVID-19-related hospitalizations and death rates across administrative units over time, namely across counties in the case of Germany and across cantons in the case of Switzerland. We estimate the effect of the relative timing of lockdown measures based on measuring health outcomes relative to the region-specific start of the epidemic and comparing outcomes across regions with earlier and later start dates. The start date is defined as the day on which the confirmed regional infections per 10,000 inhabitants exceed 1 for the first time. In the analysis, we control for regional characteristics (population size and density, age structure, and GDP per capita), initial trends of the epidemic (median age of confirmed infections and initial growth rate of confirmed infections), and other policies selectively introduced prior to the major lockdowns (e.g. a ban on visits to hospitals and retirement homes in some regions).

Linear regression estimates suggest that for both Switzerland (which also includes the Principality of Liechtenstein as data point) and Germany, a relatively later exposure to the measures entails higher cumulative hospitalization and death rates on sufficiently advanced region-specific days after the outbreak of the epidemic. This suggests that an earlier imposition of measures is more effective than a later one w.r.t. our health outcomes, which is in line with findings in Amuedo-Dorantes, Borra, Garrido, and Sevilla (2020) on the effect of lockdown timing on COVID-19-related deaths in Spain. For Germany with its substantially larger number of observations, we also estimate the effect of the relative timing based on doubly robust (DR) estimation, see Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995), which is a more flexible approach than exclusively relying on a linear outcome model. For Switzerland, we also consider the synthetic control method, see Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010), to assess for two selected cantons with a relatively late exposure what their counterfactual outcomes would have been under an earlier exposure. Both the DR and synthetic control methods corroborate the findings of the linear regression. For Germany only, we also evaluate the effect of curfews that were introduced by a subset of German states in addition to the federal lockdown measures and bans of gatherings with more than two individuals. Exploiting this cross-sectional variation while controlling for observed characteristics, neither linear regression nor DR estimation suggest that curfews further reduce hospitalizations and fatalities under the lockdown measures already in place, which is in line with the findings in Bonardi, Gallea, Kalanoski, and Lalive (2020) and Banholzer, van Weenen, Kratzwald, Seeliger, Tschernutter, Bottrighi, Cenedese, Salles, Feuerriegel, and Vach (2020). Apart from this assessment of the impact of curfews on COVID-19-related death rates, our analysis does not inform about the effectiveness of single social distancing measures implemented as part of the lockdown in Germany and Switzerland. Further, a cost-effectiveness assessment of COVID-19 response measures, which is certainly of great importance for policy makers, is beyond the scope of this paper, as the long-run social and economic consequences of the lockdown cannot be credibly assessed at the current stage.

The remainder of this paper is organized as follows. Section 3.2 provides an overview of the timeline of COVID-19 measures in Switzerland and Germany. Sections 3.3 and 3.4 describe the data and econometric methods used in the analyses. Section 3.5 presents and interprets the results. Section 3.6 concludes.

3.2 Timeline of COVID-19 Response Measures

Both Germany and Switzerland are federal states with competencies in epidemic control partly belonging to the 26 cantons in Switzerland and the 16 federal states (Länder) in Germany. The German states themselves are comprised of all in all 401 counties (Kreise) which also have certain competencies in handling epidemic outbreaks. With competencies fragmented across the federal governments and sub-federal authorities, not all measures were implemented in all regions and, if so, not always at the same time. However, decisions on key COVID-19 response measures were made at the federal level in both countries.

In Switzerland, the first COVID-19 response measure, a ban of events with more than 1000 visitors, was announced and implemented at the federal level on February 28th when there were some 25 confirmed COVID-19 cases (0.03 per 10,000 inhabitants) in Switzerland. Several measures at the cantonal level followed. For instance, many cantons introduced a ban on visits to retirement homes. Some 2.5 weeks after the first measure was implemented, the Federal Council decided to close all schools and childcare centers in Switzerland as well as non-essential shops, cafes, bars, and restaurants on March 16th. In the following, we will refer to these measures as lockdown measures. At that point in time, the rate of confirmed infections in Switzerland was at 4.2 per 10,000 inhabitants. The schedule of response measures in the Principality of Liechtenstein (LI) was similar to that in Switzerland with the lockdown entering into force two days later. Due to the two countries' similar schedules of COVID-19 response measures, their geographic proximity and their economic, cultural and political interconnection, we include LI as additional data point when investigating the impact of the lockdown measures in Switzerland.

In Germany, first measures at the federal level were implemented between March 9th and March 12th. On March 8th, when there were some 1000 reported COVID-19 cases (0.12 per 10,000 inhabitants) in Germany, the federal government advised against events with more than 1000 visitors. This recommendation was translated into a ban by most federal states, while others implemented it as recommendation only. As in Switzerland, schools and childcare centers in most German states closed on March 16th, the remaining states followed within two days. The closure of all non-essential retailers, bars and public events of any kind and the restriction of restaurant opening hours was decided at the federal level on March 16th when the overall rate of confirmed infections reached 1.1 per 10,000 inhabitants. The states implemented these measures between March 17th and March 20th. Other than in Switzerland and LI, these measures were further enhanced later on. On March 22nd, a ban of groups with more than two individuals was decided at the federal level and several states additionally implemented curfews. Since April 17th, more and more states have made wearing face masks in shops and public transport compulsory, resulting in a nationwide requirement to wear masks in public from April 27th on. Meanwhile, lockdown measures have been lifted gradually in Switzerland and Germany, with distinct schedules and exit strategies across countries and states. For instance, curfews ended in the respective German states around April 27th, with the exception of Bavaria, where they ended on May 5th. On May 6th, a so-called "emergency mechanism" was put in place in Germany requiring counties to re-impose lockdown measures locally if the rate of new confirmed infections over 7 days exceeds 5 per 10,000 inhabitants.

3.3 Data

For Switzerland and LI, data on confirmed COVID-19 infections as well as on COVID-19-related hospitalizations and deaths are amalgamated by the Swiss Federal Office of Public Health (FOPH) and made available to the interuniversity research consortium of the Swiss School of Public Health (www.ssphplus.ch). For each confirmed case, the FOPH gathers information on the reporting canton, test date, as well as patient's age and gender from laboratory declarations. For our analysis, we aggregate the number of confirmed infections, hospitalizations and fatalities by canton and test date, compute the respective cumulative numbers by canton and date, and complement the data with socio-demographic variables at the cantonal level (and for LI) from the statistical offices of Switzerland and LI. For each of the 26 Swiss cantons and LI, we calculate the rate of cumulative confirmed infections, hospitalizations and fatalities per 10,000 inhabitants, as well as the median age of those tested positively for COVID-19 prior to the lockdown measures in Switzerland and LI. Furthermore, we construct indicators for whether a canton has introduced certain additional measures not imposed by the federal government along with variables providing the start date of such canton-level measures as stated in press releases of the respective cantons.

In Germany, all confirmed infections and deaths are reported to the Robert Koch Institute (RKI), a federal government agency and research institute for disease control and prevention. The RKI publishes data on the age group, gender, test date and county of residence of each validated COVID-19 case reported to the institute. Only for the county of Berlin with 3.6 million inhabitants, the RKI also reports the urban residential district of confirmed cases. All in all, there are 401 counties in Germany and 12 residential districts in Berlin. Similar to Switzerland, we aggregate the data by county (or residential district, respectively) and test date, and compute cumulative confirmed cases and fatalities by county and date. We complement the data with socio-demographic variables at the county/district level from the Federal Office of Statistics, the statistical offices of the federal states and the statistical office of the city of Berlin. As most measures in Germany were implemented at the state or even county level and at different points in time, we generate variables for all measures indicating whether and when they were imposed in each county.

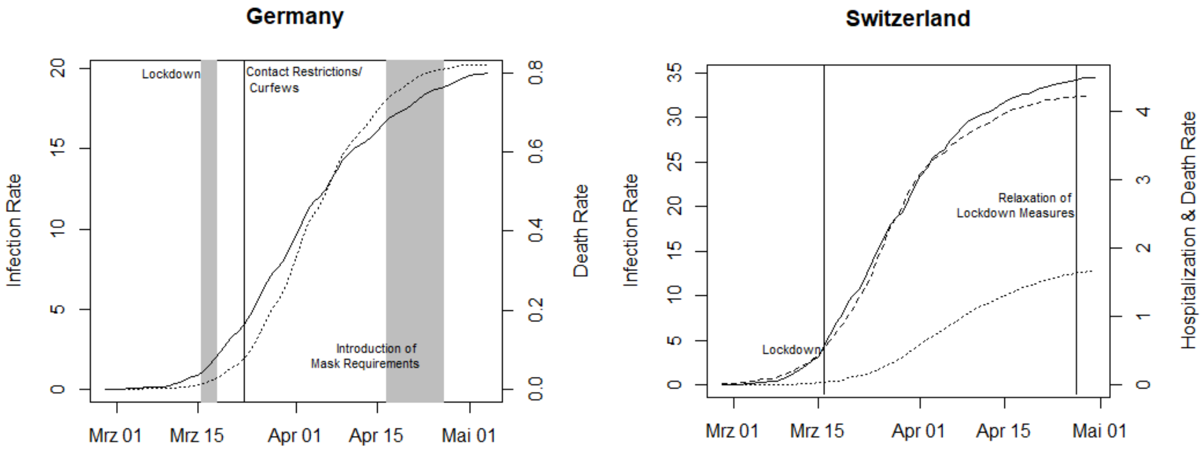


Figure 3.3.1: Cumulative confirmed infections (solid line), deaths (dotted line) and hospitalizations (dashed line) per 10,000 inhabitants in Germany and Switzerland.

Figure 3.3.1 provides the cumulative numbers of confirmed COVID-19 infections and COVID-19-related deaths per 10,000 inhabitants in Germany (left) as well as cumulative numbers of confirmed infections, hospitalizations and deaths in Switzerland (right). The figure suggests a flattening of the COVID-19 epidemic curve in both countries after the main COVID-19 measures have been in place for some weeks, which does, however, not necessarily exclusively reflect the causal impact of the measures. As a further descriptive statistic, Figure 3.3.2 provides the overall deaths per 10,000 inhabitants (thus including COVID-19-related mortality) by calendar week in Germany and Switzerland since January 1st 2020 (provisional data). While the increase in mortality in March and April can be linked to the COVID-19 epidemic (a finding that also holds when controlling for the average mortality over 2015-2019), we cannot directly infer how large the increase would have been with and without the lockdown measures. For this reason, our analysis aims at shedding light on the causal effect of the measures.

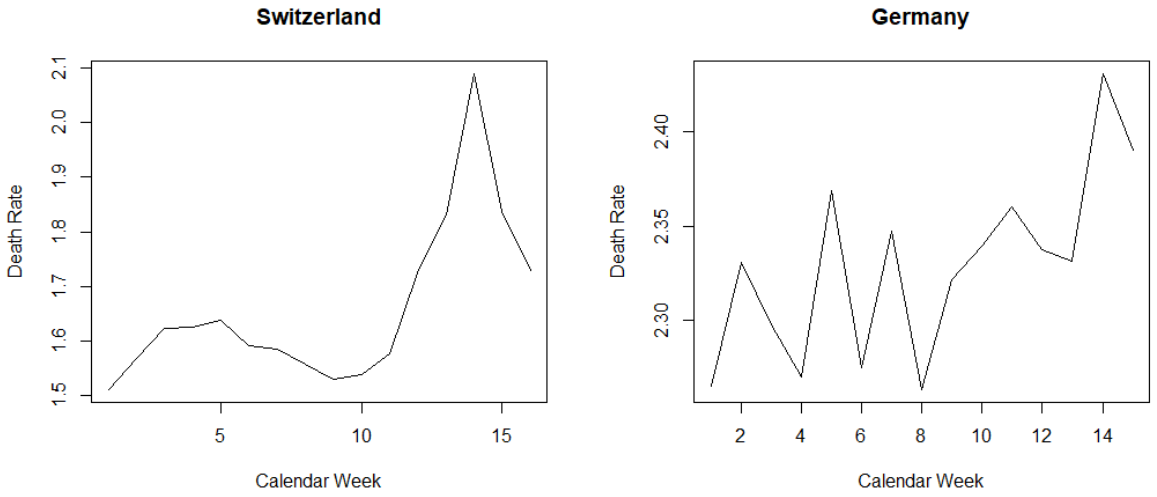


Figure 3.3.2: Overall deaths per 10,000 inhabitants by calendar week in Switzerland (left) and Germany (right). Source: federal statistical offices of Switzerland (www.bfs.admin.ch) and Germany (www.destatis.de), retrieval date: May 6th.

3.4 Econometric Approach

In our analysis, we exploit the fact that the epidemic was more advanced in some regions than in others when the key control measures came into force. In Switzerland, for instance, Basel-Stadt had already more than 1 confirmed case per 10,000 inhabitants 12 days before the federal lockdown measures were implemented, while other cantons such as St. Gallen were at an earlier stage, reaching 1 confirmed infection per 10,000 inhabitants on the day of the lockdown. In Germany, the county of Heinsberg recorded more than 1 confirmed infection per 10,000 inhabitants already 19 days before the lockdown. In several other counties this level of infections was reached only after the lockdown.

For Germany, we investigate the impact of the lockdown measures as well as the curfew on cumulative deaths per 10,000 inhabitants. For Switzerland and LI, we assess the causal effect of the lockdown on both cumulative hospitalizations and deaths per 10,000 inhabitants. The idea is to quantify the epidemic stage of each canton/county when measures were implemented by defining dates on which the health outcomes are measured relative to the day a canton/county first reached a certain rate of confirmed infections. For both Germany and Switzerland, we define the start date of the epidemic as the day when the rate of infections first reached or exceeded 1 infection per 10,000 inhabitants. In Switzerland, for instance, the start date of the epidemic in Basel-Stadt is on March 5th (late exposure to measures) while in St. Gallen the epidemic started on March 16th (early exposure to measures). Appendix 3.A provides the start states for all Swiss cantons and LI.

Besides their obvious relevance for health care, a further motivation to consider hospitalization and death rates as outcomes is that their measurement is likely more robust to differences in testing strategies across regions than the measurement of confirmed COVID-19 infections. While the share of infections with mild symptoms being detected ceteris paribus likely rises with increased testing, the number of hospitalizations and fatalities gives a better estimate of the severeness of the epidemic in terms of human loss and strains for the health care system. As both Germany and Switzerland maintain a system of mandatory health insurance and neither country generally saw their hospitalization capacities exhausted, we would suspect that the number of COVID-19-related hospitalizations in general mirrors well the number of individuals infected with COVID-19 that are in need of hospitalization. Nevertheless, a potential concern in our analysis is that the criteria for hospitalizations might not be uniform across regions. The same may apply to the measurement of fatalities, i.e. the definition of criteria according to which a decease is attributed to COVID-19. If such measurement issues in health outcomes are not systematically associated with the region-specific start date of the epidemic (or more generally, with the policy interventions considered), they do not bias the results of our analysis. However, if for instance regions with an earlier start date and a more advanced epidemic systematically applied more stringent rules for hospital admissions (e.g. to prevent capacity constraints), this could also entail an underestimation of COVID-19 fatalities due to underreporting deceases at home. In this case, our analysis of the relative timing of measures presented below would likely provide a lower bound of the true effect on (capacity-unconstraint) hospitalizations and fatalities.

3.4.1 OLS Approach

We compare the average number of cumulative hospitalizations and fatalities per 10,000 inhabitants on canton/county-specific epidemic days across three groups of cantons/counties. These groups are defined by the canton/county-specific epidemic day when lockdown measures came into place. For Switzerland and LI, we distinguish the groups of cantons as follows. Cantons reaching or exceeding 1 confirmed infection per 10,000 inhabitants at most 4 days before the lockdown measures are exposed to the measures at a relatively early stage of the epidemic and constitute the reference group (sample size $N = 8$). Those cantons with at least 1 confirmed infection per 10,000 inhabitants between 5 and 8 days before March 16th (or March 18th in the case of LI) are the intermediate intervention group ($N = 11$). Those with a canton-specific start date at least 9 days before March 16th are the late intervention group ($N = 8$).

For Germany, we proceed analogously and define the treatment groups based on the days between the county-specific start of the epidemic and the lockdown according to the retail closures between March 17th and 20th, but with somewhat different time brackets. Counties with at least 1 confirmed infection per 10,000 inhabitants not earlier than 3 days after the implementation of lockdown measures make up the reference group. The specified start dates are later than the lockdown, which may at first glance raise endogeneity concerns. However, any effect of the measures can materialize in the outcomes only with a substantial time lag of more than 1.5 weeks (due to incubation time and reporting lags), as also confirmed in our analysis. Therefore, confirmed infection rates are not yet influenced by the measures even several days after the lockdown. Yet, we exclude 4 counties having start dates as late as 9 days after the lockdown or later, leaving us with a reference group of $N = 52$. The intermediate intervention group is comprised of all counties with at least 1 confirmed infection per 10,000 inhabitants between 3 days before and 2 days after the lockdown ($N = 275$). The late intervention group consists of counties with at least 1 confirmed infection per 10,000 inhabitants more than 3 days before the lockdown ($N = 81$).

We estimate the difference in cumulative death rates, as well as hospitalization rates for Switzerland and LI, between either of the two treatment groups (intermediate and late intervention group) and the reference group by means of an OLS regression with treatment indicators. We also control for the following canton-/county-specific covariates: population size and density, income per capita, age distribution, age structure of positively tested up to the lockdown, the initial canton-/county-specific growth trend for confirmed cases, and canton-specific bans on visits in hospitals and retired homes entering into force prior to the lockdown. The large number of counties in Germany allows us to further control for past mortality by age group, past mortality rate related to respiratory diseases and hospital capacities (beds/1000 inhabitants). We also control for state-specific measures entering into force prior to the general lockdown, like bans of or recommendations against events with more than 1000 visitors, as well as curfews imposed in some states only a few days after the general lockdown. Appendix 3.B provides descriptive statistics of the covariates used in the analysis of the German and Swiss measures for the respective total samples as well as separately for the various intervention groups.

Though aiming to control for confounders jointly affecting the region-specific epidemic and the health outcomes in a comprehensive way, we cannot completely rule out that some important characteristics are omitted in our analysis. For instance, we cannot directly control for the amount of inter-generational interactions, which is according to Bayer and Kuhn (2020) correlated with the ratio of deaths over confirmed cases and could potentially differ across regions. We, however, point out that the results for the relative timing of measures are quite robust to (not) controlling for covariates. Since the lockdown measures in Germany and in Switzerland have been eased starting with April 20th and April 27th, respectively, we evaluate the effect of the relative timing of measures on the health outcomes in these countries until April 23rd and April 30th, respectively.

For Germany, we also investigate the impact of curfews, as introduced in some federal states between March 21st and 23rd on top of the federally imposed contact restriction that banned groups of more than 2 individuals. The OLS regression contains a binary treatment indicator for curfews as well as a range of control variables. The latter include the previously mentioned county-specific characteristics, growth trends and COVID-19 response measures, and in addition the cumulative confirmed infections and death rates on several days prior to the curfews, in order to make regions exposed and not exposed to curfews as similar as possible. The OLS specification is provided in Appendix 3.C, descriptive statistics for counties with and without curfews in Appendix 3.B.

3.4.2 Doubly Robust Estimation

The larger number of regions in Germany allows us to also consider a more flexible (so-called semiparametric) evaluation approach based on doubly robust (DR) estimation, see Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995). It is based on (i) estimating a logit model for the treatment probability as a function of the covariates as well as a linear model for the outcome as a function of the treatment and the covariates and (ii) using the respective model predictions as plug-in parameters for the estimation of the treatment effects. DR provides consistent effect estimates if at least one of the plug-in models is correctly specified and thus relies on less stringent assumptions than OLS. Using the 'drgee' package of Zetterqvist and Sjölander (2015) for the statistical software 'R', we apply DR for estimating the average effect of a binary intervention separately to subsets of counties consisting of the reference group and either the intermediate intervention group or the late intervention group.

3.4.3 Synthetic Control Approach

For Switzerland, we complement the regression analysis with a synthetic control approach, a quantitative case study method suggested in Abadie and Gardeazabal (2003). To this end, we compare cumulative hospitalization and fatality rates in a specific canton with a late exposure to the lockdown to the rates of an artificially (or synthetically) created counterfactual canton. This synthetic canton should be comparable to the original reference canton in terms of covariates outlined in Section 3.4.1 and pre-treatment health outcomes (measured 2 and 5 days after the start date), but characterized by an earlier exposure to the lockdown.[2] To this end, the synthetic canton is generated as a weighted average of control cantons with an earlier exposure using the 'Synth' package of Abadie, Diamond, and Hainmueller (2011) for the statistical software 'R', where the weights depend on how close their characteristics and pre-treatment outcomes match the values of the reference canton with the later exposure. The control pool includes all in all 11 cantons that reached 1 confirmed infection per 10,000 inhabitants at most 3 days before the lockdown.

3.5 Results

3.5.1 Germany

Figure 3.5.1 reports the mean differences in cumulative fatalities per 10,000 inhabitants between either treatment group and the early intervention group (reference group) per day up to 28 days after the county-specific start date (solid lines) based on the OLS approach.[3] It also includes 90% confidence intervals (dashed lines). The mean differences in fatality rates between the late and the early intervention groups (left) remain close to zero during the first 2.5 weeks of the county-

[2] In contrast to the OLS specification provided in Appendix 3.C, squared variables (i.e. the squares of the population share aged 65+ and of the median age of confirmed infections prior to the lockdown) are not included. In addition, the dummy for the number of inhabitants being smaller than 60,000 is replaced by the actual number of inhabitants.

[3] The motivation for the 28 days window is that we would like to include all (but 4) counties while at the same time only considering the period when the lockdown measures were fully implemented. As the last county we include in our evaluation sample saw its start of the epidemic 8 days after the lockdown, the time range considered in the analysis is limited to this specific window not including any effects of the first easing of lockdown measures starting with April 20th.

specific epidemic but show a positive and statistically significant tendency thereafter. The point estimates suggest that after one month, fatalities per 10,000 inhabitants are reduced by 0.6 cases under an earlier lockdown. Also the difference in death rates between the intermediate and the early intervention groups are statistically significant at the 10 percent level, but (expectedly) smaller in magnitude. Overall, the results suggest that the relative timing of measures had a perceptible impact on COVID19-related fatalities in Germany. We note that Appendix 3.C provides the OLS specification with the full list of coefficients on treatments and covariates along with standard errors 28 days after the start of the epidemic. Concerning the robustness of our findings, we note that estimations without controlling for observed covariates yield qualitatively similar results, see Appendix 3.D.

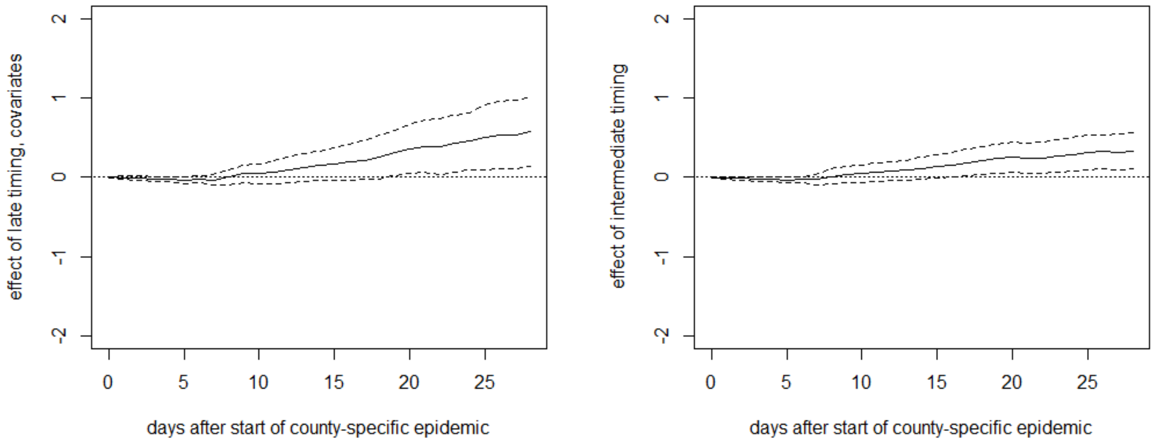


Figure 3.5.1: OLS effects of late (left) and intermediate (right) timing of measures on cumulative deaths per 10,000 inhabitants in Germany.

Figure 3.5.2 reports the estimates of DR, which are generally similar to OLS, though suggesting an even stronger effect of a late timing of lockdown measures on the death rate. The point estimate suggests that an earlier lockdown reduces fatalities by roughly 1 case per 10,000 one month after the start of the epidemic.

With 27% of the German population living in counties with late lockdown timing, a rough back-of-the-envelope calculation based on the OLS point estimates suggests that some 1283 COVID-19-related deaths (2080 when using the DR results) could have been prevented in Germany over the first four weeks after lockdown implementation if the counties with late timing had implemented the lockdown early, meaning no later than 3 days before reaching or exceeding the level of 1 confirmed infection per 10,000 inhabitants. If all 275 states with intermediate lockdown timing had implemented the lockdown early, the death toll could have been further reduced by some 1816 (1580 based on DR results).

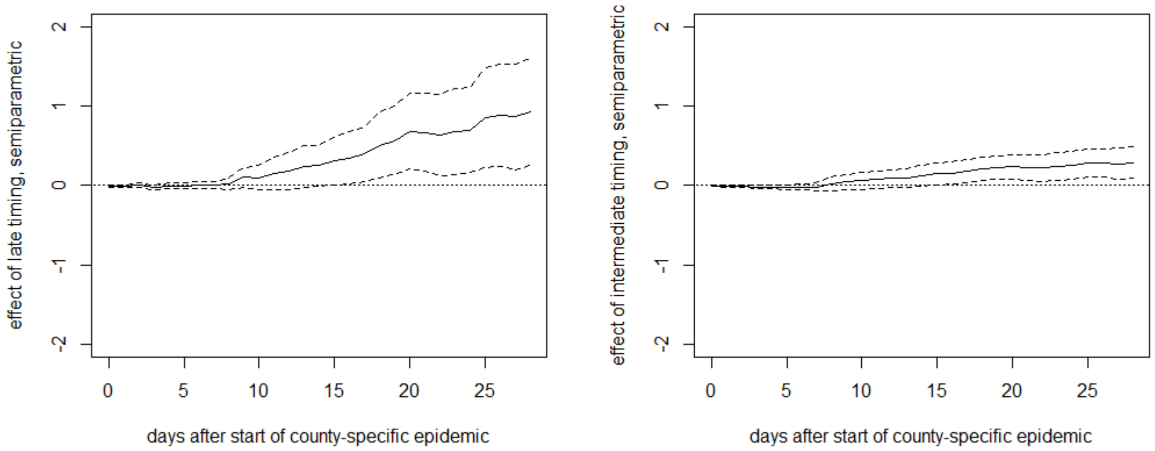


Figure 3.5.2: DR effects of late (left) and intermediate (right) timing of measures on cumulative deaths per 10,000 inhabitants in Germany.

Figure 3.5.3 reports the results of a further OLS regression, in which the treatment indicators for the intermediate and late intervention groups are replaced by the time lag between the county-specific start date of the epidemic and the lockdown, in order to (linearly) estimate the effect of the lag. This can be interpreted as the average effect of waiting an additional day before implementing the measures. The point estimates suggest that each additional day without lockdown entails on average 0.04 to 0.05 additional fatalities per 10,000 inhabitants after one month of the epidemic, even though the confidence intervals are rather wide (but yet do not include a zero effect). Again, these results are quite robust to not controlling for covariates, see Appendix 3.D.

Our results also appear interesting with respect to one key element in the German exit strategy, the so-called "emergency mechanism" requiring counties to re-impose lockdown measures locally if the rate of new confirmed infections over 7 days exceeds 5 per 10,000 inhabitants. Though the local epidemic start date is based on the cumulative rate of confirmed infections and the threshold of the German policy is based on the 7-day running infection rate, one may want to assess the appropriateness of this threshold in the light of our findings about the importance of lockdown timing. In fact, the threshold for re-implementing lockdown measures can be regarded as late rather than intermediate or early intervention with respect to our definition, which seems worth considering given the threat of a second wave. However, the situation during the early phase of the epidemic is most likely not comparable to that in a later point in time, where the hope is that larger testing capacities and better policy response lead to an earlier detection and containment of local COVID-19 outbreaks and that the increased awareness in the population entails an adoption of social distancing and hygiene measures that sufficiently slow down the transmission.

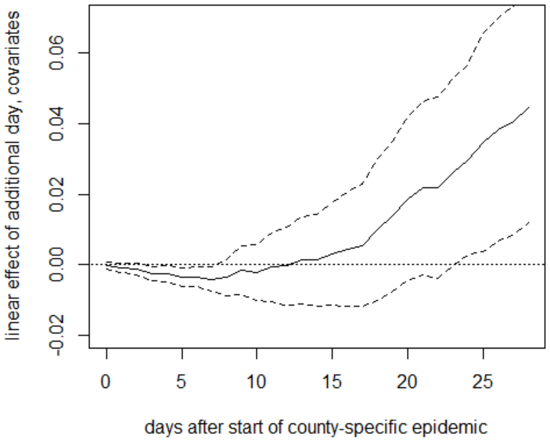


Figure 3.5.3: OLS effect of delaying lockdown by one day on deaths per 10,000 inhabitants in Germany.

Furthermore, the left graph in Figure 3.5.4 provides the OLS-based effects of curfews relative to contact restrictions, i.e. bans of gatherings with more than 2 persons, under all other lockdown measures already in place. The estimates have a positive sign, which appears counter-intuitive as curfews are more restrictive than contact restrictions, but are never statistically significantly different from zero throughout the evaluation window which starts on March 23rd and ends 35 days later. The same finding applies to estimation results based on DR, which are shown in the right graph of Figure 3.5.4. Therefore, we do not find evidence that curfews are more effective than banning groups for reducing fatality rates.

3.5.2 Switzerland and Liechtenstein

Figure 3.5.5 reports the OLS estimates of the mean differences in cumulative hospitalizations (left) and fatalities (right) per 10,000 inhabitants between the late and the early intervention groups up to 44 days after the start of the canton-specific epidemic (solid line), as well as 90%

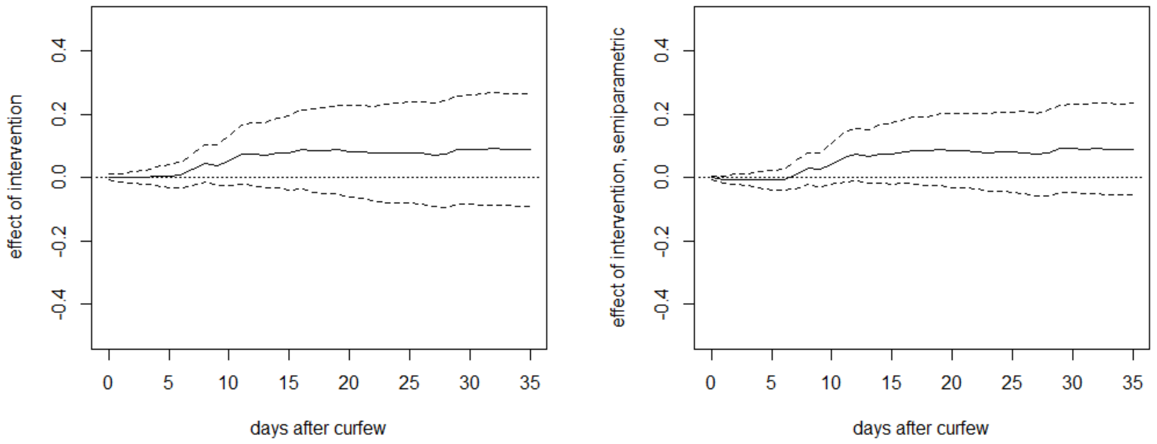


Figure 3.5.4: OLS (left) and DR (right) effects of curfews on deaths per 10,000 inhabitants in Germany.

confidence intervals (dashed lines). See Appendix 3.C for the full OLS specification with the coefficients on treatments and covariates on the last day of the evaluation window and fatalities as outcome variable.

We note that the canton of Ticino is excluded from this analysis due to its comparably strong economic and social ties with Northern Italy (which was particularly severely affected by the COVID19 crisis), as this could arguably have affected the canton's hospitalizations and fatalities. However, our findings are quite similar when including Ticino in the regression, as well as when not controlling for covariates, see Appendix 3.E.

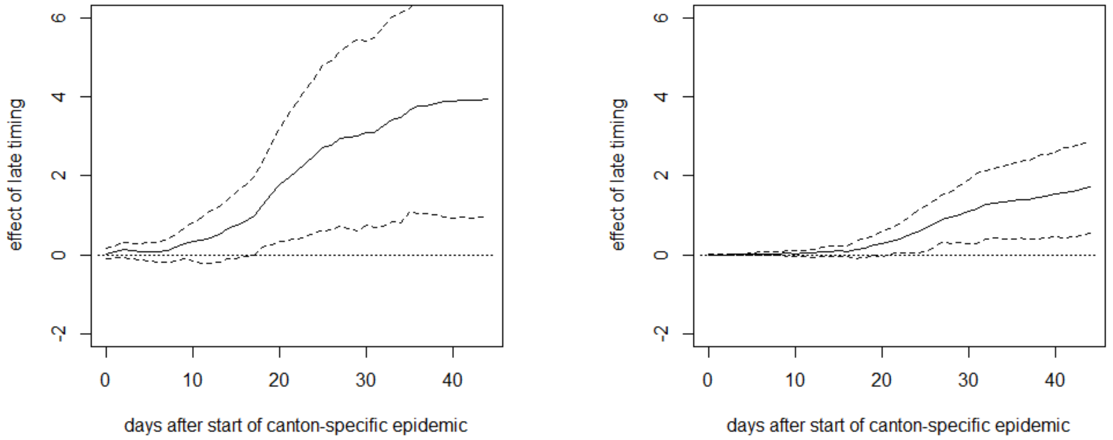


Figure 3.5.5: Effect of late timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants in Switzerland.

As for Germany, we see no immediate effect of the relative timing of measures on the health outcomes right after their introduction. However, after about two weeks, there is a positive tendency in the effect on cumulative hospitalizations that becomes statistically significant at the 10% level about 2.5 weeks after the start of the canton-specific epidemic. The point estimates suggest that after 1.5 months, cumulative hospitalizations per 10,000 inhabitants increase by almost 4 cases when introducing the measures later rather than earlier, even though the estimates are not very precise (i.e. confidence intervals are wide). A qualitatively similar pattern is observed for the effect on cumulative deaths, which becomes statistically significant after about 3 weeks. The point estimates suggest an increase of 1 to 2 fatalities per 10,000 inhabitants in the case of a later lockdown, but precision is again low. Figure 3.5.6 reports the same analysis for a comparison of the groups with intermediate and early timing. As these two groups are more similar in terms of the relative timing of the measures, differences are less pronounced and never statistically significant in all but one case, which might be due to low statistical power related to the small number of cantons.[4]

A rough back-of-the-envelope estimation based on these point estimates suggests that some 333 COVID-19-related deaths and some 764 hospitalizations could have been prevented during the time of the lockdown in Switzerland if the cantons with late timing had implemented the lockdown at most 4 days after reaching or exceeding the level of 1 confirmed infection per 10,000 inhabitants.

Finally, we report the results of the synthetic control method for two cantons experiencing the lockdown rather late relative to their start date of the epidemic. Figure 3.5.7 plots the difference in cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants on a daily base after the canton-specific start date between Basel-Stadt, which was on day 12 of the epidemic when the measures came into force, and its synthetic counterfactual. The latter is generated from a control group of 11 cantons with an earlier timing (with start dates between 3 days before and 1 day after the lockdown). Dots on the solid line imply that the differences are statistically significant at the 10% level according to placebo tests in the control group, in which each of the 11 cantons is considered as (pseudo-)treated in a rotating scheme in order

[4] For cumulative fatalities, we also run the OLS regression using an alternative data source based on calculations of the statistics office of the canton of Zurich, available at https://statistik.zh.ch (retrieved on May 15th). We obtain a comparable pattern. Namely, the late intervention effect turns statistically significant after about 3 weeks with even somewhat higher point estimates (approaching 3) at the end of the evaluation window. The intermediate intervention effect is again insignificant.

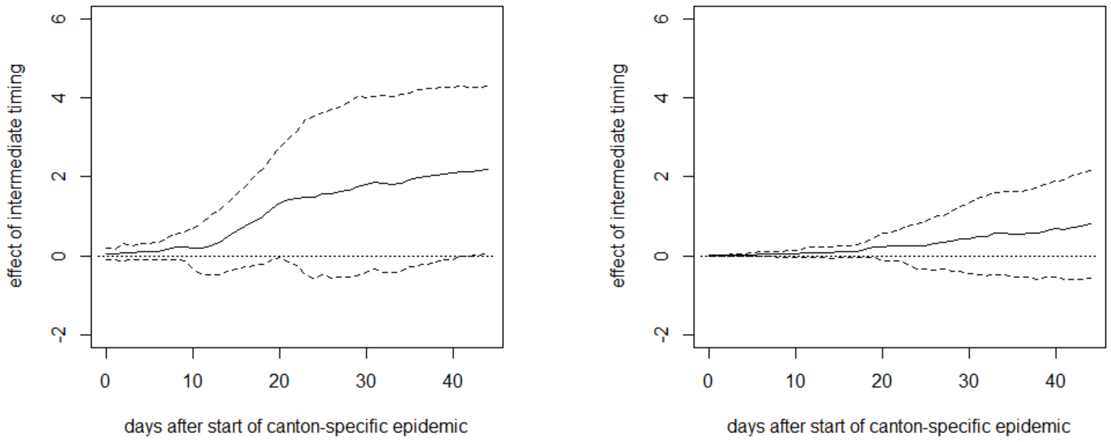


Figure 3.5.6: Effect of intermediate timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants.

to estimate its (pseudo-)counterfactual based on the remaining 10 cantons. We, however, note that the estimation of p-values might be imprecise, due to the low number of control cantons available for the placebo tests.

Again, the relative timing of measures shows no immediate effect on hospitalizations but the difference becomes statistically significant after roughly 2.5 weeks. The point estimates suggest that the hospitalization rate in Basel-Stadt could have been reduced by more than 4 hospitalizations if the lockdown measures had been introduced earlier. Similarily, the fatalities per 10,000 inhabitants could have been reduced by 1 to 2 cases about 1.5 months after the start of the epidemic. As for the OLS analysis, the exact numbers should, however, be interpreted with caution, as they are imprecisely estimated and canton-specific factors not considered in the analysis could play a role as well.

Figure 3.5.8 reports the results for Neuchâtel, another canton with a relatively late timing, which was on day 10 of the epidemic when the measures came into force. Concerning the effect of the lockdown timing on hospitalizations, we find a similar pattern as for Basel-Stadt. Albeit the effect on COVID-19-related fatalities is somewhat less pronounced, it turns statistically significant in the final periods of the evaluation window.

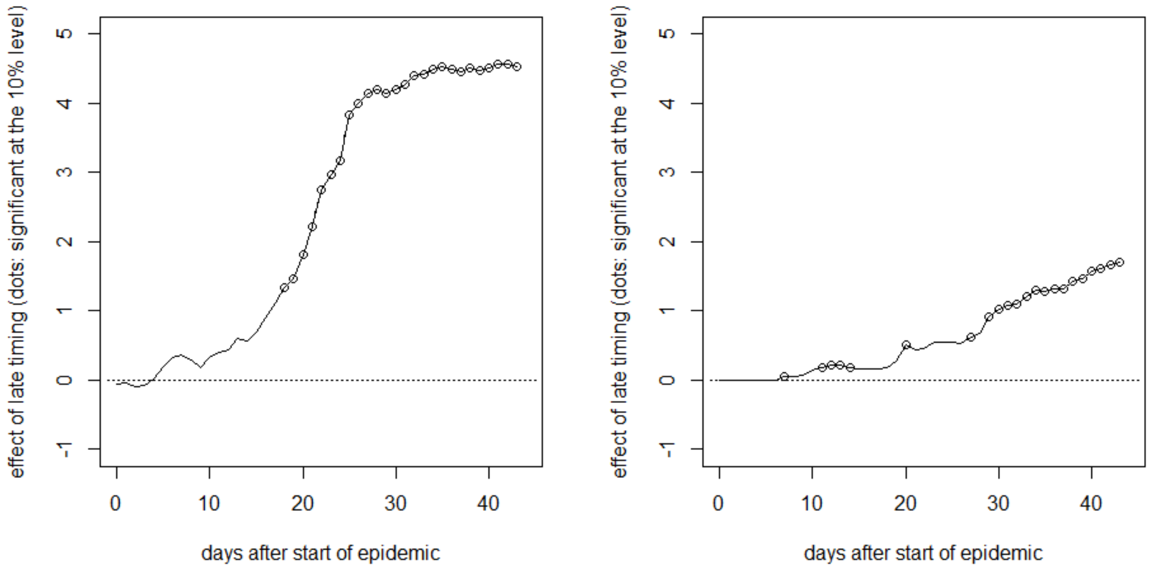


Figure 3.5.7: Effect of late timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants in Basel-Stadt.

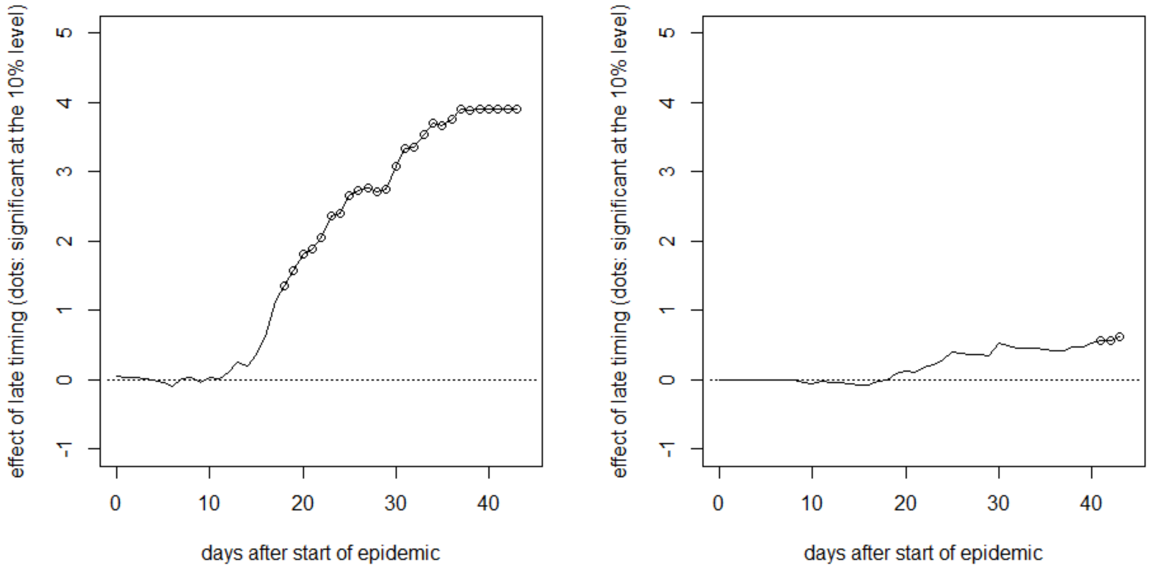


Figure 3.5.8: Effect of late timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants in Neuchâtel.

3.6 Conclusion

In this paper, we analyzed the impact of lockdown timing on COVID-19 related fatalities and hospitalizations in Germany and Switzerland. For doing so, we exploited the fact that measures differed across regions and that the epidemic was more advanced in some regions than in others when certain measures came into force. Using OLS and doubly robust estimation, we compared the development of COVID-19-related hospitalization and death rates - two indicators which are arguably rather robust to regional differences in COVID-19 testing policies - across regions that have been at different epidemic stages when exposed to the lockdown measures. For Switzerland, we also applied a synthetic control approach to investigate the impact of the relative timing of the lockdown in two selected cantons. In addition, we analyzed the impact of curfews as implemented in some German states on top of the federal ban on gatherings of more than 2 persons based on a cross-regional comparison.

For both countries, we found an earlier lockdown to be more effective than a later one, as cumulative hospitalization and fatality rates measured relative to the region-specific start date of the epidemic were higher in regions with a more advanced spread of COVID-19 when the measures came into force. In contrast, our results did not provide evidence for curfews being more effective than bans on gatherings under the other lockdown measures already in place.

Appendix

3.A Start Dates of Canton-Specific Epidemics

Canton	Start Date
Aargau (AG)	03/16
Appenzell Innerrhoden (AI)	03/13
Appenzell Ausserrhoden (AR)	03/13
Bern (BE)	03/14
Basel-Landschaft (BL)	03/11
Basel-Stadt (BS)	03/05
Fribourg (FR)	03/11
Genève (GE)	03/09
Glarus (GL)	03/12
Graubünden (GR)	03/09
Jura (JU)	03/10
Luzern (LU)	03/16
Neuchâtel (NE)	03/07
Nidwalden (NW)	03/09
Obwalden (OW)	03/11
St. Gallen (SG)	03/16
Schaffhausen (SH)	03/17
Solothurn (SO)	03/16
Schwyz (SZ)	03/12
Thurgau (TG)	03/16
Ticino (TI)	03/05
Uri (UR)	03/17
Vaud (VD)	03/09
Valais (VS)	03/12
Zug (ZG)	03/13
Zürich (ZH)	03/12
Principality of Liechtenstein (LI)	03/09

Table 3.A.1: 2020 dates on which 1 confirmed infection per 10,000 inhabitants was reached in the Swiss cantons and LI.

3.B Descriptive Statistics of Covariates

Variable	Total Sample	Late Timing	Intermediate Timing	Early Timing	Curfew	No Curfew
	N = 408	N = 81	N = 275	N = 52	N = 149	N = 259
Population	203,103	276,529	197,295	119,444	158,786	228,598
Population Density	671	929	665	301	440	804
Income per Capita (Euro)	37,224	41,686	36,505	34,076	38,325	36,591
Share of Population Aged 65+	0.222	0.208	0.221	0.244	0.226	0.219
80+ Mortality Rate (per 1000 Inhabitants), 2017	6.52	5.96	6.52	7.36	6.68	6.42
Share of Respiratory-Disease-Related Deaths, 2016	0.07	0.069	0.071	0.067	0.066	0.072
Hospital Beds per 1000 Inhabitants	6.31	6.08	6.25	6.97	6.69	6.09
Share of Confirmed Infections Aged 80+ prior to Lockdown	0.019	0.024	0.018	0.014	0.022	0.017
Initial Growth Trend for Confirmed Cases in Log Points	0.209	0.23	0.234	0.049	0.185	0.224
Ban of events with >1000 Participants	0.917	0.889	0.924	0.923	1	0.869
Curfew	0.365	0.247	0.378	0.481	1	0
Ban of Groups of >5 Persons (prior to Contact Ban/Curfew)	0.223	0.21	0.236	0.173	0	0.351
Permission to Meet with 1 Non-Household-Member	0.711	0.802	0.698	0.635	0.262	0.969

Table 3.B.1: Mean of covariates considered in the estimations using the German data in the total sample, the late intervention group, the intermediate intervention group and the early intervention group, respectively.

Variable	Total Sample	Late Timing	Intermediate Timing	Early Timing
	N = 27	N = 8	N = 11	N = 8
Population	315,648	286,649	268,466	409,524
Population Density	503	1,046	278	271
Income per Capita (CHF)	80,404	102,840	73,134	67,964
Share of Population Aged 65+	0.192	0.193	0.19	0.193
Median Age of Confirmed Infections prior to Lockdown	50.19	49.56	49.09	52.31
Initial Growth Trend of Confirmed Cases in Log Points	0.235	0.239	0.21	0.266
Ban on Visits to Retirement Homes	0.593	0.5	0.727	0.5

Table 3.B.2: Means of covariates considered in the estimations using the Swiss (and LI) data in the total sample, the late intervention group, the intermediate intervention group, the early intervention group, the group of counties with curfew and the group of counties without curfew respectively.

3.C OLS Specifications for Germany and Switzerland

	Estimate	Standard Error
Intercept	-1.1628	0.6661
Intermediate Timing	0.3348	0.1403
Late Timing	0.5729	0.2663
Share of Population Aged 65+	-6.4132	2.8281
Population: 0 - 105,878	0.4388	0.2112
Population: 105,879 - 158,080	0.2848	0.135
Population: 158,081 - 251,534	0.0665	0.0985
Population Density: 0 - 117.3	0.0801	0.1425
Population Density: 117.3 - 206.7	0.1201	0.1454
Population Density: 206.7 - 779.7	0.0613	0.1347
Income per Capita: 0 - 27,934	-0.1437	0.1561
Income per Capita: 27,935 - 33,109	-0.1721	0.1439
Income per Capita: 33,110 - 40,506	0.0568	0.1749
Share of Confirmed Infections Aged 80+ prior to Lockdown	4.4466	2.1463
80+ Mortality Rate (per 1000 Inhabitants), 2017	0.2066	0.091
Share of Respiratory-Disease-Related Deaths, 2016	0.9538	3.7197
Hospital Beds per 1000 Inhabitants	-0.0329	0.0184
Initial Growth Trend for Confirmed Cases in Log Points: 0 - 0.14	-0.1188	0.1852
Initial Growth Trend for Confirmed Cases in Log Points: 0.14 - 0.21	-0.089	0.1407
Initial Growth Trend for Confirmed Cases in Log Points: 0.21 - 0.28	-0.0369	0.136
Confirmed Infections per 10,000 Inhabitants on Epidemic Day 4	0.2556	0.0805
Recommendation against Events with >1000 Visitors	0.1594	0.0985
Ban of Events with >1000 Visitors	0.7132	0.141
Curfew	0.2403	0.1111

Table 3.C.1: OLS estimates for Germany 28 days after the start of the county-specific epidemic with fatalities per 10,000 inhabitants as outcome variable.

	Estimate	Standard Error
Intercept	39.2105	45.0916
Intermediate Timing	0.7961	0.7712
Late Timing	1.7187	0.6681
Share of Population Aged 65+	-337.2691	362.2737
Squared Share of Population Aged 65+	848.3766	950.0775
Population: 0 - 59,999	-0.5647	1.1326
Population Density	4e-04	4e-04
Income per Capita	0	0
Median Age of Confirmed Infections prior to Lockdown	-0.2783	1.308
Squared Median Age of Confirmed Infections	0.003	0.0131
Initial Growth Trend for Confirmed Cases in Log Points	6.3784	8.0649
Confirmed Infections per 10,000 Inhabitants on Epidemic Day 4	0.0172	0.6938
Ban on Visits to Retirement Homes	0.153	0.4966

Table 3.C.2: OLS estimates for Switzerland and LI 44 days after the start of the canton-specific epidemic with fatalities per 10,000 inhabitants as outcome variable.

	Estimate	Standard Error
Intercept	-0.5481	0.5532
Curfew	0.089	0.1081
Share of Population Aged 65+	-5.6962	2.9762
Income per Capita: 0 - 27,934	-0.0998	0.1872
Income per Capita: 27,935 - 33,109	-0.0444	0.1598
Income per Capita: 33,110 - 40,506	-0.056	0.1298
Population Density: 0 - 117.3	0.0077	0.1547
Population Density: 117.3 - 206.7	0.1532	0.1558
Population Density: 206.7 - 779.7	0.0388	0.1315
Population: 0 - 105,878	0.1964	0.1917
Population: 105,879 - 158,080	0.1198	0.1565
Population: 158,080 - 251,534	-0.048	0.1067
Share of Confirmed Infections Aged 80+	0.6616	2.0497
80+ Mortality Rate (per 1000 Inhabitants), 2017	0.2029	0.0692
Share of Respiratory-Disease-Related Deaths, 2016	3.5314	3.6582
Hospital Beds per 1000 Inhabitants	-0.0201	0.016
Confirmed Fatalities per 10,000 Inhabitants 10 days before Curfew	-4.3731	4.1915
Confirmed Fatalities per 10,000 Inhabitants 5 days before Curfew	-2.7013	3.4591
Confirmed Fatalities per 10,000 Inhabitants 4 days before Curfew	1.3937	3.7116
Confirmed Fatalities per 10,000 Inhabitants 3 days before Curfew	-2.8829	3.6353
Confirmed Fatalities per 10,000 Inhabitants 2 days before Curfew	5.058	2.5642
Confirmed Fatalities per 10,000 Inhabitants 1 day before Curfew	2.1268	2.1477
Confirmed Cases per 10,000 Inhabitants 25 days before Curfew	2.478	4.2755
Confirmed Cases per 10,000 Inhabitants 20 days before Curfew	0.9095	1.5009
Confirmed Cases per 10,000 Inhabitants 15 days before Curfew	0.0804	0.4324
Confirmed Cases per 10,000 Inhabitants 10 days before Curfew	-0.3862	0.2614
Confirmed Cases per 10,000 Inhabitants 5 days before Curfew	0.0339	0.2059
Confirmed Cases per 10,000 Inhabitants 4 days before Curfew	-0.3237	0.3682
Confirmed Cases per 10,000 Inhabitants 3 days before Curfew	0.1382	0.3992
Confirmed Cases per 10,000 Inhabitants 2 days before Curfew	-0.148	0.2767
Confirmed Cases per 10,000 Inhabitants 1 day before Curfew	0.3193	0.2064
Initial Growth Trend for Confirmed Cases in Log Points	0.0158	0.0264
Recommendation against Events with >1000 Visitors	0.2291	0.075
Ban of Events with >1000 Visitors	0.6488	0.1937
Ban of Groups of >5 Persons (prior to Contact Ban/Curfew)	0.1391	0.1265
Permission to Meet with 1 Non-Household-Member	-0.1802	0.1271

Table 3.C.3: OLS estimates for the impact of curfews (compared to contact restrictions) 35 days after the imposition of curfews with fatalities per 10,000 inhabitants as outcome variable.

3.D Estimations for Germany Without Covariates

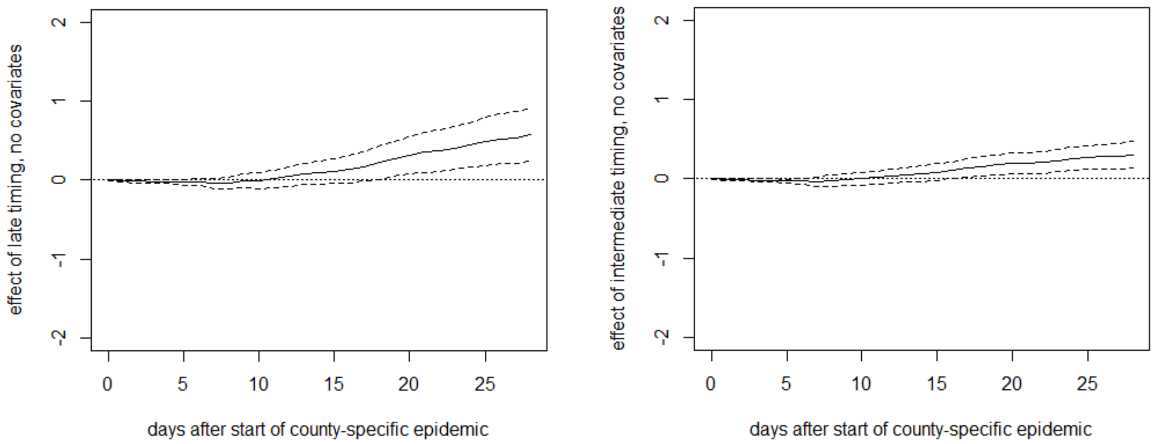


Figure 3.D.1: OLS effects of late (left) and intermediate (right) timing of measures on cumulative deaths per 10,000 inhabitants without covariates.

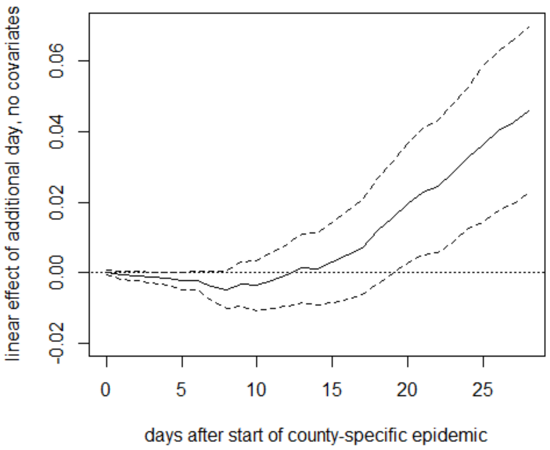


Figure 3.D.2: OLS effect of delaying lockdown by one day on deaths per 10,000 inhabitants in Germany without covariates.

3.E Estimations for Switzerland Without Covariates and Including Ticino

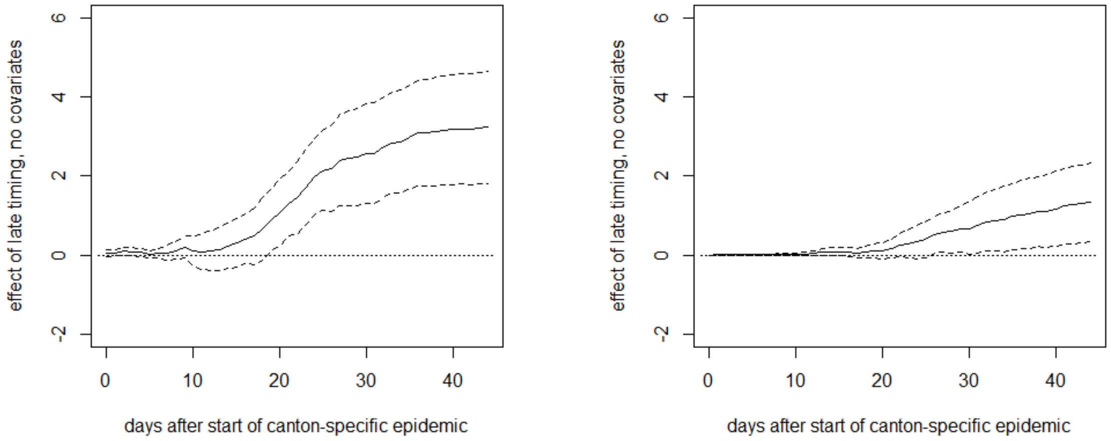


Figure 3.E.1: OLS effect of late timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants without covariates excluding Ticino.

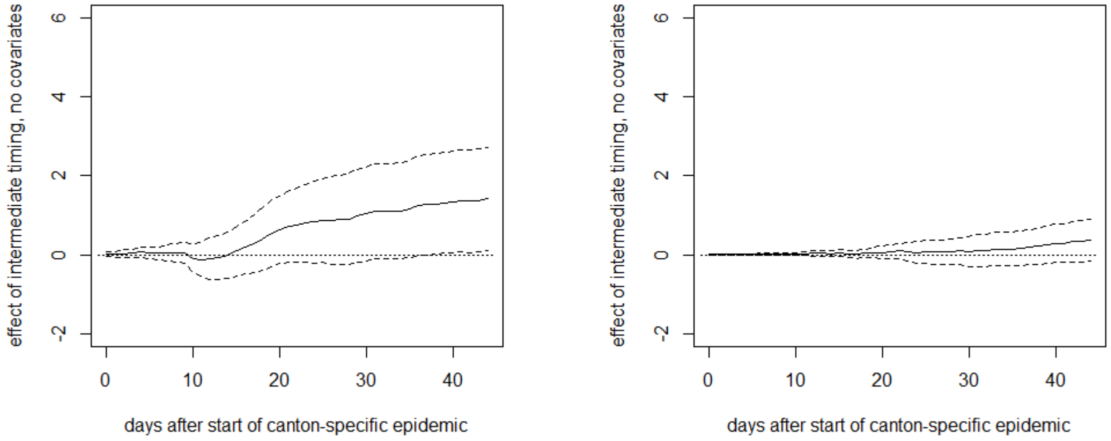


Figure 3.E.2: OLS effect of intermediate timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants without covariates excluding Ticino.

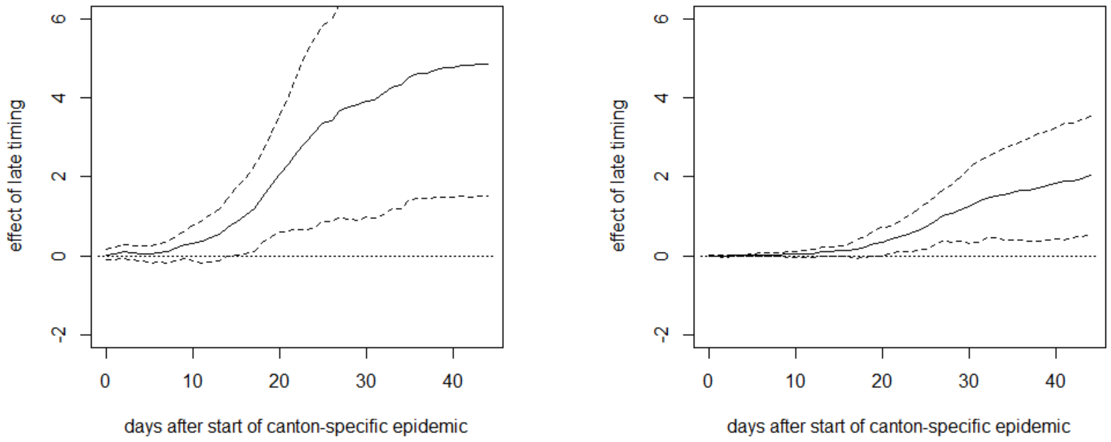


Figure 3.E.3: OLS effect of late timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants with covariates including Ticino.

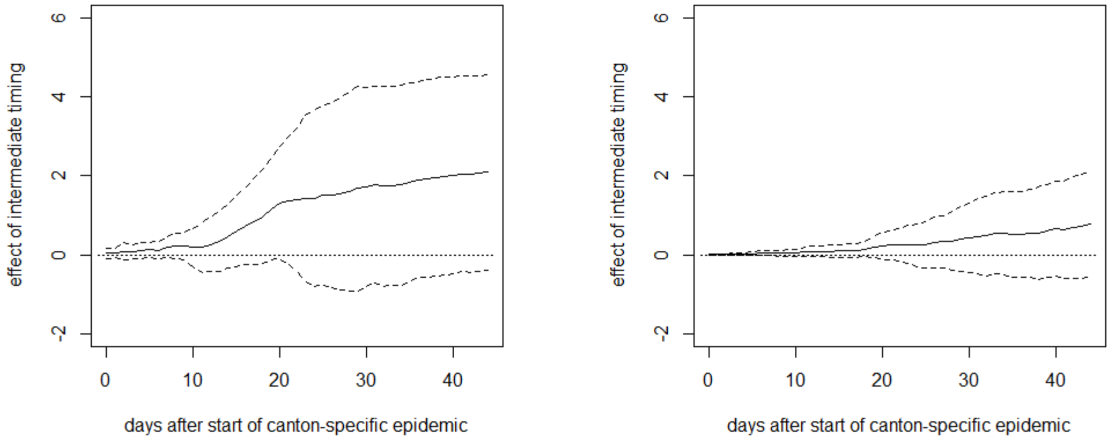


Figure 3.E.4: OLS effect of intermediate timing of measures on cumulative hospitalizations (left) and deaths (right) per 10,000 inhabitants with covariates including Ticino.

Chapter 4

The Impact of the #MeToo Movement on Language in Court - A text-based causal inference approach

Abstract: This study assesses the effect of the #MeToo movement on different quantifiers of the 2015-2020 judicial opinions in sexual violence related cases from 51 U.S. courts. The judicial opinions are vectorized into bag-of-words and tf-idf vectors in order to study their development over time. Further, different indicators quantify to what extent the judges use a language that implicitly shifts some blame from the victim(s) to the perpetrator(s). These indicators measure how the grammatical structure, the sentiment and the context of sentences mentioning the victim(s) and/or perpetrator(s) change over time. The causal effect of the #MeToo movement is estimated by means of Difference-in-Differences comparing the development of the language in opinions on sexual violence and other interpersonal crime related cases as well as a Panel Event Study approach. The results do not clearly identify a #MeToo-movement-induced change in the language in court but suggest that the movement may have accelerated the evolution of court language slightly, causing the effect to materialize with a significant time lag. Additionally, the study considers potential effect heterogeneity with respect to the judge's gender and his/her political affiliation. The study combines causal inference with text quantification methods that are commonly used for classification as well as with indicators from the fields of sentiment analysis, word embedding models and grammatical tagging.

Keywords: text analysis, metoo, difference-in-differences, causal effect.

JEL classification: K49, C21, C23.

I have benefited from valuable comments by participants in the PhD seminar at the University of Fribourg, the Brownbag seminar at Central European University, and various conferences.

4.1 Introduction

After starting as an online campaign against sexual harassment, #MeToo soon evolved into a movement that led to extensive and sustained media coverage of the prevalence of sexual violence in society. In the months and years following its emergence, it shaped public discourse and brought about changes in attitudes and responses to sexual harassment and assault. Several studies have examined the movement‘s impact on various societal, cultural and political dimensions, such as women's perception of safety (Ait Bihi Ouali and Graham (2021)), gender norms in tweets (Moricz (2019)) and the propensity to report a sexual offense (Levy and Mattsson (2021)). To date, however, no study has examined how the movement influenced treatment of sexual violence cases and their victims in court.

The present study assesses how the #MeToo movement affected the way sexual offenses are handled in the U.S. justice system by analyzing the language used in judicial opinions. It examines the effect of the movement by means of a Difference-in-Differences (DiD) and a panel data-based event study approach, where the DiD approach compares the development of language in opinions on sexual offenses to that in opinions on other crimes against persons while the panel event study approach assesses how the language used by judges changed as a result of the #Metoo movement. For quantifying the development of language in judicial opinions, the study develops novel text-based indicators for measuring the amount of victim blaming and introduces an approach for assessing treatment effects based on context and feature vectors. Thereby, it contributes to the recently growing literature on text-based causal inference.

The study proceeds as follows. Section 4.2 provides background information on the #MeToo movement and the U.S. justice system. It reviews the current state of literature on text analysis, particularly the emerging field of text-based causal inference. Section 4.3 describes the corpus of judicial opinions examined in this study. The following section, Section 4.4, illustrates how the judicial opinions are quantified by means of various indicators and text vectorization approaches. Section 4.5 outlines the identification strategy underlying this paper and discusses the assumptions necessary to identify the impact of the #MeToo movement using a DiD model and a panel data-based event study approach. The subsequent section (4.6) summarizes the results and Section 4.7 concludes.

4.2 Background

4.2.1 The #MeToo Movement and its societal, cultural and political impact

The phrase "Me Too" was initially coined by social justice activist Tarana Burke who began using this phrase in 2006 to campaign for the empowerment of sexual violence victims, particularly among women of color. At the end of 2017, following the public exposure of sexual misconduct allegations against Hollywood producer Harvey Weinstein, the phrase spread virally through the hashtag #MeToo (Fileborn and Loney-Howes (2019)).

On October 5th, 2017, The New York Times published an exposé by Kantor and Twohey (2017) detailing decades of sexual harassment allegations against Harvey Weinstein, followed just days later by an investigative article in The New Yorker in which many more women accused Weinstein of sexual harassment and abuse (Farrow (2017)). A few days later, on October 15th, actress Alyssa Milano tweeted a post encouraging victims of sexual harassment and assault to come forward by using the hashtag #MeToo in order to raise awareness about the prevalence of sexual violence in society. The #MeToo hashtag rapidly spread as more and more people around the world shared their experiences with sexual harassment and assault. On Twitter, the hashtag was tweeted about 300,000 times on the day after Milano's post, reached a peak of 750,000 tweets within 24h and was used on average more than 55,000 times per day during the year following the initial tweet (Anderson and Toor (2018)). On Facebook, the #MeToo conversation peaked at 4.7 million participating users within 24 hours, who engaged with over 12 million posts, comments, and reactions (Santiago and Criss (2017)).

The discussion on social media grew into a movement that led to multiple protest marches in the U.S. and around the globe, resulted in extensive and sustained media coverage of the issues of sexual harassment and assault, and shaped the public discourse in the months following October 2017. Caputi, Nobles, and Ayers (2019) estimate that in the first 8 months after the movement's emergence, the number of google searches on sexual harassment and assault exceeded the expected amount by some 86%. According to the Women's Media Center, the number of articles on sexual assault and harassment in a sample of 14 leading U.S. newspapers was more than double the pre-#MeToo average in November 2017 and still exceeded the pre-#MeToo average by 30% some 10 months after the onset of the movement (Ennis and Wolfe

(2018)). Time Magazine (2017) named the "Silence Breakers" - victims of sexual harassment or assault who came forward and thereby started the global dialogue on sexual violence - as its 2017 "Person of the Year", i.e., as the person or group that most influenced the events of the year, according to the magazine.

In the months and years after its onset, the #MeToo movement has brought about changes in attitudes and responses to sexual harassment and assault but also provoked some disillusionment and backlash. Several surveys among both men and women suggest that many are attempting to avoid sexually harassing behavior in response to #MeToo and have perceived others to do so, with this change observed in surveys of workplace behavior as well as behavior in public or private spaces (Ksenia Keplinger and Barnes (2019); Jackson and Newall (2018); Careerarc (2020); Greenfield (2018)).

Other surveys, on the other hand, suggest that many employers expect or have experienced men to become more reluctant towards interactions with female coworkers and subordinates in response to the sexual-harassment discourse triggered by the #MeToo movement – even if such interactions are crucial for the woman's professional advancement (see, e.g., Atwater, Tringale, Sturm, Taylor, and Braddy (2019); French, Mortensen, and Timming (2021); McGregor (2019); Bertotti and Maxfield (2018); NBC News and Wall Street Journal (2017)). In addition, the #MeToo movement has also been blamed to have gone "too far" and it has nurtured the narrative of false and exaggerated accusations against men, resulting in counter-movements with the #HimToo as the most prominent example. The hashtag, originally used by male sexual assault victims, has turned into one for expressing that men are often victims of false accusations (Dejmanee, Zaher, Samantha, and Papa (2020)).

Furthermore, there are several articles arguing that #MeToo (almost) exclusively benefited affluent white women while other groups, such as women from lower socioeconomic backgrounds and women of color, lack the public outreach, support systems and financial security necessary for generating public attention and dealing with potential backlashes (see, e.g., Fileborn and Loney-Howes (2019); Kagal, Cowan, and Jawad (2019); Taub (2019)). The criticism of demographic disparities in victim outreach is also supported by several studies, with Mueller, Wood-Doughty, Amir, Dredze, and Nobles (2021) finding that a disproportionate number of #MeToo tweets were authored by white women, and a study on coverage of the movement in The New York Times revealing an overrepresentation of white victims of sexual violence (Evans (2018)).

However, the underrepresentation of sexual violence victims from certain demographic groups does not necessarily translate into a lesser impact of the movement on those groups. Palmer, Fissel, Hoxmeier, and Williams (2021) assess survey data from a university in the Mid-Atlantic region finding that the number of black students who disclose unwanted sexual activity increased between 2017 and 2019, while white students' disclosure behavior did not change substantially. Levy and Mattsson (2021) find that the effect of the movement on reporting sexual offenses to the police is similar across racial, income and educational groups in the U.S. Further, survey data from the Pew Research Center show that across all demographic groups, the majority of social media users encountered sexual violence content in the months following the #MeToo movement (Pew Reasearch Center (2018)).

To sum up, #MeToo elicited both supportive and defensive reactions, and may have influenced some specific groups more than others. However, all these studies agree that the movement shaped the public debate across demographic groups.

The societal, cultural and political impact of the #MeToo movement has been analyzed extensively with mixed findings about the movement's influence. There are studies showing that the #MeToo movement did not significantly increase self-reported interest in political participation (Castle, Jenkins, Ortbals, Poloni-Staudinger, and Strachan (2020)), that the portrayal of male entrepreneurs in Swedish media changed only marginally after the onset of the movement (Jernberg, Lindbäck, and Roos (2020)), that the #MeToo movement caused a significant increase in the propensity of female dating platform users in South Korea to decline dating requests (Yoon, Choe, Han, and Kim (2020)) and that the #MeToo movement induced a change in gender norms in Swedish-language tweets (Moricz (2019)). A study by Klar and McCoy (2021) reveals that support for the #MeToo movement is associated with a stronger belief in the sexual misconduct allegations against Donald Trump among Democrats but not among Republicans. Ait Bihi Ouali and Graham (2021) examined how the movement affected women's perception of safety by comparing the development of men's and women's perceptions of safety in subway stations of 25 cities around the world with a DiD approach. They found a significant decline in women's perceptions of safety after #MeToo, with the effect of the #MeToo movement being greater the more the country's mass media covered the movement.

In the study most relevant to the present one, Levy and Mattsson (2021) assess the effect of the #MeToo movement on the propensity to report a sexual offense to the police. They do so by applying a triple-difference approach over time, across 31 OECD countries, and between sexual and non-sexual offenses. For countries where the #MeToo movement attracted a great deal of attention, they find that it caused a significant increase of about 10% in the number of reported sexual offenses during the first six months after the movement began. A DiD analysis comparing the development of reported sexual and non-sexual crimes in the U.S. reveals a similar long-term effect for the 15 months following Alyssa Milano's tweet.

The study further shows that the effect of the movement on the reporting of sexual offenses was similar across regions and socioeconomic groups in the U.S., but that it did not affect the reporting of all types of sexual assault equally. Rather, it led to an increase in the number of reports of rape and fondling while not showing any effect on the number of reports of statutory rape and sodomy. Then, in an attempt to identify the mechanism responsible for the increase in reports, the authors assess data from different surveys, which allows them to rule out, among other things, that the rise in reported crimes in the United States is due to an increase in the number of sex offenses. Finally, the study assesses the impact of the movement on sexual offense arrests, finding that the increase in reported sex offenses in the U.S did not bring about a similar surge in the number of arrests. Rather, the authors estimate that the movement raised the number of arrests by only about 6%, which they attribute to the fact that the movement had a stronger effect on cases with a low likelihood of arrest due to being reported more than a month after they occurred, being milder offenses, and/or lack of sufficient evidence. The study does not provide any conclusions about the effect of the movement on convictions. The results of this study will be incorporated into both the identification and design of robustness checks as described in Section 4.5.

Despite the vast number of studies on the societal, cultural and political impact of the #MeToo movement, there is, to the best of my knowledge, not yet a quantitative study on the impact of the #MeToo movement on the justice system. Yet this is certainly interesting, as the movement has created new unofficial reporting pathways for victims who choose not to address a crime through the official legal process for reasons such as potential retraumatization in court and low likelihood of conviction. The #MeToo movement has thereby exposed some difficult issues in the justice system's handling of sexual violence cases. It is now to be examined whether and to what extent the movement has brought about a change in the way cases of sexual violence are addressed in court and how the justice systems deals with victims.

4.2.2 Victim Blaming

Besides raising awareness about the prevalence of sexual harassment and assault in society, the #MeToo movement has also fueled discussions about the acceptance of rape myths. For this reason, the judicial opinions under examination in this study are assessed specifically with respect to the reinforcement of rape myths and victim blaming.

The term “rape myths” refers to stereotypical and inaccurate beliefs about sexual assaults that are prevalent in society. They are often used to shift the blame for a sexual assault from the perpetrator to the victim and to downplay the seriousness of the incident. Rape myths do not necessarily have to be explicitly expressed, but can also be reinforced through rhetorical devices or the mention of specific details when describing sexual assault. For example, studies show that sexual violence reports that focus on external circumstances and the victim’s behavior can increase the likelihood of readers to accept rape myths and view the victim as partly responsible for the assault (see, e.g., Bohner (2001), Franiuk, Seefelt, Cepress, and Vandello (2008) and McCoy (2004)).

Victim blaming and reinforcement of rape myths can be found in a variety of contexts, albeit with varying degrees of explicitness. While often more bluntly expressed in social media and other informal contexts (see, e.g., Suvarna, Bhalla, Kumar, and Bhardwaj (2020) and Suvarna and Bhalla (2020) for studies on identification of victim-blaming language in informal contexts), victim blaming and reinforcement of rape myths are also present in traditional media (see, e.g., Sacks, Ackerman, and Shlosberg (2018), Northcutt Bohmert, Allison, and Ducate (2019) and Franiuk, Seefelt, Cepress, and Vandello (2008)) as well as in the justice system. In court, general victim rights and sexual assault-specific rules, such as rape shield laws, prohibit to some extent explicit blaming, stigmatization and stereotyping. Yet, several studies based on testimonies from various actors, such as victims, barristers, sex crime investigators and independent observers, suggest that victim blaming and rape myths are still prevalent in courts (see, e.g., Temkin, Gray, and Barrett (2018), Spencer, Dodge, Ricciardelli, and Ballucci (2018), Smith and Skinner (2012), Temkin (2000)). While it is often the defense that brings up rape myths for strategical reasons, some judges tend not to intervene and sometimes even take up the argument (Temkin, Gray, and Barrett (2018), Ehrlich (2012)). In addition, many judges tend to employ terminology of affection and consensual sex in cases where the perpetrator is familiar to the victim (Ehrlich

(2012)).

4.2.3 The U.S. Justice System

The legal documents examined in this paper are judicial opinions from 51 U.S. state and federal appellate courts. Appellate courts review legal cases that have already been heard in a lower court (trial court) after one of the parties appeals the trial court's decision. In civil cases, both parties have the right to appeal the trial court's decision while rulings in criminal cases can only be appealed by the defendant in most states. To appeal a decision, the appealing party (the appellant) must file a brief, i.e., a written argument setting forth the facts and arguing why the trial court's decision was erroneous, to which the other party (the appellee) must respond with an appellee's brief.

The appellate court does not usually admit new evidence or witnesses; it may rule solely on the basis of the written briefs or after hearing oral arguments. The appellate court often issues what is called a judicial opinion, i.e., a written decision outlining the court's reasoning; it is usually written by a single judge and reviewed by the other judges on the panel. In cases where one judge does not agree with the majority opinion, she may issue a dissenting opinion; judges who disagree with the reasoning of the majority opinion but do agree with the result may issue a concurring opinion. In some cases, the appellate judges issue an unsigned opinion called a per curiam opinion (American Bar Association (2019)).

The U.S. legal system is based on common law, meaning judges not only apply the laws formulated by legislatures but also take into account how these laws were interpreted in previous comparable cases, with the court decisions on those cases referred to as precedents. A common law system can therefore constantly - without the intervention of legislators - produce new doctrines and let others die out. For this to happen, judges need only adapt to newly emerging circumstances and fill possible gaps in the legislation by interpreting the law accordingly and creating precedents (Harper (2016)). The U.S. judicial system does not consider all opinions to be potential precedents, but distinguishes between precedential opinions, which are opinions that the authoring court believes have sufficient precedential value and are therefore published so that others may cite them as precedents, and nonprecedential opinions, which are written primarily for the parties involved in the case and may be cited by others only as persuasive rather than binding authority.

4.2.4 Text-Based Causal Inference

While there are several socio-economic and legal studies on text classification based on Natural Language Processing (NLP) and also some NLP-based analyses on language development, literature on text-based causal inference is scarce. The social science literature on text classification ranges from studies on differences in the linguistic style between posts in different online communities (Khalid and Srinivasan (2020)) and between comments on #MeToo articles in different news outlets (Rho, Mark, and Mazmanian (2018)) to studies that develop classifiers for political speeches in order to predict the speaker's ideology (Yu, Kaufmann, and Diermeier (2008)) or identify his/her sentiment towards the topic discussed in the speech (Abercrombie and Batista-Navarro (2018)). In the legal domain, Hausladen, Schubert, and Ash (2020) developed a document classifier for judicial opinions from U.S. circuit courts that classifies the opinions according to the predicted ideological direction (conservative vs. liberal) of the decision. In addition, there are several studies on classifying legal documents by topic (see, e.g., Undavia, Meyers, and Ortega (2018), Filtz, Kirrane, Polleres, and Wohlgenannt (2019) and Alekseev, Katasev, Kirillov, Khassianov, and Zuev (2019)).

The evolution of language over time has been analyzed using quantifiers for the context in which words are used (see, e.g., Kulkarni, Al-Rfou, Perozzi, and Skiena (2015), Hamilton, Leskovec, and Jurafsky (2016) and Frermann and Lapata (2016)) as well as based on indicators for the sentiment of that context (see, e.g., Jatowt and Duh (2014) and Hellrich, Buechel, and Hahn (2018)). Nguyen and Rose (2011) analyze how new members in a medical forum gradually adapt their language to the forum's linguistic standards during their first year of forum participation. They quantify the language in posts by using indicators that measure lexical features of posts, such as the number of colloquial words used and the number of words belonging to various psychological, topical and linguistic categories as identified by the Linguistic Inquiry and Word Count (LIWC) tool. Studies on language development either construct indicators to assess their development over time or they rely on vectorizing words or language features and calculating the distance between the resulting vectors in different periods of time.

Literature on text-based causal inference has emerged only in recent years. For one, there are some studies that integrate NLP elements into causal inference, such as some recent studies on text-based confounding adjustment (see Keith, Jensen, and O'Connor (2020) for a review).

Roberts, Stewart, and Nielsen (2020) for example develop a framework for estimating treatment effects which combines text-based matching and confounding adjustment based on text. They apply this framework to control for and match based on the content of publications in order to estimate how a scholar's gender affects the number of citations of his/her publications. Other studies that rely on text-based confounding adjustment include those by Sallin (2021) and Veitch, Sridhar, and Blei (2020). Mozer, Miratrix, Kaufman, and Anastasopoulos (2020) and Field, Park, and Tsvetkov (2020) propose a text matching approach based on distance metrics rather than text classification. Wood-Doughty, Shpitser, and Dredze (2018) integrate text classifiers into causal inference in order to tackle problems with missing data and measurement error.

Other studies use text as treatment or outcome. Ornaghi, Ash, and Chen (2019) analyze how a judge's score on an indicator of gender-stereotyped language affects his/her decisions on women's rights' issues. Tan, Lee, and Pang (2014) assess how wording in tweets affects the number of re-tweets. To do so, they apply Bag-of-Words (BoW) vectorization in order to identify words that in-/decrease re-tweet propensity and quantify wording by means of different indicators of tweet features such as sentiment, lexical distinctiveness and readability. Similarly, Deshpande, Li, and Kuleshov (2022), Pryzant, Card, Jurafsky, Veitch, and Sridhar (2020), Wang and Culotta (2019), Fong and Grimmer (2016) and Feuerriegel, Heitzmann, and Neumann (2015) evaluate how to estimate causal effects of wording, semantics and lexical choices in texts.

Egami, Fong, Grimmer, Roberts, and Stewart (2018) developed a sample splitting framework for estimating treatment effects with text as outcome, building on the classification of outcome texts based on a model trained in the training sample. Sobolev (2018) assesses how troll activity that promotes a pro-government agenda on Russian social media affects the evolution of online discussions using a regression discontinuity approach. To do so, he models the development of conversations on social media as changes in the mixture of topics with topics identified through NLP-based classification. Other examples of studies with texts as an outcome include a study by Chandrasekharan, Pavalanathan, Srinivasan, Glynn, Eisenstein, and Gilbert (2017) on how Reddit's 2015 anti-harassment policy affected the usage of hate speech, as well as an analysis by Pavalanathan, Han, and Eisenstein (2018) on the effect of tagging articles as not written in a "neutral point of view" on the development of lexical patterns in the labeled articles.

Among the studies on text-based causal inference with text as the outcome, there is, to my knowledge, no study yet using a panel event study approach, and so far only one study that applies a DiD approach, namely the study by Chandrasekharan, Pavalanathan, Srinivasan, Glynn, Eisenstein, and Gilbert (2017) on the effects of Reddit's anti-harassment policy on hate speech use. This study compares the development of an indicator that measures the frequency of terms typically associated with hate speech posts between individuals who were members in a group where the anti-harassment policy was violated and those who were not. Hate speech-related terms are identified from conversations in groups that were banned in the context of the introduction of the 2015 anti-harassment policy.

4.3 Text Data

The judicial opinions used in this study were obtained from CourtListener, an archive of court data operated by The Free Law Project (2020). The CourtListener database collects judicial opinions issued by state and federal courts from various sources. The body of opinions for this study is restricted to precedential opinions from state and federal appelate courts, i.e., opinions that the authoring court believes have sufficient precedential value and are therefore made public so that they can be cited as precedents. The reason for this choice is that the corpus of precedential opinions available in the CourtListener database is complete for the available courts, unlike that of non-precedential opinions, i.e., by restricting the sample to precedential opinions, no selection problems arise. In addition, precedents reflect developments in courts and case law. Since precedents are part of the body of law, they contribute to the continuous development of the legal system.

4.3.1 The Body of Judicial Opinions

The body of judicial opinions examined in this paper includes opinions from 51 appelate courts. Most judicial opinions of appellate courts have a similar structure. They usually begin with a summary of the trial court's ruling, followed by the arguments of the defense and the prosecution that were presented and admitted before the trial court, as well as the appellate brief filed by the defense. They typically conclude with reasoning and the decision of the appellate judges. Thus, the opinions reflect the atmosphere during the trial court hearing and the appeal proceedings, as well as the attitude of the defense and the judges towards the victim.

To single out the opinions on sexual offenses and those on other crimes against persons from the full body of precedential opinions available in the CourtListener Database, I take advantage of the fact that the appelate court opinions have a similar structure. The opinions usually begin with an introduction that specifies the offense(s) for which the offender was convicted in trial court and can therefore be reliably classified based on term search in the introduction. To distinguish opinions on sexual-violence related cases and those on other cases of interpersonal crimes, I rely on regular-expression-based (`regex`-based)[1] search of the legal terms for such crimes[2] in the introduction of the respective opinion.

As the headers of the opinions differ between courts, their divisions and even the judges who author the opinions, the introductions cannot be reliably identified based on `regex` rules. Therefore, the first third of each opinion but no more than 5,000 characters are defined as the introduction. The introductions of all opinions available at CourtListener are searched for matches to legal terms related to sexual violence; the remaining opinions are then searched for matches to other interpersonal-violence related legal terms. While certainly not adequate to accurately classifying opinions into crime categories, the identification procedure described above ensures that opinions are selected into the sample based on the same rules throughout the observation period. Focusing on the introduction of opinions prevents erroneously classifying opinions to one of the two groups when an opinion mentions crimes from a party's past.

Using the `regex` procedure described above, I can identify 43,088 opinions on crimes against persons published between January 2015 and November 2020, including 15,312 on sexual offenses and 27,781 on non-sexual offenses against persons. The number of opinions per court is provided in Table 4.A.1.

4.4 Text Quantification

In order to assess the impact of the #MeToo movement on language in court, the judicial opinions need to be quantified. In this section, I outline the different approaches I use to do so,

[1] A regular expressions is a string of characters that is used to specify a search pattern. Regular expressions have a syntax that allows to match a set of different character strings. To identify the term "sexual assault", for example, a regular expression can be constructed that matches that exact term, its plural as well as the same term but with more than one space between "sexual" and "assault", with a line or page break between both words or with either of the two words capitalized

[2] The legal terms for sexual offenses differ strongly across U.S. states (see https://apps.rainn.org/policy/#report-generator for state-specific terms and definitions); therefore, different state-specific sets of legal terms are used. The legal terms for offenses of interpersonal violence are more homogeneous across the states; consequently, the same set of terms is used for all courts where the legal terms are obtained from the following site: https://www.criminaldefenselawyer.com/topics/crimes-against-persons)

where the quantifiers described in Section 4.4.1 aim at quantifying the amount of victim blaming in each opinion, while the text vectorization methods outlined in Section 4.4.2 are later used to capture the general evolution of language in judicial opinions.

4.4.1 Victim Blaming Indicators

To quantify the extent of victim blaming in judicial opinions, three indicators are constructed based solely on sentences in which the victim or perpetrator is named. These indicators aim at capturing the extent to which opinions contain wording that implicitly shifts some blame from the perpetrator onto the victim, where such wording may come from the defense, the prosecution, or the judges involved in the case.

In the appelate court opinions, the victim is often referred to by his/her name or by the term “victim”. The appealing party, i.e., the person found guilty in trial court, is usually called the “appellant” or by his/her name, but may also be referred to as the “petitioner”. As the victims’ and the appellants’ names are not clearly identifiable, only sentences containing the words “victim”, “appellant” and “petitioner” as well as inflected forms of these words are considered in the construction of the indicators.

The Semantic Role of Victim: Subject vs. Object

The first victim-blaming indicator captures the semantic structure of the sentences in which the victim is mentioned. The reason for this is that various studies from the field of psychology show that grammatical structure can, on the one hand, provide information about how the author perceives a fact and, on the other hand, influence the perception of the recipients. Niemi and Young (2016) asked participants in an experiment on Amazon’s Mechanical Turk (MTurk) to read fictional reports of rape in which either the victim or the perpetrator was the subject of some 75% of the sentences. Participants who read reports in which the victim was primarily the grammatical subject were more likely to shift some responsibility for the assault to the victim. These findings are in line with a study by Strickland, Fisher, Keil, and Knobe (2014) in which study participants were asked to judge the intentionality of the grammatical object and subject in a set of sentences that were ambiguous in terms of intentionality. Study participants attributed significantly more intentionality to the grammatical subject than to the object.

The above findings also apply to sentences and texts written in passive voice. In an experiment by Bohner (2001), study participants were asked to describe an uncommented video showing a rape scene and to complete a questionnaire measuring rape myth acceptance. The study reveals that describing the scene primarily in the passive voice (with the victim as the subject) is positively correlated with attributing responsibility to the victim. Henley, Miller, and Beazley (1995) confronted study participants with fabricated news reports on violence against women written in either active or passive voice. They found that males but not females rated the perpetrator's responsibility higher after reading reports in the active voice.

These findings suggest developing a victim-blaming indicator that captures the semantic role of the victim in judicial opinions. I construct an indicator that measures the relative frequency of sentences in which the victim functions as grammatical subject, where the semantic roles of all words are identified using the Python package `spacy`.

Sentiment Orientation

The second indicator aims to capture the sentiment orientation of sentences mentioning the perpetrator. Jatowt and Duh (2014) developed such an indicator based on SentiWordNet, a database with information on word sentiments. The English SentiWordNet 3.0 contains more than 100,000 words, each of which is assigned sentiment scores for positivity and negativity. Since many words have different meanings/senses depending on the context in which they are used, the SentiWordNet dataset contains a separate entry for each meaning of a word. The different word meanings are ranked according to how frequently the word is used with the different meanings.

Jatowt and Duh (2014) propose to calculate sentiment scores for a word of interest as the average of the positivity or negativity scores of all context words (the words surrounding the word of interest). As the meaning of the context words is not identifiable without deeper content analyzes, they take the weighted average of the scores for all meanings of each context word, with the weights calculated based on the meaning ranks from the SentiWordNet dataset.

To assess the sentiment orientation of sentences in which the perpetrator is named, I determine the negativity score as suggested by Jatowt and Duh (2014). I consider the words that are at most five words away from the mention of the perpetrator and occur in the same sentence. The reason for focussing on the context words surrounding the mention of the perpetrator is that positively connotated context words of the victim may not point at the absence of victim-blaming language, as, e.g., comments on the victim's clothing or attitude towards the perpetrator are common examples of rape myth reinforcement.

Word Embedding

In a third approach to identifying the effect of the #MeToo movement on the use of victim-blaming language, the development of the context in which the words "victim" and "appellant"/"petitioner" appear is assessed by means of Word2Vec, a neural network-based word embedding method. Word2Vec is a method for vectorizing words in such a way that each word vector of predefined length captures the context in which the represented word usually appears, and thereby its semantic and syntactic properties. There are two approaches to learning the vector representation for each word in a text corpus. In the Common Bag of Words (CBOW) approach, the Word2Vec neural net takes context words (the words surrounding the unknown target word in a sentence) as input and returns probabilities for each word in the model vocabulary to appear in the given context. In the Skip-Gram approach, the neural network takes single words as inputs and predicts their context. In both approaches, the inputs are passed through a hidden layer that is constantly updated in order to optimize the returned prediction probabilities. Once the model is trained, the context vectors can be extracted from the hidden layer (Mikolov, Sutskever, Chen, Corrado, and Dean (2013)).

The judicial opinions are first decomposed into sentences. In a second step, these sentences are tokenized and lemmatized, i.e., they are split into individual words which are then converted to their base form using Python's `nltk` package. These preprocessed sentences are fed into the Word2Vec algorithm. Since the analysis aims to compare the development of the embedding of the words "victim" and "complainant"/"appellant" in opinions on sexual and non-sexual crimes, the Word2Vec model is trained for each quarter and crime group separately. In order to make the Word2Vec models comparable, they are aligned using Compass Aligned Distributional Embeddings (CADE). As suggested by Jatowt and Duh (2014), I calculate the cosine similarity between the vectors of the first quarter of 2015 (for the words "victim" and "complainant"/"appelant") and those for each other quarter, both for the group of opinions on sexual offenses and the group of non-sexual offenses, in order to be able to assess the development of the word embedding vectors.

Since the word embedding approach only identifies one vector per time period and group, this approach only allows for estimating the effect of the #MeToo movement, but not the uncertainty of the estimate, i.e., the standard errors.

4.4.2 Text vectorization

For the analysis of the linguistic development in the judicial opinions, the opinions are vectorized by means of the Bag of Words (BoW) and the term frequency - inverse document frequency (tf-idf) models, both of which are frequently used for text classification. The resulting vectors are then used to quantify the development of the language in court over time, as described further below.

Before applying any of the two vectorization methods to the set of judicial opinions, all opinions are cleared of so-called stop words, i.e., function words with ambiguous meaning such as determiners (e.g., "the", "a", "another"), coordinating conjunctions (e.g., "but", "yet", "so") and prepositions (e.g., "in", "under", "before"). For doing so, I use the predefined set of such stop words provided in the `nltk` package. Further, the words remaining in the corpus are tokenized and lemmatized, also by use of the `nltk` package. Through pre-processing, each judicial opinion is reduced to a corpus of lowercased words that have lexical meaning and are set to their base form.

The described pre-processing of text documents is widely used in the literatur on topic extraction and text classification. It aims to reduce the dimensionality of the feature space and to increase efficiency. Some studies on stop word removal in text classification, however, suggest that its impact is small (see, e.g., HaCohen-Kerner, Miller, and Yigal (2020) and Uysal and Gunal (2014)); some studies in the field of sentiment classification even suggest that stop word removal might negatively impact classification accuracy (Ghag and Shah (2015), Kharde, Sonawane, et al. (2016)). For this reason, I only remove stop words for text vectorization but not for calculating the victim blaming indicators.

The Bag of Words Approach

The Bag of Words (BoW) model is a representation of text that converts text documents into fixed-length vectors, where each vector represents the relative frequencies of words in one document. The set of unique words occuring in the corpus of all text documents to be analyzed, hereafter referred to as the model vocabulary, determines the length of the vectors representing the text documents. Formally, the model can be described as follows: Each dimension of the vector v_d^{BoW} representing text document d corresponds to one word w from the model vocabulary W. Vector element $v_{d,w}^{BoW}$ thus describes the relative frequency of word w in document d, i.e., the number of appearances of word w in document d divided by the total number of words in document d. Applied to the corpus of judicial opinions, each vector v_d^{BoW} represents one judicial opinion d.

There are a number of words in judicial opinions, such as "court", "trial" or "judge", that are not included in the list of stop words, but nevertheless appear in almost all opinions and do not provide any insight into the evolution of the language used in court. To place greater emphasis on salient words that may reflect linguistic developments in court opinions, I not only determine the BoW representation of opinions based on the entire vocabulary, but also construct a second set of BoW vectors based only on a vocabulary that includes words that occur in less than 95% of all cases. This alternate BoW approach will in the following be referred to as reduced-corpus BoW. The tf-idf model described in the next section goes in a similar direction as this reduced-corpus BoW, since both aim at reducing the influence of very frequently used words in the vector representations of judicial opinions.

The Term Frequency - Inverse Document Frequency Approach

The tf-idf model represents text documents as fixed-length vectors just as the BoW model, only that the vector elements of the tf-idf vectors $v_{d,w}^{tfidf}$ are computed as the relative frequency of word w in document d multiplied by the logarithmically scaled inverse proportion of documents in the corpus that contain word w. The vector element $v_{d,w}^{tfidf}$ can be calculated as

$$v_{d,w}^{tfidf} = wf_{d,w} \times log(\frac{N}{df_w}),$$

where $wf_{d,w}$ denotes the relative frequency of word w in document d, N the number of documents in the corpus and df_d the number of documents containing word w. Thus, the tf-idf vector element $v_{d,w}^{tfidf}$ is comparably small if the corresponding word w occurs in (almost) every document, such as "court" or "judge", and large if the word w appears in document d but is not contained in multiple other documents. This way, the words that are characteristic of one or a set of document(s) are given a higher weight.

When applied to the corpus of judicial opinions without further text processing, the tf-idf model yields the problem that names of persons, places and organizations that naturally only occur in one or a small set of opinion(s) are given particularly large weights, even though they are not relevant to the documents' content. To circumvent this, I apply the Stanford PoS Tagger, a probabilistic conditional log-linear model that - based on lexical features of words as well as the context in which they appear - tags each word in a text as corresponding to a grammatical category such as verb, noun, proper name, etc. This way, I identify and exclude those words that correspond to names of persons, places and organizations before removing stop words, lemmatizing the texts and finally applying the tf-idf model.

Assessing the Development of Text Vectors over Time

The high-dimensional vectors determined with the three approaches outlined above are then projected using principal component analysis (PCA) in order to reduce the dimensionality and thereby the computational complexity, while preserving as much of the variance of the original text vectors as possible. Since the dimensionality reduction is applied over time to all text vectors together, the resulting vectors are suitable for computing similarity between the opinions they represent. Jatowt and Duh (2014) take a similar approach to evaluate the evolution of word context vectors over time.

The data is reduced to as many principal components as needed to capture 90% of the variation in the original opinion vector dataset. The resulting reduced vectors have 64 (BoW), 65 (reduced corpus BoW) and 36 (tf-idf) dimensions respectively. Studies on text classification usually reduce the data to only 1-5 dimensions to avoid capturing noise. The goal of the present study, however, is not to classify texts, but rather to detect even minor changes in the language used in judicial opinions. These minor changes may indicate, for example, a change in the treatment of victims in court, while in the context of the thematic classification of the comments, they may be considered noise.

After this data reduction step, I average the vectors of all opinions from the first half of 2015 for each court separately and calculate the distance between these 2015 average vectors and all other opinion vectors of the corresponding court (from opinions published between July 2015 and November 2020) in order to quantify each opinion by its distance from the average of the first half of 2015. To do this, I choose the L_1 distance metric (also known as Manhattan distance) as proposed by Aggarwal, Hinneburg, and Keim (2001). They show that for data of 20 dimensions and more the L_1 metric is the best distance measure in terms of contrasting two points. Other distance measures as well as other dimensionality choices will also be considered as part of the robustness checks.

4.5 Identification

This section first outlines how I attempt to disentangle the impact of #MeToo on the text quantifiers described above from other potential external influences on court language. The estimation strategy based on the DiD and panel event study approach is discussed further below in Section 4.5.2.

The purpose of this study is to assess the impact of the #MeToo movement on language in court when dealing with cases of sexual violence. The outcome, courtroom language, is expected to reflect changes in the treatment of sexual offenses and their victims, as well as in the atmosphere at such trials, that are not due to directly measurable, exogenously imposed reforms, but rather to changes in the attitudes of the parties involved. Language in judicial opinions can shift as the parties involved in a trial change the way they describe sex offenses and how they address victim and offender, where these changes may be both a conscious or unconscious expression of attitudinal shifts. Judges may also change to more conscientiously apply rules of court, codes of conduct and rape shield statutes, with rape shield statutes being laws designed to protect victims of sexual offenses by, for example, prohibiting evidence relating to the victim's past sexual behavior. Moreover, as the U.S. legal system is a common law system, judges may begin to interpret other laws differently, providing the impetus for new doctrines. Such changes in the interpretation of law are most visible in precedential opinions and are likely to be reflected in the language used in those opinions. All of these changes in the way cases of sexual offenses and their victims are handled in court due to changing attitudes are not directly measurable, but rather must be gleaned from written opinions.

By examining various indicators that quantify the extent of victim blaming in court, I place particular attention on how the #MeToo movement has affected the treatment of victims in court. Just as with the more general language quantifiers, victim blaming indicators can capture both conscious and unconscious manifestations of attitudinal shifts, with the semantic indicator likely capturing primarily unconscious change, while the other two indicators of victim blaming capture both types of attitudinal shifts. The #MeToo movement may be defined as the collective action of sexual violence victims who used the phrase and hashtag "Me Too" for publicizing their experiences of sexual harassment and assault in order to point out the prevalence of sexual violence in society.

4.5.1 Disentangling the Effect of #MeToo on Language in Court from External Factors

A major difficulty of estimating the effect of the #MeToo movement is disentangling the movement's impact from other external influences that may induce changes in the language used in judicial opinions on sex offense cases.

One such confounding influcence could be legislative reforms. The language in court is likely to be affected by reforms that, for example, bring about changes in how certain crimes are sanctioned, what evidence is admitted in court, or what role parties and witnesses may take in the court process. Since this study aims at estimating the effect of the #MeToo movement on court language and victim treatment, rather than on legislative reforms, such reforms would bias the treatment effect estimate.

To the best of my knowledge, and as noted by Levy and Mattsson (2021), there have been no major legislative changes on crimes against persons (neither on sexual nor on non-sexual offenses) in the United States during the years under study. There are only some states that have enacted laws prohibiting the use of non-disclosure agreements which prevent victims of sexual harassment or assault from speaking out. It is unlikely that this legislative change had a direct effect on the language used in sexual offense lawsuits, as such non-disclosure agreements while designed to prevent sexual harassment charges, have no bearing on what victims say in court when a crime is tried. The prohibition of non-disclosure agreements, however, may have increased the number of sexual offense reports, as victims may have come forward who would have been prevented from speaking out in the absence of these laws. This issue of changes in the number of reported cases of sexual violence is discussed further below.

Then, some states have introduced policies on workplace harassment such as implementation of anti-harassment trainings and laws requiring transparency about sexual harassment investigations (Johnson, Sekaran, and Gombar (2020), National Conference on State Legislature (2019), Myers (2020)). Since these policies only address how employers should handle sexual harassment, and not how harassment should be handled once it is litigated in court, these policies are also unlikely to have any effect on the language in courts. Other than these changes, there were no other major changes in state legislation on interpersonal crimes between 2015 and 2020. I can therefore rule out the possibility that the changes in language are due to legislative reforms.

Another source of potential bias in the effect estimates are other exogenous changes in the judges' attitudes toward sexual harassment and assault cases. Such exogenous changes in attitudes may result from other sexual harassment or assault scandals, other movements, or even the release of a book or film that addresses sexual violence. All of these events could potentially draw public attention to the issue of sexual violence and lead to a change in attitudes toward sexual violence cases and their victims. These concerns can be mitigated by looking at search history, traditional media and social media data from the United States. All of these data show that at no other time during the period studied was as much attention drawn to sexual harassment and assault as in the weeks following the onset of the #MeToo movement.

The #MeToo movement was very effective in raising awareness about the prevalence of sexual violence. About 65% of social media users report having regularly encountered at least some content related to sexual harassment or assault on social media platforms in the months following the start of the #MeToo movement, with little difference across demographic groups (Anderson and Toor (2018)). The observation that a large share of society has been confronted with the issue of sexual violence is also reflected in Google search data. Following Caputi, Nobles, and Ayers (2019) and Levy and Mattsson (2021), I look at how often the terms "sexual assault" and "sexual harassment" were searched for on Google in the United States during the study period. Figure 4.5.1 shows that public interest in these topics has never been higher than at the start of the #MeToo movement. Additionally, Figure 4.5.1 also shows how often these terms were searched for on Google News, with a similarly pointed peak of searches in October 2017. The onset of the movement also saw a sharp increase in the number of articles in traditional media covering sexual harassment and assault, as shown both in an analysis of four major U.S. newspapers conducted by Levy and Mattsson (2021) and in a study published by the Women's Media Center (Ennis and Wolfe (2018)). The fact that the topic of sexual violence has been given great prominence in both traditional media and social media, and that all socio-economic groups have been reached, it is very likely that judges have also been confronted with the #MeToo movement and may have been influenced in their attitudes.

Another issue with identifying the treatment effect is that it is not possible to separate the effect of the #MeToo movement itself from the impact of the sexual assault scandal that triggered it, i.e., the effect of the social media attention to sexual violence and that of the revelations about Harvey Weinstein and about other prominent cases in the immediate aftermath of the Weinstein scandal. However, the caveat that any potential impact of the movement on

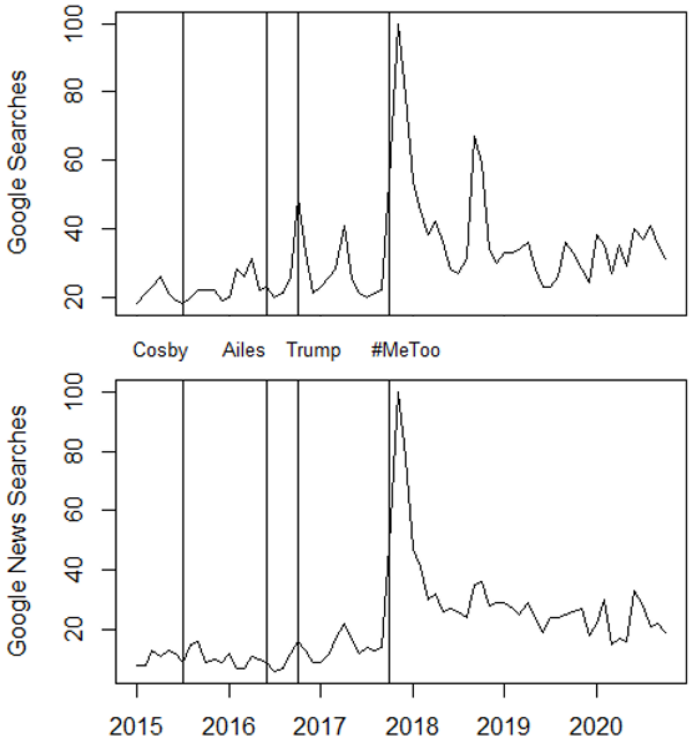


Figure 4.5.1: Searches on Google and Google News in relation to the highest point in the period January 2015 to November 2020. The vertical lines indicate when the sexual harassment/assault allegations against stand-up comedian Bill Cosby, former U.S. President Donald Trump and Fox News CEO Roger Ailes became public.

attitudes and language can be attributed to the Weinstein scandal can be partially debunked. During the study period, there were many other sexual assault or harassment scandals involving many men with similar or even higher profile than Harvey Weinstein, none of which attracted nearly as much public attention to the issue of sexual harassment and assault as the #MeToo movement. Following Levy and Mattsson (2021), Figure 4.5.1 indicates the timing of three exemplary scandals involving prominent men that did not lead to a comparable increase of search interest as did the #MeToo movement. I argue, therefore, that changes in attitudes and language are unlikely to be due to the Weinstein sexual assault scandal itself because, in the absence of a #MeToo movement, the scandal would likely have attracted no more public attention than the other prominent cases.

Finally, identifying the causal effect of #MeToo on language in court entails a third major issue, namely, that the composition of sexual offense cases heard in court may have changed as the #MeToo movement led to an increase in the reporting of such crimes. Levy and Mattsson (2021) show that the #MeToo movement has brought about an increase in reports of sexual offenses, which in turn induced a slight, albeit statistically significant, rise in arrests related to sexual offenses, where arrests are defined as cases in which the suspect is taken into custody or summoned to court. Some of these additional arrests (that arguably would not have happened in absence of the #MeToo movement) may have resulted in convictions, some of which in turn may have been appealed and thus become part of the sample. Similarly, there may be cases in my sample that went to trial only because of the prohibition on non-disclosure agreements that some states enacted in response to #MeToo.

Although the increase in cases in itself is not problematic for identifying the effect of #MeToo on language in court, it may have led to changes in the composition of sex offenses addressed in the judicial opinions in the sample. This, in turn, would be troublesome, as such compositional changes would likely result in language shifts that cannot be attributed to a change in how a given case is treated in court. I must therefore disentangle the direct effect of the movement on court language from the indirect effect mediated through compositional changes in the set of sexual opinions.

However, concerns about #MeToo-induced compositional changes in my sample can be debunked to some extent as the increase in reports of sexual offenses triggered by the movement is unlikely to have caused many additional cases in the set of appelate court opinions under study. For one, the findings by Levy and Mattsson (2021) suggest that the set of additional sexual offense reports includes a disproportionate number of comparatively lighter crimes or cases with less pressing evidence. Based on survey data, the authors argue that the movement encouraged sexual offense reports through changing the victims' perception of the severity of the experienced sexual offense. In addition, the authors note that the movement had a particularly strong effect on the reporting of cases that occured at least one month before being brought to the police, i.e., in cases that are more difficult to prove in court. In both cases, i.e., low severity sexual offenses and cases without pressing evidence, it is unlikely that there will be a criminal trial and thus that they may end up in an appeals court and in my sample.

The argument that most MeToo-induced reports will not end up in my sample can also be substantiated by looking at the authors' results on the types of sexual offenses for which the number of reports has increased. They show that the increase in reports of sexual offenses and resulting arrests/summonses is mainly driven by an increase in sexual harassment reports, which very seldom result in criminal trials. The estimated effect of #MeToo on the number of sexual assault reports is substantially smaller but still statistically significant for two of four sexual assault subcategories, namely rape and fondling. Compared to cases of sexual harassment, arrest/summonses related to these two categories of sexual assault are more likely to result in criminal proceedings, which in turn can lead to compositional changes in my sample and must be considered when assessing the movement's impact on court language.

	2015	2016	2017	2018	2019	2020
Share of Sexual Offense Opinions	0.361	0.356	0.353	0.353	0.357	0.352
		(0.503)	(0.646)	(0.865)	(0.888)	(0.275)
Composition of Sexual Offenses:						
Sexual Assault	0.834	0.815**	0.786	0.817	0.824	0.827*
		(0.932)	(0.015)	(0.644)	(0.773)	(0.098)
Sexual Assault on a Minor/Child	0.028	0.028	0.028	0.032	0.026	0.036
		(0.745)	(0.26)	(0.502)	(0.506)	(0.345)
Statutory Sexual Assault	0.134	0.151	0.148	0.141	0.134	0.155
		(0.982)	(0.616)	(0.977)	(0.42)	(0.731)
Sodomy	0.091	0.114	0.102	0.084	0.103	0.113
		(0.034)	(0.316)	(0.964)	(0.639)	(0.361)
Fondling	0.306	0.315	0.325	0.339	0.32	0.338
		(0.483)	(0.533)	(0.241)	(0.49)	(0.454)
Sexual Harassment	0.002	0.006	0.007	0.006	0.004	0.006
		(0.57)	(0.405)	(0.553)	(0.846)	(0.438)
# Opinions (Total)	6474	7685	7111	7223	7883	6713
# Sexual Offense Opinions	2333	2737	2510	2548	2812	2367

Table 4.5.1: Shares of sexual-violence related opinions that deal with different crime types (by year), as well as the p-value for the differance between the 2015 share and the share in the respective year (with court fixed effects). Note: For 2016-2020, the year y is defined as November y-1 to October y in order to have a clear cut at the onset of the #MeToo movement in November 2017; the year 2015 only consists of the months January-October 2015. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

A look at Table 4.5.1 reveals that the share of sexual offense opinions across the sample is constant over time, with 35-36% of opinions relating to sexual offenses each year. The table also reports the share of sexual offense opinions that address different sexual offenses, categorizing opinions by which sexual offense-related terms could be identified in the opinion's introduction, i.e., one opinion may be counted in more than one category. To draw on the results of Levy and Mattsson (2021) for robustness checks, the terms identified in the introductions are grouped into six categories of sex offenses similar to those defined by the FBI National Incident-Based Reporting System, which Levy and Mattsson (2021) also rely on. In addition to the five categories that match those used in the above study I also included a subcategory of the broad category of sexual assault, namely sexual assault of children and minors, to examine whether an increase in reports from a particular group may have been offset by a decrease in reports from another group. The table indicates that the composition of sexual offense cases does not show any substantial shifts over time, with none of the sexual offense types having a constantly in- or decreasing share over time, also when particularly looking at the years 2017 to 2019 for which Levy and Mattsson (2021) find a significant increase in reports. The table also reports the p-values for the difference proportion of each category of sexual offense in 2015 and the proportion in every other year, controlling for court fixed effects. The p-values indicate that there are no statistically significant differences between the years under study. When not controlling for court fixed effect, a few of these differences are moderately statistically significant, both before and after the #MeToo movement, suggesting that the slight intertemporal differences in the composition are attributable to general variations in the number of opinions per court and differences across courts in the composition of opinions. I can therefore conclude that if #MeToo induced compositional changes in the composition of sexual offense opinions in the sample, they are small and not statistically significant.

4.5.2 Empirical Strategy

The effect of the #MeToo movement on the language at court is assessed by means of a DiD approach and an event study approach both of which are outlined further below. To account for the fact that the process between the initial trial court hearing and the publication of a judicial opinion from an appelate court often takes several months, sometimes years, I particularly look at the effect of the #MeToo on court opinions published at least one year after the movement began; that is, while the estimates for the year following the movement's onset are also reported, a stronger emphasis is put on subsequent years, particularly November 2018 through April 2020, when the Covid-19 pandemic hit the United States. While appeals must be filed promptly after publication of the trial court's decision (usually within 30 days), the process in appeal courts often takes considerably longer: the U.S. Courts of Appeals report the median disposition time, i.e., the time between the filing of an appeal and the appelate cour's decision, to be 8.6 months for 2015.[3] Thus, opinions published at least one year after the movement's onset are likely to contain closing arguments, trial court judgements, appellee and appelant briefs, as well as appelate court reasoning and decisions written after the #MeToo movement began. However, as

[3] see https://www.uscourts.gov/news/2016/12/20/just-facts-us-courts-appeals

trial court proceedings can also last for many months or even years, opinions on cases with such lengthy trial court proceedings may also contain citations from or references to early trial court hearings that took place before #MeToo. However, the fact that the opinions may include parts of hearings from before #MeToo should be taken into account when interpreting the estimated effect, as this could lead to an underestimation of the movement's impact on language in the courts.

Difference-in-Differences Approach

I apply a DiD approach both with two and multiple time periods in order to identify the Average Treatment Effect on the Treated (ATET) (see, e.g., Snow (1856), Card and Krueger (1994) and Acemoglu and Angrist (2001)). The sample consists of all judicial opinions on cases that can be classified as "crimes against persons". Opinions classified as treating sexual crimes form the treatment group, i.e., the set of opinions that are affected by the #MeToo movement, while opinions on other crimes against persons serve as control group. One advantage of choosing the sample of "crimes against persons" is that all crimes against persons involve a victim and a perpetrator, which is important for detecting victim-blaming language. Further, the comparison of sexual offenses and other crimes against persons is also common in the literature on victim blaming and rape-myth acceptance (see, e.g., Reich, Pegel, and Johnson (2021), Bieneck and Krahé (2011) and Levy and Mattsson (2021))).

In order to identify the ATET of the #MeToo movement using a DiD approach, certain assumptions must hold. For simplicity, I will discuss these assumptions for the simple case of two time periods, but the discussion is easily transferable to the multiple time period approach.

The common support assumption states that for each post-treatment observation from the treatment group, there must be comparable observations in the other three groups, i.e., the pre-treatment observations from the treatment and the control group, as well as the post-treatment observations from the control group. Translated to the present study, there must be opinions in these three groups that are comparable to post-treatment opinions on sexual offenses in terms of court and judge characteristics. This assumption is likely to hold because I only consider judicial opinions from courts that handle both sexual offense cases and cases of other crimes against persons. Further, judges are usually assigned their cases randomly, which is why the characteristics of judges in all 4 groups should be similar.

Then, the no anticipation assumption states that the treatment may not have any effect on the outcome in pre-treatment periods, which would have been the case if judges and other actors in court had anticipated the movement and changed their behavior accordingly before the movement went viral. This assumption is also likely to hold as the #MeToo movement was launched immediately after the sexual assault allegations against Harvey Weinstein came to light, and to an extent that no one anticipated.

The third assumption, finally, is the common trend assumption. It states that in absence of the treatment, outcomes in the treatment and control group would follow a parallel trend, or in other words, that the gap between the outcome in the two groups would be constant over time. This assumption is supported by the fact that between 2015 and 2020, there have not been any major legal change in either area, i.e., that of sexual offenses and other crimes against persons, that is likely to affect procedures and language at court (see Section 4.5.1). In addition, all crimes against persons are relatively "old" crimes, i.e., crimes that are unlikely to change in nature and have long been considered crimes, so there is no need to continually re-interpret the corresponding laws (unlike, e.g., cyber crimes). The above discussion suggests that, in the absence of external interferences, the language of judicial opinions would evolve at a similar pace for all types of crimes against persons.

However, this argumentation for the common trend can be criticized in that recent decades have seen a trend toward more gender-sensitive language in parts of society, which is likely to have a greater impact on language related to sexual offenses than on language related to other crimes against persons. If this trend toward more gender-sensitive language is also perceptible in court, the parallel trend assumption may be violated, as then the language in opinions on sexual offenses would generally, i.e., even in absence of the #MeToo movement, evolve more rapidly than that in other opinions. Further, other feminist developments in society over the past few decades may have encouraged judges to re-interpret the law on sex offenses more frequently than that on other offenses, even if it was not for the #MeToo movement. Both these issues would lead to a violation of the parallel trend assumption and certainly suggest interpreting the results from the DiD approach with some reservation. However, the parallel trend assumption can and will be tested by means of placebo tests, in order to rule out serious violations of the common trend assumption.

Another potential source of bias would be changes in the composition of the sample. With Table 4.5.1 not showing any significant changes in the composition of opinions on sexual offenses over time, and particularly between 2017 and 2019, I can rule out the possibility that the movement caused substantial changes in the composition of opinions regarding categories of sexual offenses. To check for robustness of my estimates against the small changes in the sample composition regarding courts, the DiD analysis is complemented by an Inverse Probability Weighting (IPW) DiD approach (Abadie (2005)), in which I weight all control observations and the pre-movement sexual offense opinions to have the same distribution of courts as the group of post-treatment sexual offense opinions. [4] Further, I conduct a second robustness check, building on the results of Levy and Mattsson (2021). The authors show that the #MeToo movement only affected the reporting of some types of sexual offenses, while others, namely sodomy and rape, were not affected. I therefore re-run my analyses, considering only opinions on sexual offenses where the number of reports was not affected by the #MeToo movement.

However, there may still be #MeToo-induced shifts in the composition of cases that cannot be controlled for: the composition of offenses of a given type could still change in terms of the strength of the evidence and/or the severity of the offense. This would be the case if the #MeToo movement had led to an increase in trial court cases involving sexual offenses and, at the same time, a decrease in the proportion of convicted offenders who appeal their convictions. In this case, the proportion of appeal cases with inconclusive evidence may have increased, which could have accelerated the development of language in the sex offense sample and thus biased upward the estimated effects of movement on language development, i.e., the effect estimates from the text vectorization approaches. In contrast, the estimates for the impact of the movement on the use of victim-blaming language would in this scenario constitute lower bounds of the actual decline in the use of such language, since the more reasons there are to doubt the credibility of the victim or the seriousness of the incident, the more likely it is that victim-blaming language will be used.

In the DiD approach, I control for court fixed effects, since the court- or state-specific laws and rules, as well as the terminology therein, are likely to differ. In addition, I control for the word count to account for the fact that there are some very short and formal opinions in the sample that have little or no flexibility in how they are written. However, I also report the estimates for when no controls are included. The reported standard errors are heterogeneity-robust and

[4] IPW based on offense categories would be problematic in that the offense categories of the control and treatment groups do not overlap by design.

clustered at the court level (Zeileis (2004), Bertrand, Duflo, and Mullainathan (2004)).

Event Study Approach

I complement the DiD analysis with a simple event study approach, in which I assess the development of the text quantifiers before and after the #MeToo movement in a panel setting. In doing so, I attempt to address the problem that the set of opinions on non-sexual offenses may represent an imperfect control group, e.g., because of generally faster evolving language in opinions on sexual offenses. In the event study approach, I examine the judge-specific developments in the text quantifiers and victim blaming language indicators before and after the onset of the #MeToo movement and intentionally chose not to include opinions on non-sexual offenses as control observations, as this part of the study aims to deal with possible criticism of the choice of control group. Thus, the purpose of this event study application is not to identify the causal effect of the movement, but rather to observe whether there was a shift in the overall development of language in sex offense cases and the use of victim-blaming language from before to after the movement.

To apply the event study approach, I took advantage of the fact that for 9 courts, accounting for roughly 30 % of the opinions in my sample, the names of the judge who wrote the opinion is provided in the CourtListener database. For the remaining sexual offense opinions, I obtained the judge's name based on a semi-automatic approach, i.e., by formulating court-specific `regex`-rules to identify the authoring judge, which I then checked manually based on the context from which they were drawn. In cases where more than one judge is named as the author of an opinion, the name mentioned first is selected as the authoring judge, as is the case in the CourtListener database. Opinions written per curiam, i.e., in the name of the court rather than the judge(s), as well as opinions in which no author is named are excluded from the sample in the event study setting (they account for less than 1% of sexual offense opinions).

The names of the so identified judges are cleaned up, i.e., spellings of the same name and title in opinions of the same court are aligned. Of course, the entire process of identifying the authoring judge has many potential sources of error. For one thing, there could be two judges with the same name in the same court whose opinions will be attributed to one and the same individual. Second, when an opinion is authored by more than one judge, the order in which the judges are named does not necessarily say anything about the writing share of the judges, thus the judge named first and selected by me is not necessarily the primary author. It can be stated, however, that in the vast majority of opinions one single judge is named as the author and opinions with more than one authoring judge are the exception. Finally, both, the judge names identified by CourtListener and those identified by me may occasionally be incorrect. However, since there is no reason to believe that the number of erroneous judge names is time-dependent, this solely affects the estimation by introducing some additional noise, but does not lead to a systematic bias in the estimators.

I identify 1382 judges, who, on average, publish roughly 11 sexual opinions during the observation period, with the median number of opinions being 4, i.e., there are a few judges who authored a large amount of opinions (up to 184 opinions) while many others only published a handful opinions during the study period. For each judge in the sample, I calculate the six-month average of each text quantifier and victim blaming indictator[5] to obtain a panel data structure with one or no observation per individual and time period. To avoid losing too many observations, I do not exclude all observations for which data is missing in any time period, but keep all judges who published at least one observation before the onset of the #MeToo movement and at least one observation a year or more after the start of the movement. Then, I apply a Fixed Effects (FE) approach while weighting the observations by the number of opinions they were calculated from. Through weighting the observations, I account for the fact that the judges differ greatly in how many sex offense opinions they publish per six-month period, which makes some judges much more important for the development of language in court than others.

Although some information is lost by averaging the observations per judge and 6-month period, the panel approach might eliminate some of the noise typical of text analyses by increasing the amount of text per observation. On the other hand, however, it requires me to exclude several observations from the sample (3843 opinions authored by 785 judges) because I do not have observations from either before or after MeToo for the authoring judges. By excluding judges who, for whatever reason, do not frequently publish precedential opinions on sex offenses, important information may be lost. Further, this panel approach captures only part of the evolution of language in court. Changes due to retirement or dismissal, and replacement of judges are obviously not captured in this panel approach.

[5] The 6-month text quantifiers are calculated as the 6-month means of the judge's opinions from the respective court's opinion vectors in H1-2015, expressed relative to the median distance. The 6-month victim blaming indicator is calculated as the sum of mentions of the victim as the subject of a sentence divided by the total number of mentions of the victim in each 6-month period.

4.5.3 Assessing Effect Heterogeneity

Finally, I also assess whether there is evidence of effect heterogeneity with respect to a judge's gender or political affiliation, as well as with regard to the political orientation of the state in which a court is located. This is because different studies on victim blaming and rape myths acceptance show that females are less likely to accept rape myths that shift the blame for an assault upon the victim (e.g., Pinciotti and Orcutt (2021), Russell and Hand (2017), Davies, Rogers, and Whitelegg (2009), Schneider, Mori, Lambert, and Wong (2009)). Boux (2016) finds that this is particularly true for female Democrats.

Most studies in which participants were confronted with a sexual assault scenario found that men were more likely than women to blame the victim and show signs of rape myth acceptance, while other studies found no significant effect of gender on the likelihood of victim blaming (see Gravelin, Biernat, and Bucher (2019), Grubb and Turner (2012) and Suarez and Gadalla (2010) for reviews). Other research shows that study participants with politically conservative views are more likely to (partially) blame the victim for a sexual assault (Anderson, Cooper, and Okamura (1997), Lambert and Raichle (2000)). For the judicial context, Boux (2016) finds that female Democratic judges are less likely to use rape myths than male judges (regardless of political affiliation), while her results show no significant difference between female Republican judges and male judges.

In addition to the differences in victim blaming and rape myth acceptance noted above, there is also evidence that the perception of and reaction to the #MeToo movement differ across genders and political camps. Castle, Jenkins, Ortbals, Poloni-Staudinger, and Strachan (2020) found in a poll that Democrats were more likely than Republicans to say they were aware of the movement and mobilized by it. An analysis of the members of Congress' communications on their public Facebook pages reveals that in the wake of the #MeToo movement, far more female than male members addressed the issue of sexual violence in their posts, with this pattern evident in both political parties (Anderson and Toor (2018)).

In light of these findings, it is interesting to assess whether judges' language in court is affected differently by the #MeToo movement depending on their gender and political affiliation. The research cited above suggests that female and/or politically liberal judges were more receptive to the #MeToo movement and more willing to change their behavior. Then again, these judges

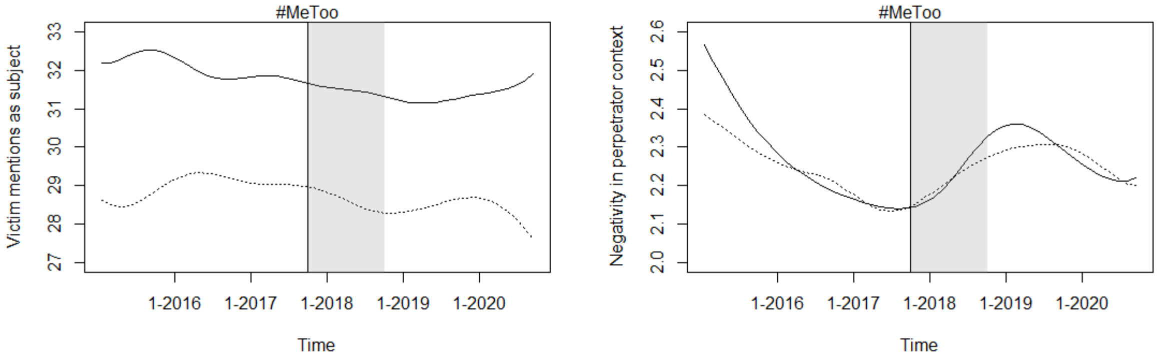


Figure 4.6.1: Kernel-smoothed plot of victim blaming indicators, with the solid line representing sexual offense opinions and the dotted line representing the control group. Curves were smoothed using the default settings of the `sm.regression` function in R. The semantics indicator measures mentions of the victim as a subject as a percentage of total mentions, and the sentiment indicator measures the negativity score of the context words of mentions of the offender, ranging from 0 to 100, with the words being more negative the higher the score.

may have been more cautious in their choice of words prior to the movement and may have already avoided language that implied victim blaming, leaving them little room for change toward language that attributed less blame to the victim.

The analysis of treatment effect heterogeneity by judge characteristics is restricted to courts for which the names of the authoring judge is available on CourtListener. Information on the judges' gender and political affiliation is obtained from ballotpedia, an online encyclopedia on American politics and elections. Ballotpedia only provides the political affiliation of some judges. To determe the political affiliation of the other judges, I draw on the party for which the judge ran in the judicial election or the political affiliation of the politician who appointed the judge, depending on the process used to select judges. For about 21% of the judges no political affiliation can be determined. To assess effect heterogeneity with respect to a state's political orientation, I categorize as predominantly Democratic (Republican) those states that were won by the Democratic (Republican) party in at least three of the four 2008-20 presidential elections. Opinions from swing states and courts at the supra-state level are excluded for this analysis.

4.6 Results

4.6.1 Victim Blaming Indicators

Figure 4.6.1 shows the development of the victim blaming indicators in opinions on sexual offenses and on other crimes against persons. Neither plot suggests any substantial differences

	Share of Victim as Subject			Negativity of Offender Context		
sex cr. x post	-0.687 (1.087)			-0.079 (0.056)		
sex cr. x '19		-0.446 (1.088)			0.002 (0.064)	
sex cr. x '20		-0.872 (1.340)			-0.087 (0.057)	
sex cr. x H1-'19			-1.214 (1.528)			0.000 (0.062)
sex cr. x H2-'19			0.273 (1.144)			0.003 (0.117)
sex cr. x H1-'20			-1.155 (1.407)			-0.107 (0.098)
sex cr. x H2-'20			-0.621 (1.587)			-0.069 (0.098)
post	X			X		
year FE		X			X	
half-year FE			X			X
court FE	X	X	X	X	X	X
# words	X	X	X	X	X	X

Table 4.6.1: DiD estimates of effect heterogeneity for victim blaming indicators. Significance levels: * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

between the development of these two indicators in the treatment and the control group. Further, the DiD estimates provided in Table 4.6.1 do not indicate a significant impact of the #MeToo movement on either of the two indicators. The DiD estimates for the semantics indicator suggest a slight, but not statistically significant, #MeToo-induced decline in the number of victim mentions as a grammatical subject and hence a decline in victim blaming. The estimates for the sentiment indicator are also not statistically significant and even indicate, contrary to the research hypothesis, a decrease in the use of words with negative connotations in the context of mentions of the perpetrator. Likewise, the event study approach (see Figure 4.6.2) shows no substantial changes in either of the two victim blaming indicators.

While assessing the impact of the #MeToo movement on the victim blaming indicators does not yield significant results, the plot on the development of the semantics estimator is nevertheless interesting. It shows that the use of the victim as grammatical subject is generally substantially higher in opinions on sexual assault cases than in cases on other crimes against persons. Given that several studies indicate higher prevalence of victim blaming in sexual assault cases than in cases on other crimes against persons (see Section 4.2.2) and that the indicator is constructed based on scientific findings (see Section 4.4.1), it seems reasonable to explore

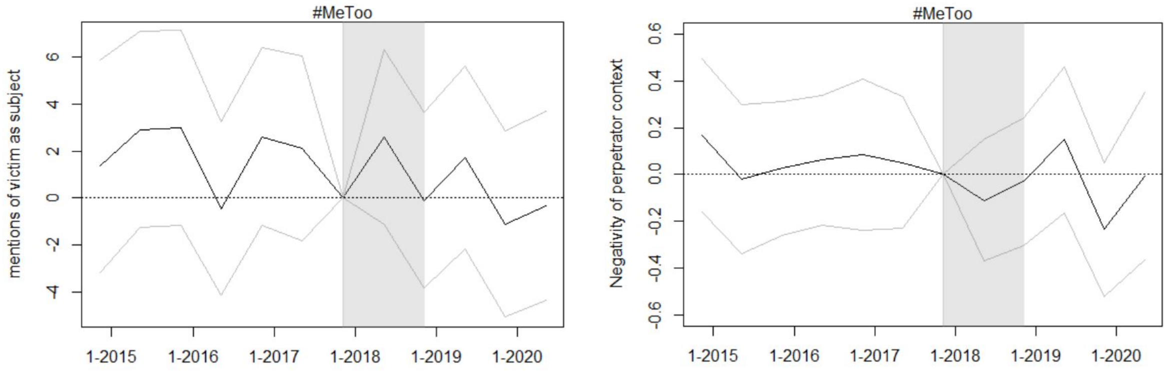


Figure 4.6.2: Event study estimates for the evolution of the victim blaming indicators in sexual offense opions. The gray lines indicate the 90 percent confidence interval around the estimates.

whether this indicator may be useful for measuring the extent of victim blaming in judicial opinions - as long as the purpose is to classify opinions or quantify the status quo rather than measure changes over time. The sentiment indicator, on the other hand, seems to vary little both over time and crime types. It generally takes very low values, which may be because judges deliberately avoid sentiment-charged language. It therefore does not seem appropriate for use in the context of court opinions and may be better suited for contexts with more colloquial language such as social media.

The results of the robustness checks for the DiD with victim blaming indicators are provided in Appendix 4.B. Given that no significant effects on either indicator can be identified using the DiD approach, the fact that the placebo test does not indicate a violation of the parallel trend assumption is not particularly informative. For the sentiment indicator, both the IPW estimates as well as the DiD estimates based on sodomy and rape cases only indicate a significant decrease in negatively connoted words in the context of mentions of the perpetrator. In the case of the semantics indicator, both estimates have different signs and are not statistically significant.

The effect heterogeneity estimates obtained using the DiD and event study approach, respectively, can be found in Appendices 4.C and 4.D. For the semantics indicator, the estimates suggest a larger decline in mentions of the victim as subject among female and Democratic judges as well as in predominately Democratic states, although again the differences are not statistically significant. The DiD estimates for the sentiment indicator imply a greater decline in the use of words with negative connotations among females and no difference in the effect of #MeToo on this indicator between Democrates and Republicans. The event study approach, on the other hand, suggests that Democrates have increased their use of negatively connoted words

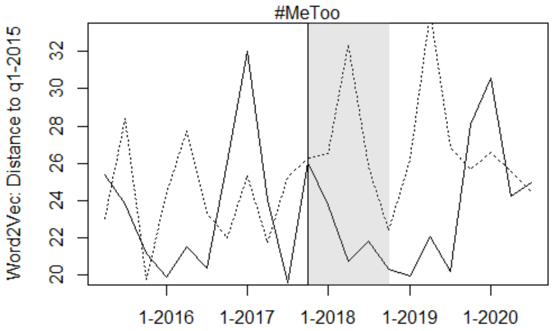


Figure 4.6.3: Development of the word2vec representation of the word "victim", with the solid line representing sexual offense opinions and the dotted line representing the control group.

when compared to the development of this indicator among Republicans.

Finally, Figure 4.6.3 displays the development of the word2vec representation of the word "victim" in both the control and treatment group. Contrary to the research hypothesis, the context in which victim is mentioned does not evolve faster for opinions on sexual offenses than for those on other crimes against persons. Again, this approach may be better suited to contexts with more flexible and rapidly evolving language.

4.6.2 Text Vectorization

Figure 4.6.4 illustrates the evolution of the BoW[6] and the tf-idf text quantifiers in opinions on sexual offenses and on other crimes against persons. Both charts suggest that language in opinions on sexual offenses evolves more rapidly than that in opinions on other crimes against persons between the onset of the movement and 2019. However, the graphs also indicate that there may be problems with the parallel trend assumption. Moreover, the BoW graph shows a narrowing of the distance between the opinion vectors and their H1-2015 average in 2016 and 2017, as well as in 2020, which could be due to changes in the composition of courts, but also to other unobservable factors, which in turn would be critical for identifying the causal effect [7].

The DiD estimates for the text vectorization-based opinion quantifiers can be found in Table 4.6.2. The results point at a slight #MeToo-induced change in courtroom language, which

[6] The reduced sample BoW quantifier evolves similarly to the BoW quantifier in both groups, which is why it is not shown here.

[7] The use of other smoothing methods and smaller bandwidths yields similar curves. Thus, it does not seem to be an (over-)smoothing issue.

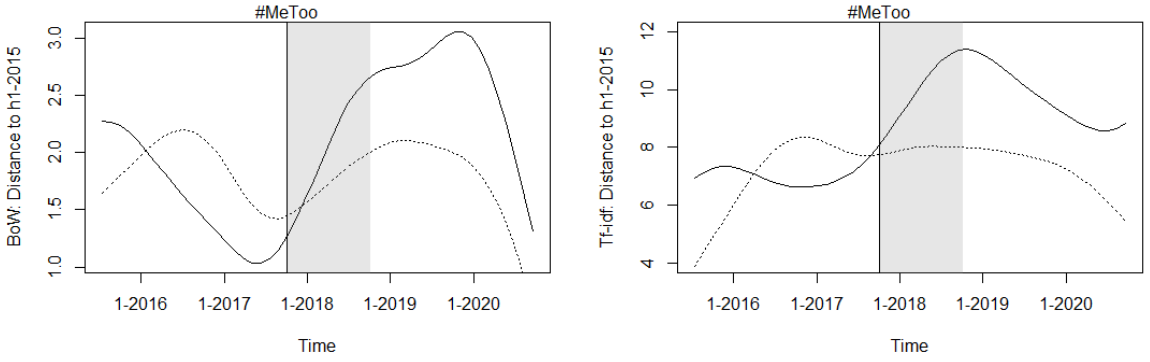


Figure 4.6.4: Kernel-smoothed plot of the distance of text vectors from their H1 2015 average, with the solid line representing sexual offense opinions and the dotted line representing the control group.

however materializes with a substantial time lag. The estimates suggest that the language in sexual offense opinions deviates more quickly from the 2015 average than in opinions on other cases of crimes against persons. The DiD estimate for the second half of 2020 is statistically insignificant for all three text quantifiers and numerically small for the two BoW quantifiers, indicating that the langugage change in sexual offense opinions is not likely due to the COVID-19 crisis.

The event study estimates in Figure 4.6.5 show a decrease in the distance to the H1-2015 average in 2016 that is similar to, though less pronounced than, that observed in Figure 4.6.4. Further, the event study estimates do not indicate a stronger deviation of language from the H1-2015 average in the years after #MeToo than in the years before #MeToo, suggesting that the effect estimated with the DiD approach may be attributable to changes in case composition or personnel rather than changes in judges' attitudes.

The results of the robustness checks are presented in Appendix 4.B. While the placebo tests do not reveal a significant violation of the common trend assumption, they do not allow me to rule out such a violation, especially because the estimated effects of the placebo treatment on the (reduced sample) BoW quantifiers are positive, just like the observed effect in the main DiD analysis. I therefore also estimated the effect of placebo treatments at other points in time during the pre-#MeToo period, all of which turned out to be statistically insignificant, with some of them having a positive and others a negative sign. The fact that the IPW-based DiD and the reduced sample DiD estimate positive effects for all quantifiers, some of which

		BoW			Reduced Sample BoW			tf-idf	
sex cr. x post	0.815*			0.700			1.665		
	(0.469)			(0.511)			(1.562)		
sex cr. x '19		0.629			0.269			1.338	
		(0.535)			(0.507)			(1.732)	
sex cr. x '20		0.919**			1.165**			1.106*	
		(0.454)			(0.496)			(0.646)	
sex cr. x H1-'19			0.658			0.474			0.940
			(0.715)			(0.718)			(1.835)
sex cr. x H2-'19			0.603			0.079			1.716
			(0.563)			(0.564)			(1.813)
sex cr. x H1-'20			1.461			1.806*			0.281
			(0.905)			(0.955)			(0.966)
sex cr. x H2-'20			0.317			0.459			1.958
			(0.692)			(0.670)			(1.312)
post	X			X			X		
year FE		X			X			X	
half-year FE			X			X			X
court FE	X	X	X	X	X	X	X	X	X
# words	X	X	X	X	X	X	X	X	X

Table 4.6.2: DiD estimates for text vectorization-based opinion quantifiers. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

are statistically significant, support the finding that the #MeToo movement has caused a slight increase in language development in sexual assault opinions, with the IPW results ruling out that the observed effect in the main analysis is due to changes in the composition of courts, while the latter rules out that it is due to an increase of reports of sexual offenses. When considering these results in conjunction with the event study results, one possible explanation for the observed slight linguistic change in sexual offense opinions is the change in judicial appointments toward judges with more progressive views on sex offenses. However, given the numerically small effect estimates, the noise in these quantifiers, and the fact that the effect does not appear until two years after the movement began, it is difficult to attribute the observed effect to the #MeToo movement.

A look at the effect heterogeneity estimates in Appendix 4.D and 4.C does not give a clear picture. While the DiD estimates for the (reduced sample) BoW suggest a faster evolution of language in opinions written by females and Democrates, the event study plots point to a similar evolution of language in opinions written by female and male or Democratic and Republican judges, respectively. For the tf-idf approach, the DiD estimates suggest that the language of female and Democratic judges evolves less rapidly than that of their counterparts.

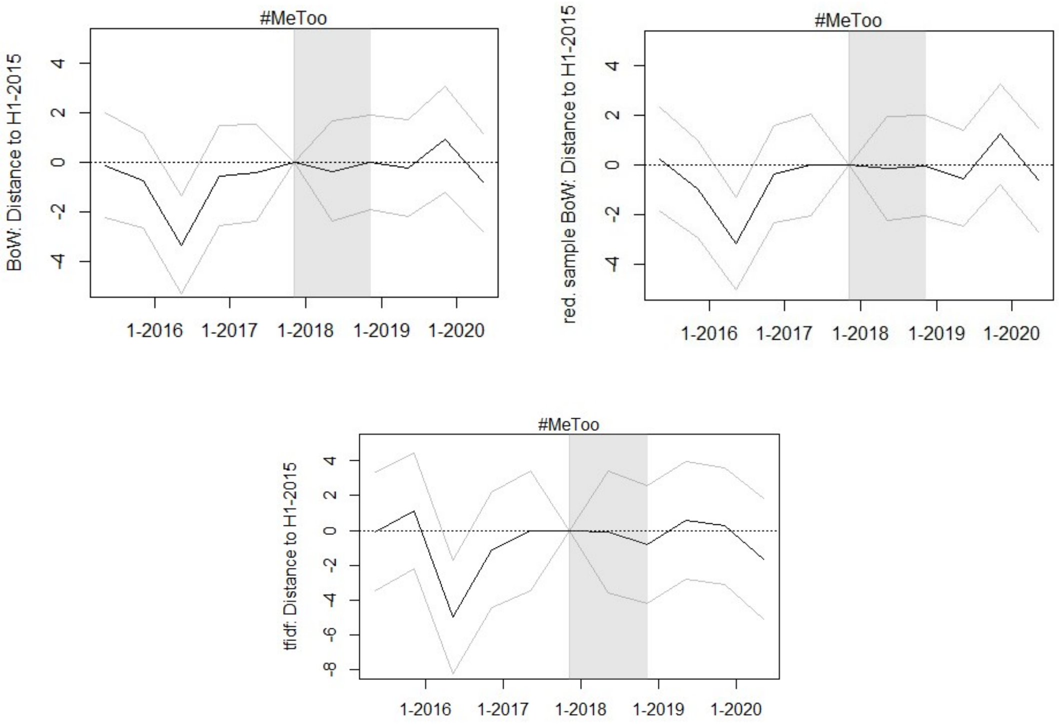


Figure 4.6.5: Event study estimates for the evolution of the text vectorization-based opinion quantifiers. The gray lines indicate the 90 percent confidence interval around the estimates.

4.7 Conclusion

In this study, I quantified judicial opinions by means of different indicators and text vectorization methods to assess how the #MeToo movement has affected the evolution of language in sexual assault opinions and whether it has led to a decrease in victim blaming in such cases. Although I do not obtain statistically significant estimates for the impact of the movement on most quantifiers, with a few exceptions of mildly statistically significant estimates, the point estimates suggest a faster evolution of language in sexual assault opinions as well as a decline in victim blaming. The reasons for not obtaining statistically significant results may be manifold. For one, the language in judicial opinions is generally not very flexible, i.e., there is regulation on what different parties are allowed to say in court and the structure of court opinions, particularly in certain paragraphs, is highly formalized. Therefore, any treatment should be expected to have a smaller effect on language in court opinions than on language in any other more flexible text body. Moreover, the text vectorization methods in particular, but also the victim blaming indicators, capture a lot of noise, leading to large standard errors in the estimators.

Despite not revealing strong statistically significant effects, the study may be a valuable contribution to the growing literature on text-based causal inference: for one, I have developed indicators that can be useful proxies for victim blaming and may be used as a treatment, outcome or control, as long as their purpose is either to measure the status quo rather than development over time, or they are applied to a text body with more flexible language. Then, I have also introduced an approach to quantifying language development in a body of text that is based on text vectorization methods originally designed for categorization and clustering purposes. This enables the use of text vectorization methods in the context of DiD analyses or panel data methods. Again, this approach may be useful for assessing text bodies with more flexible language, analyzing larger text bodies, or evaluating longer-term effects of a treatment in a panel setting.

Appendix

4.A Descriptives

Court	# Non-sexual Offenses	# Sexual Offenses
Appellate Court of Illinois	668	192
Army Court of Criminal Appeals	10	46
California Court of Appeal	912	361
Commonwealth Court of Pennsylvania	247	114
Connecticut Appellate Court	436	164
Court of Appeals for the D.C. Circuit	60	24
Court of Appeals for the Eighth Circuit	349	194
Court of Appeals for the Eleventh Circuit	179	64
Court of Appeals for the Federal Circuit	14	3
Court of Appeals for the Fifth Circuit	259	108
Court of Appeals for the First Circuit	147	70
Court of Appeals for the Fourth Circuit	167	61
Court of Appeals for the Ninth Circuit	242	98
Court of Appeals for the Second Circuit	115	44
Court of Appeals for the Seventh Circuit	367	134
Court of Appeals for the Sixth Circuit	188	95
Court of Appeals for the Tenth Circuit	137	51
Court of Appeals for the Third Circuit	98	36
Court of Appeals of Alaska	47	31
Court of Appeals of Arizona	43	31
Court of Appeals of Arkansas	192	156
Court of Appeals of Georgia	403	375
Court of Appeals of Iowa	1244	680
Court of Appeals of Kansas	48	47
Court of Appeals of Minnesota	1	19
Court of Appeals of Mississippi	426	197
Court of Appeals of North Carolina	274	147
Court of Appeals of Tennessee	121	85
Court of Appeals of Texas	4082	2576
Court of Appeals of Virginia	92	39
Court of Appeals of Washington	204	137
Court of Criminal Appeals of Tennessee	1949	797
Court of Criminal Appeals of Texas	252	120
District Court of Appeal of Florida	1123	324
District Court, District of Columbia	366	133
District of Columbia Court of Appeals	154	34
Idaho Court of Appeals	48	42
Indiana Court of Appeals	1781	1146
Massachusetts Appeals Court	45	98
Michigan Court of Appeals	85	67
Missouri Court of Appeals	354	285
Navy-Marine Corps Court of Criminal Appeals	76	252
Nebraska Court of Appeals	99	113
New Jersey Superior Court	65	35
New Mexico Court of Appeals	60	33
New York Court of Appeals	93	40
Ohio Court of Appeals	3283	2091
Superior Court of Delaware	159	36
Superior Court of Pennsylvania	5995	3246
United States Air Force Court of Criminal Appeals	3	28
United States Court of Federal Claims	19	8

Table 4.A.1: Number of opinions per court.

4.B DiD: Robustness Checks

4.B.1 Victim Blaming Indicators

	Victim as Subject			Neg. of Offender Context		
	(1)	(2)	(3)	(1)	(2)	(3)
sex cr. x placebo	-0.716			-0.012		
	(0.810)			(0.066)		
sex cr. x post		-0.939			-0.100*	
		(0.926)			(0.058)	
sex cr. x post			1.399			-0.093**
			(1.564)			(0.042)
post	X	X	X	X	X	X
court FE	X	X	X	X	X	X
# words	X	X	X	X	X	X

Table 4.B.1: Robustness tests: (1) DiD was performed using only pretreatment observations and a placebo treatment in the middle of the pretreatment period; (2) DiD with IPW based on court distribution, performed for the entire sample using the `didweight` function from the `causalweight` package in the statistical software R (R Core Team (2022)); and (3) DiD performed only for the sample of sodomy and sexual assault cases. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

4.B.2 Text Vectorization

	BoW			Reduced Sample BoW			tf-idf		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
sex cr. x placebo	0.220			0.166			-0.941		
	(0.584)			(0.675)			(0.831)		
sex cr. x post		0.535			0.442			1.697**	
		(0.424)			(0.438)			(0.821)	
sex cr. x post			0.806			0.836*			4.148*
			(0.534)			(0.505)			(2.254)
post	X	X	X	X	X	X	X	X	X
court FE	X	X	X	X	X	X	X	X	X
# words	X	X	X	X	X	X	X	X	X

Table 4.B.2: Robustness tests: (1) DiD was performed using only pretreatment observations and a placebo treatment in the middle of the pretreatment period; (2) DiD with IPW based on court distribution, performed for the entire sample using the `didweight` function from the `causalweight` package in R; and (3) DiD performed only for the sample of sodomy and sexual assault cases. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

4.C DiD: Effect Heterogeneity

4.C.1 Victim Blaming Indicators

	Victim as Subject			Neg. of Offender Context		
female x sexual x post	-1.103			-0.236		
	(3.329)			(0.394)		
dem. judge x sexual x post		-2.099			-0.015	
		(3.729)			(0.225)	
dem. state x sexual x post			-0.276			-0.106
			(2.967)			(0.162)
post	X	X	X	X	X	X
court FE	X	X	X	X	X	X
# words	X	X	X	X	X	X

Table 4.C.1: DiD estimates of effect heterogeneity for victim blaming indicators. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

4.C.2 Text Vectorization

	BoW			Reduced Sample BoW			tf-idf		
female x sexual x post	0.891			1.170			-1.740		
	(1.171)			(1.194)			(2.207)		
dem. judge x sexual x post		0.340			0.764			-4.144	
		(1.938)			(2.145)			(2.993)	
dem. state x sexual x post			0.357			-0.152			2.169
			(1.335)			(1.207)			(1.841)
post	X	X	X	X	X	X	X	X	X
court FE	X	X	X	X	X	X	X	X	X
# words	X	X	X	X	X	X	X	X	X

Table 4.C.2: DiD estimates of effect heterogeneity for text vectorization-based opinion quantifiers. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

4.D Event Study Approach: Effect Heterogeneity

4.D.1 By Gender

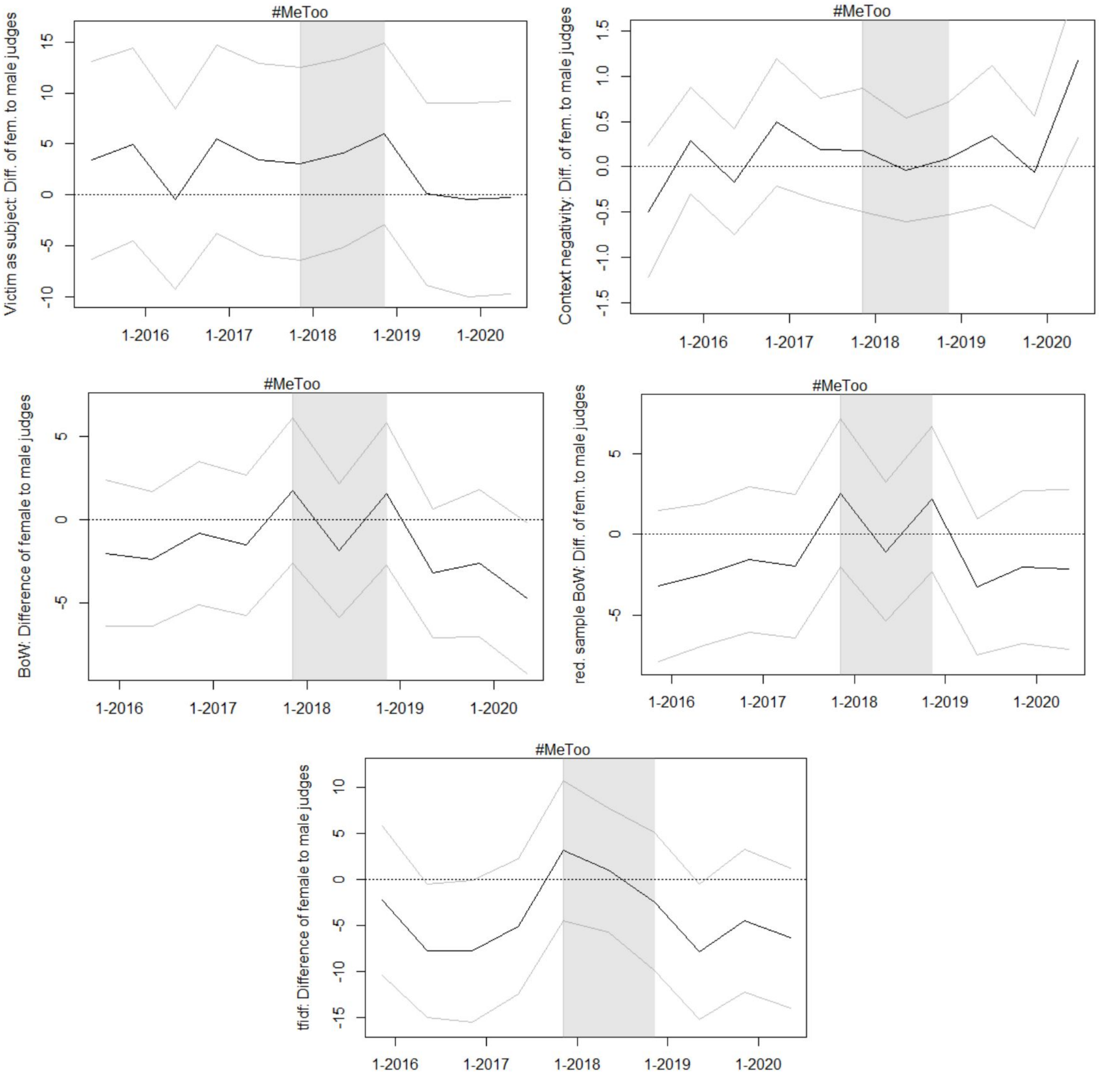


Figure 4.D.1: Event study estimates of the difference in indicator development of female vs. male judges. The black line represents the FE estimates for the interaction terms of 6-month period identifiers and a dummy variable indicating whether a judge is female. The gray lines indicate the 90 percent confidence interval around the estimates.

4.D.2 By Political Affiliation

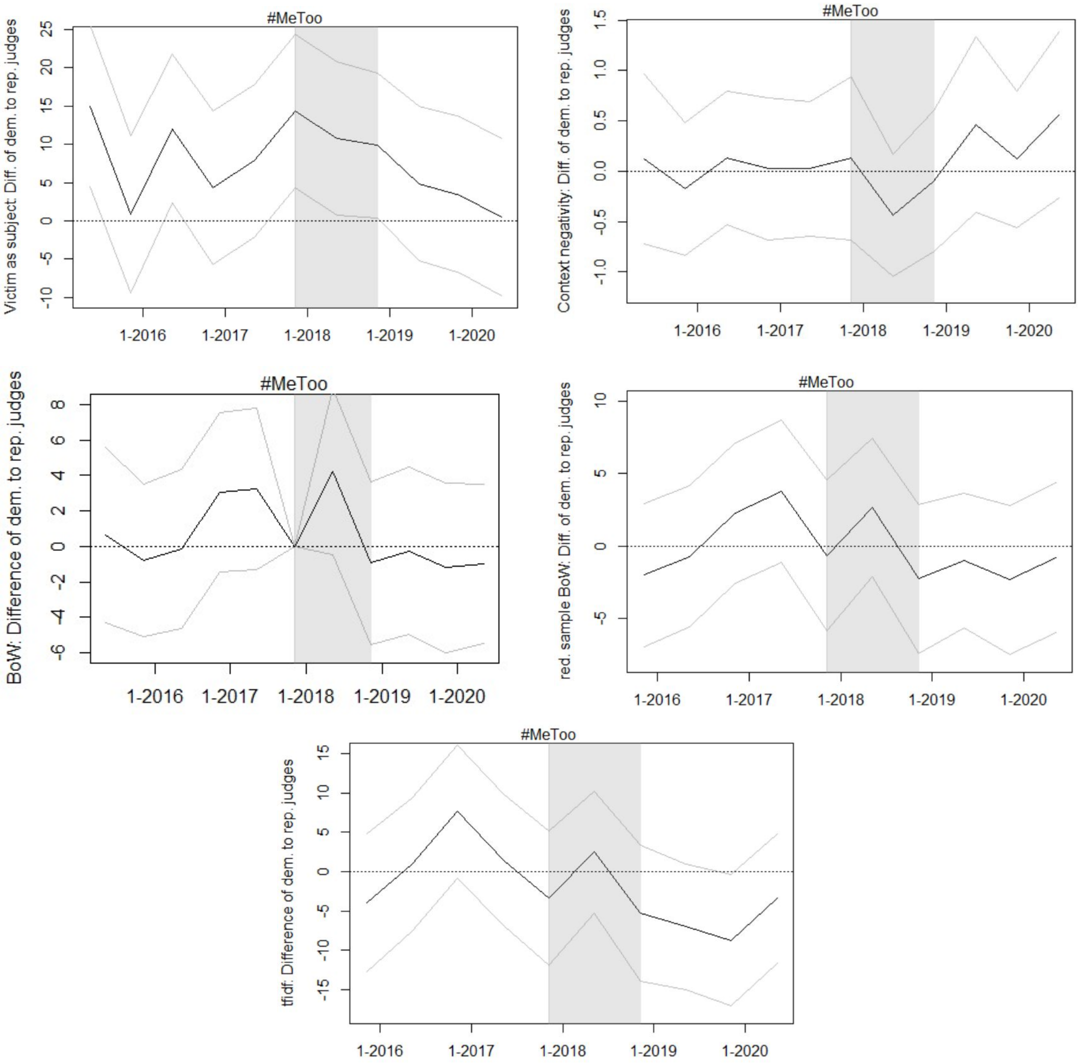


Figure 4.D.2: Event study estimates of the difference in indicator development of Democratic vs. Republican judges. The black line represents the FE estimates for the interaction terms of 6-month period identifiers and a dummy variable indicating whether a judge is affiliated with the Democratic party. The gray lines indicate the 90 percent confidence interval around the estimates.

Chapter 5

How causal machine learning can leverage marketing strategies: Assessing and improving the performance of a coupon campaign

with Martin Huber

Abstract: We apply causal machine learning algorithms to assess the causal effect of a marketing intervention, namely a coupon campaign, on the sales of a retailer. Besides assessing the average impacts of different types of coupons, we also investigate the heterogeneity of causal effects across different subgroups of customers, e.g., between clients with relatively high vs. low prior purchases. Finally, we use optimal policy learning to determine (in a data-driven way) which customer groups should be targeted by the coupon campaign in order to maximize the marketing intervention's effectiveness in terms of sales. We find that only two out of the five coupon categories examined, namely coupons applicable to the product categories of drugstore items and other food, have a statistically significant positive effect on retailer sales. The assessment of group average treatment effects reveals substantial differences in the impact of coupon provision across customer groups, particularly across customer groups as defined by prior purchases at the store, with drugstore coupons being particularly effective among customers with high prior purchases and other food coupons among customers with low prior purchases. Our study provides a use case for the application of causal machine learning in business analytics to evaluate the causal impact of specific firm policies (like marketing campaigns) for decision support.

Keywords: causal machine learning, coupon campaign, marketing.

JEL classification: M30, C21.

This article is published as working paper on arXiv.

5.1 Introduction

Over the last two decades, the amount of customer data available to marketers has increased dramatically with new data types such as social media, clickstream, search query and supermarket scanner data on the rise. The increasing availability of customer Big Data has spawned a new stream of literature on machine learning (ML) methods and tools in the field of business and marketing. The ML literature on designing marketing campaigns ranges from research on modeling customer behavior (e.g. Xia, Chatterjee, and May (2019), Hu, Dang, and Chintagunta (2019)), price sensitivity (e.g. Arevalillo (2021)) and purchase decisions (e.g. Donnelly, Ruiz, Blei, and Athey (2021)) to studies on the development of personalized product recommendation systems (e.g. Ramzan, Bajwa, Jamil, Amin, Ramzan, Mirza, and Sarwar (2019), Anitha and Kalaiarasu (2021)), customer churn management (e.g. Gordini and Veglio (2017)) and acquisition of new customers (e.g.Luk, Choy, and Lam (2019)).

A common feature of these studies is that they are based on predictive ML, i.e., on identifying patterns of variables in the data in order to use them for predicting an outcome of interest (e.g., sales). This is done by training predictive models in one part of the data and determining the best performing model (with the smallest possible prediction error) in the other part of the data. Under some commonly used ML algorithms, the identified model serves as a black box, i.e., it is based on functions that are too complex for any human to understand (as in so-called deep learning), while in other cases, the model has an explicit (and thus comprehensible) structure. In any case, however, such predictive ML models generally do not provide insights into the causal effects of specific variables or interventions (such as a marketing campaign) on the outcome of interest. Thus, predictive ML, although appropriate for making educated guesses about outcomes based on certain patterns observed in the data, is not well suited for determining and comparing the effectiveness of possible courses of action, which would be relevant for decision support, e.g. for optimally designing a marketing campaign.[1]

[1] To predict an outcome of interest based on predictor variables, ML aims at minimizing the prediction error by optimally trading off prediction bias and variance. When multiple variables capture the same relevant predictive feature, i.e., are correlated with that feature, ML algorithms may identify some of these variables as relevant predictors while attaching little importance to others, regardless of the variables' causal effect on the outcome. For instance, variables that do not directly or only modestly affect the outcome may enter the prediction model as relevant predictors, simply because they are correlated with other variables that actually affect the outcome. For this reason, it may happen that these other variables play little or no role in the predictive model, even though they have a causal impact on the outcome, simply because they provide little additional information for the prediction. Therefore, predictive ML is generally not suitable for the causal analysis of 'what if' questions,

To improve on the shortcomings of predictive ML in evaluating the impact of implementing vs. not implementing a specific intervention, a fast growing literature in econometrics and statistics has been developing so-called causal ML algorithms. In this paper, we demonstrate the application of such methods in the context of business analytics for decision support, that is, for evaluating a marketing intervention. More precisely, we make use of the so-called causal forest approach by Athey, Tibshirani, and Wager (2019) to assess the causal effect of marketing campaigns, in which customers were provided coupons for different product types, on customers' purchasing behavior, i.e. the difference in their expected behavior with and without being targeted by a coupon campaign. While predictive ML algorithms are not able to isolate the causal effects of coupons on customers' purchasing behavior from the influence of background characteristics (e.g. socio-economic characteristics and price sensitivity) which jointly influence coupon reception and purchasing behavior, the causal forest approach can do so under certain assumptions.

One crucial condition is that all variables that jointly affect coupon reception and purchasing behavior are observed in the data and can thus be controlled for. This condition is known as selection-on-observables or unconfoundedness assumption. Under further conditions on the quality of the ML models estimated as part of the causal forest approach for predicting purchasing behavior and coupon reception as a function of the observed variables, the causal forest approach permits evaluating the mean impact of the coupons on all customers, as well as across specific subgroups or customer segments (e.g. different age groups). Our results suggest, for instance, a positive overall effect of coupons for drugstore items. For coupons applicable to ready-to-eat food as well as meat and seafood, on the other hand, we do not find a statistically significant overall effect. An analysis of the effect of drugstore coupons across different customer subgroups reveals that these coupons particularly affect customers with high pre-campaign spending as well as low- to middle-income customers.

Furthermore, we apply optimal policy learning based on ML as proposed by Athey and Wager (2021), in order to learn from the data which customer segments should be optimally targeted by coupon campaigns such that the overall (or net after cost) effect is maximized. In contrast to predictive ML, optimal policy learning allows, under certain conditions, identifying the coupon provision policy which is most effective in terms of its impact on sales. This is done

such as how a change in a coupon campaign strategy will affect customer behavior.

by first assessing the expected effects in different customer segments and then selecting those segments as target groups in which the effects are sufficiently high. The estimated optimal policy for coupons applicable to meat and seafood, for instance, suggests that such coupons should be issued to low-income customers whose pre-campaign spending did not exceed a certain level, to middle-to-high-income customers aged 46 years or older who purchased something from the store in the period prior to the campaign, as well as to middle-to-high-income households with at least five members who did not purchase anything from the store in the pre-campaign period.

The paper proceeds as follows: Section 5.2 outlines the current state of quantitative research in the marketing literature and motivates the application of causal ML methods in the field of marketing. Section 5.3 introduces and describes the retailer's sales data to be analyzed. Section 5.4 defines the causal effects of interest based on so-called counterfactual reasoning, discusses the conditions required for applying causal ML (such as the selection-on-observables assumption) and describes the algorithms for causal analysis and optimal policy learning. Section 5.6 provides the results of the evaluation of the retailer's coupon campaigns as well as the optimal coupon allocation. Section 5.7 concludes.

5.2 Motivation

The evaluation of the causal impact of discount campaigns plays a significant role in the earlier marketing literature from the 'pre-Big-Data era', see e.g. Inman and McAlister (1994), Raju, Dhar, and Morrison (1994), Leone and Srinivasan (1996) and Krishna and Zhang (1999) for studies on causal effects of coupon provision. However, the last two decades have seen a surge of predictive ML applications in business analytics, which appear to increasingly dominate causal analysis in marketing as well. In a keyword-search-based literature review, Mariani, Perez-Vega, and Wirtz (2021) find that the number of publications on predictive ML and Artificial Intelligence (AI) in marketing, consumer research and psychology has grown exponentially in the past decade (2010-21). The systematic literature reviews by Mustak, Salminen, Plé, and Wirtz (2021) and Ma and Sun (2020) paint a similar picture, with the latter stating that the rise of ML in marketing began with applications of support vector machines, a specific type of ML algorithm. This was then followed by studies that introduced text analysis, topic modelling and reinforcement learning into marketing research, as well as by marketing applications of deep learning, and network embedding. Questions about the impact of marketing campaigns, the influence of certain external factors on the success of a campaign and the heterogeneity of campaign effects across customer segments appeared to become of comparatively less importance (see e.g. Hair Jr and Sarstedt (2021), Ma and Sun (2020)), even though most recently, the marketing literature saw first applications of causal ML alogithms (such as causal trees).

The following sections summarize the current state of research on discount campaigns using causal inference (Section 5.2.1) and predictive ML (Section 5.2.2). This serves as the basis for motivating the use of causal ML to evaluate and optimize discount campaigns and to approach various other marketing and business decisions in Section 5.2.3.

5.2.1 Causal Inference in Marketing

A number of studies assess the causal effects of specific marketing campaigns on consumer response to the campaigns. These studies typically rely on (field) experiments or traditional methods for causal inference based on observational data. In the latter case, researchers must assume that all variables that jointly affect the intervention and purchasing behavior are observed in the data and can thus be controlled for. Rubin and Waterman (2006) apply propensity score matching to evaluate the effect of marketing interventions aimed at physicians in order to promote the prescription of a 'lifestyle' drug. They also rank the physicians according to their estimated expected individual-level effects, which in turn can be used to derive a tailored marketing strategy. Reimers and Xie (2019) assess the effect of e-coupon provision on alcohol sales by means of a difference-in-differences approach, exploiting the fact that the restaurants in their sample issued e-coupons at different points in time. See also Xing, Zou, Yin, Wang, and Li (2020), Halvorsen, Koutsopoulos, Lau, Au, and Zhao (2016) and Zhang, Dai, Dong, Qi, Zhang, Liu, and Liu (2017) for further examples of observational and experimental studies examining the effect of coupon provision or other discount campaigns on consumer behavior.

Other contributions analyze the heterogeneity of marketing effects across customer characteristics and the circumstances under which customers are targeted by coupon and other promotional campaigns. Among them, Gopalakrishnan and Park (2021) investigate whether high- and low-consumption customers, as defined by their purchasing behavior during the 12 months prior to the experiment, differ in their responsiveness to coupon campaigns. Andrews, Luo, Fang, and Ghose (2016) study whether the level of occupancy (or crowdedness) of a subway affects passengers' response to mobile advertising campaigns and find a statistically significant positive association. Based on a field experiment, Spiekermann, Rothensee, and Klafft (2011) conclude that proximity to the location for which coupons are distributed influences coupon redemption, and that this association is much more pronounced in the city center than in suburban areas.

Furthermore, several studies evaluate how certain configurations of coupons, such as face value, distribution method and expiry date, affect consumer behavior. The experimental studies by Zheng, Chen, Zhang, and Che (2021) and Biswas, Bhowmick, Guha, and Grewal (2013) assess how the size of discounts affects consumers' perceptions of product quality and purchase intentions. Leone and Srinivasan (1996) use supermarket scanner data to analyze the effect of coupon face value on sales and profits, while Anderson and Simester (2004) study the long-term effects of discount size on the purchasing behavior of new and established customers in an experimental setting. Other contributions as e.g. Gopalakrishnan and Park (2021), Jia, Yang, Lu, and Park (2018), Choi and Coulter (2012), Krishna and Zhang (1999) and Inman and McAlister (1994) analyze how further aspects of coupon and discount campaign design affect consumer behavior.

5.2.2 Predictive ML in Marketing

In recent years, many studies have focused on ML-based prediction of coupon redemption and associated sales. They use ML algorithms to model customer behavior as a function of customers' previous transactions, their response to past coupon/discount campaigns and their socio-economic characteristics in order to predict the likelihood of customers to redeem coupons or take up discounts and make purchases.

Pusztová and Babič (2020) and He and Jiang (2017) compare the performance of different ML-based classification algorithms in predicting coupon redemption in digital marketing campaigns. The first study concludes that so-called Support Vector Machines provide the most accurate predictions, while the latter study shows that the gradient boosting framework 'XG-Boost' performs best. Greenstein-Messica, Rokach, and Shabtai (2017) introduce an algorithm that combines co-clustering and random forest classification to predict redemption of mobile restaurant coupons based on demographic and contextual variables such as the consumer's distance to the restaurant relative to the size of the coupon discount. Ren, Cao, Xu, et al. (2021) developped a two-stage model for estimating the probability of coupon redemption, consisting of a first stage in which customers are clustered based on their past purchase and redemption behavior, followed by a second stage of fitting prediction models for the different customer clusters. Furthermore, several studies such as Koehn, Lessmann, and Schaal (2020), Xiao, Li, Xu, Zhao, Yang, Lang, and Wang (2021) and Zheng, Chen, Zhang, and Che (2021) predict customer behavior in the context of coupon or other discount campaigns by means of several ML methods.

5.2.3 Causal ML in Marketing

Under certain conditions like the selection-on-observables assumption, implying that all variables that jointly affect the intervention and purchasing behavior are observed in the data and can thus be controlled for, causal ML methods allow for the evaluation of causal effects of coupon/discount campaigns as well as effect heterogeneity across customer segments. In contrast to more traditional methods of causal inference, they can leverage the full amount of information available to marketers, which may be large in the era of 'Big Data'. That is, causal ML can address research questions such as those described in Section 5.2.1 based on high-dimensional observational data containing a large set of background variables that could serve as control variables. Examples include socio-economic characteristics of customers, geographic or time-related information, weather, economic circumstances, and many more. Causal ML is based on combining causal inference approaches with ML algorithms for data-driven selection of control variables when estimating causal effects and/or their heterogeneity across customer segments.

The rise of predictive ML has prompted e.g. Anderson (2008), Lycett (2013) and Erevelles, Fukawa, and Swayne (2016) to argue that theory-based causal inference has lost some of its relevance for business decisions in light of the large datasets and sophisticated predictive ML methods available to marketers today. However, these views were soon challenged in several studies that emphasize the importance of causal reasoning and risks of basing decisions based solely on correlations, see e.g. Cowls and Schroeder (2015) and Golder and Macy (2014). In more recent years, a growing number of contributions have further stressed the importance of integrating ML and causal inference, see e.g. Hair Jr and Sarstedt (2021). Among them, Hünermund, Kaminski, and Schmitt (2021), who investigate the use of causal methods in business analytics by combining qualitative interviews and quantitative surveys among data scientists and managers in a mixed-methods research design. They document an ongoing shift in corporate decision making away from an exclusive focus on predictive ML and towards the use of causal methods, based on both observational and experimental data.

Yet, to date, applications of causal ML to marketing research appear to be relatively scarce, with the exception of large tech companies operating in the field of social media or online commerce. To the best of our knowledge, there are virtually no studies that evaluate the causal effect of coupon campaigns on customer behavior using causal ML, as we do in this paper. Smith, Seiler, and Aggarwal (2021) use predictive ML for deriving optimal coupon targeting strategies and estimate the profits that would accrue under those strategies out of sample, i.e. in parts of the data not used for deriving the strategies. The profits are estimated based on the potential outcomes framework, which is also the basis of causal ML. However, the study by Smith, Seiler, and Aggarwal (2021) is conceptually different from ours in that it uses the potential outcomes framework to compare coupon targeting strategies inferred from different predictive ML algorithms, while we apply an algorithm based on the potential outcomes framework (namely the optimal policy learning approach of Athey and Wager (2021)) to derive a coupon targeting strategy.

One study in the field of marketing which does consider causal ML is Gordon, Moakler, and Zettelmeyer (2022). They assess the performance of so-called Double Machine Learning (DML), see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), and propensity score matching, see Rosenbaum and Rubin (1983), for estimating the causal effect of conversion ads on Facebook. Such ads aim to increase online activity like page visits, sales and views on an external website. For their analysis, the authors take advantage of the fact that Facebook offers businesses the opportunity to assess their ad campaigns by means of randomized experiments. Gordon, Moakler, and Zettelmeyer (2022) compare the effect estimates based on DML and propensity score matching with those from the experiments, finding that DML outperforms propensity score matching, but that both approaches overestimate the effect substantially. This highlights the importance of observing and appropriately controlling for all factors jointly affecting the intervention and customer behavior when causally assessing marketing interventions. Also, Huber, Meier, and Wallimann (2021) consider DML when analyzing observational data to investigate whether discounted tickets induce Swiss railway customers to reschedule their journeys, e.g. to shift demand away from peak hours.

Narang, Shankar, and Narayanan (2019) apply causal forests, the causal ML framework developed by Wager and Athey (2018) and Athey, Tibshirani, and Wager (2019) also used in this study (see Section 5.5), to assess the heterogeneity across shoppers in how mobile app failures affect the frequency, volume, and monetary value of their purchases. Guo, Sriram, and Manchanda (2021) assess the effect of a law requiring pharmaceutical firms to disclose their marketing payments to physicians on the firms' payments to physicians using a Difference-in-Differences approach and assess expected individual-level effect heterogeneity by means of causal forests. Zhang and Luo (2021) incorporate causal forests in their study on modelling restaurant survival as a function of photos posted on social networks. They find that the total volume of user-generated content and the extent to which user photos are rated as helpful have a significant positive effect on the likelihood of restaurant survival. Another study from the broader field of marketing that uses causal ML is Cagala, Glogowsky, Rincke, and Strittmatter (2021). The authors apply causal ML to determine the strategy for distributing gifts among potential donors to a fundraising campaign that maximizes expected net donations. They find that the identified optimal targeting rule outperforms different non-targeted gift distribution rules, even when the optimal targeting rule is estimated based only on publicly available geographic information or on data from a previous fundraising campaign conducted in a similar sample.

In the following, we will use coupon promotions as a running example to highlight the merits of causal ML in business analytics and marketing research. In the context of coupon campaigning strategies, marketers are arguably interested not only in predicting customer behavior, but also in measuring the causal effects of alternative campaigns on customer behavior. Such effects correspond to the difference in the customers' (average) behavior when being vs. not being addressed by a particular campaign. Intuitively, this requires comparing a customer's observed behavior under the actual assigned coupon with the potential (and not directly observed) behavior that would have occurred had coupon provision been different from that actually observed, an approach commonly referred to as counterfactual reasoning. Such a causal assessment is necessary for determining whether and to which extent a campaign is effective in altering customer behavior and for understanding how customer behavior would change if coupons were distributed differently.

In a predictive ML model, however, the predictive power of coupon provision on customer behavior generally does not correspond to such a causal effect, because it is affected by so-called regularization bias, i.e., a bias that arises in the context of ML algorithms shrinking the importance of certain predictors in order to reduce the variance of the prediction and thereby improve the overall predictive performance. Regularization bias may occur, for instance, when coupon provision is strongly correlated with other (good) predictors (such as previous purchases) and/or when its effect on consumer behavior is comparably small, so that coupon provision has little predictive value. A further issue is selection bias, meaning that coupons may pick up the effects of other variables whose importance has been shrunk by the ML algorithm. The implementation of coupon campaigns should be based on estimations of causal effects that avoid regularization and selection bias, as is the case with causal ML algorithms such as DML and causal forests.

The necessity of estimating the causal effect of coupon campaigns, rather than merely predicting customer behavior, can be illustrated by means of a simple example. Suppose a retailer estimates a model to predict sales based on observational data from a previous coupon campaign in which (in an attempt to re-activate dormant customers) coupons were distributed primarily among customers who had not been in the store for a while, rather than among frequent shoppers. The estimated predictive model might indicate a negative association between coupon provision and sales, since the coupon campaign is likely to re-activate only some inactive customers, so that the (formerly) inactive customers on average purchase less than the frequent shoppers. The true effect of receiving a coupon, however, might actually be positive. A positive effect implies that when comparing two groups of (formerly) inactive customers with comparable background characteristics (like willingness to buy), where the first receives coupons while the second does not, the average purchases of the first group are higher. The predictive model therefore confuses (or confounds) the causal effect of the coupon campaign with that of being a dormant vs. a frequent shopper, thus incorrectly pointing to a negative effect.

In a second scenario, the retailer decides to issue coupons in the store. This way, frequent shoppers are regularly provided with coupons, while dormant customers rarely if ever receive any. A predictive model now detects a positive relationship between the provision of coupons and sales, although the actual effect of providing coupons could be negative, namely if frequent customers use the coupons for products they would have bought anyway. If the campaigns were evaluated using predictive methods and the results were misinterpreted as causal, marketers would come to the conclusion that the first campaign was ineffective while the second was effective. Causal methods, on the other hand, enable marketers under certain conditions to control for such biases, in our example due to differences in purchasing behavior between frequent and dormant customers, and to consistently estimate the effect of coupon campaigns. Further, these methods can also be applied to assess effect heterogeneity and identify an optimal coupon distribution scheme (or policy) that targets those customers whose average purchases are sufficiently responsive to receiving a coupon.

In causal studies on discounts, the impact of providing coupons is typically assessed either based on random experiments or observational data from previous campaigns, controlling for observed characteristics or covariates that are likely to be associated with both coupon provision and consumer behavior. Conventional, i.e., non-ML-based, causal inference methods require the researcher or analyst to manually select covariates based on theoretical considerations, domain knowledge, intuition and/or previous empirical findings. Examples for such covariates in the context of campaign evaluations include past purchasing behavior, exposure to previous campaigns, and socio-economic characteristics such as age, gender, or income. Manual selection of covariates entails the risk of omitting important control variables and may even be practically infeasible in Big Data contexts with a very large set of potential covariates (e.g., collected from online platforms), including unstructured data containing, e.g., text or clickstreams. Furthermore, conventional causal inference methods require the researcher to specify how, i.e., through which functional form (like, e.g., a linear model), the selected covariates are associated with coupon provision and purchasing behavior. Causal ML methods, in contrast, permit taking advantage of the full amount of information in the data to detect relevant covariates (which have an important influence on coupon provision and consumer behavior) in a data-driven way and control for them, as well as to flexibly estimate the functional form of statistical associations. Still, the observational data have to meet certain conditions, as described in Section 5.4.2.

The argument for counterfactual reasoning made further above also applies to efforts of optimizing the distribution of coupons across segments of customers, i.e., optimal policy learning, as discussed, e.g., in Manski (2004), Hirano and Porter (2009), Stoye (2009), and Kitagawa and Tetenov (2018). Basing optimization on predictive ML models, as advocated in several studies on predicting coupon redemption (e.g. Koehn, Lessmann, and Schaal (2020), Ren, Cao, Xu, et al. (2021), Greenstein-Messica, Rokach, and Shabtai (2017)), ignores the fact that predictive models do generally not provide information on causal effects and their heterogeneity across different customer segments. Causal ML-based policy learning as suggested by Athey and Wager (2021), on the other hand, is a causal ML approach to inferring allocation schemes which ensure that those customers for whom sufficiently large effects can be expected are targeted by the campaign. In our empirical application, we demonstrate how causal ML methods can help evaluate coupon campaigns and support marketing-related decision making. We analyze customer data from a retail store and first evaluate the average effect of providing customers with a coupon (of a certain type) on the monetary value of their purchases. In a second step, we demonstrate how optimal policy learning can be used for detecting customer segments that should or should not be targeted by coupon campaigns to maximize the effectiveness of these campaigns.

5.3 Data

In our empirical application, we analyze sales data on coupon campaigns of a retailer, which are available on the data science platform Kaggle (2019) under the denomination 'Predicting Coupon Redemption'. The dataset contains information on socio-economic characteristics of retail store customers, the coupons they have received during the campaigns as well as on their coupon redemption and purchasing behavior. The retail store ran several campaigns issuing coupons with discounts for certain products, with some coupons being applicable to individual products only and others to a range of products. In each of the 18 partially time-overlapping campaigns falling into the time span covered by the dataset, the store distributed 1 to 208 different coupon types each applicable to up to 12,000 products, most of which belong to the same product category. The coupons were distributed in such a way that each customer received 0 to 37 different coupons per campaign with the composition of this set of coupons varying between the recipients. Apart from the information on provided coupons, the dataset contains details on all purchases made by each registered customer between January 2012 and July 2013, including the date of the transaction, the redeemed coupons, the product type of each purchased product and the price paid.

For our analysis, we group the coupons into five broad categories mirroring the products they can be used for. More concisely, we distinguish between coupons applicable for ready-to-eat food items, meat and seafood, other food, drugstore items and other non-food products[2].

[2] The coupons of each category are applicable to the following product categories defined by the retailer: (a) ready-to-eat food coupons: 'Bakery', 'Restaurant', 'Prepared Food', 'Dairy, Juices & Snacks', (b) meat and seafood coupons: 'Meat', 'Packaged Meat', 'Seafood', (c) coupons applicable to other food: 'Grocery', 'Salads', 'Vegetables (cut)', 'Natural Products', (d) drugstore coupons: 'Pharmaceutical', 'Skin & Hair Care', and (e) coupons applicable to other non-food products: 'Flowers & Plants', 'Garden', 'Travel', 'Miscellaneous'

One could arguably also be interested in more fine-grained coupon categories or in a paricular coupon or discount type rather than in our broader coupon categories, which would, however, require a larger dataset to obtain satisfactory statistical power. Due to the temporal overlap of campaign periods, we need to redefine them such that each of the resulting artificially generated campaign periods coincides with the validity period of a given set of coupons. That is, all coupons which are valid in some artificial campaign period are valid during the entire period. By doing so, we can fully attribute changes in purchasing behavior from one artificial campaign period to another to the coupons valid in the respective periods. From now on, the 33 newly defined artificial campaign periods will simply be referred to as campaign periods. To account for differences in the duration of campaign periods, we consider the average per-day expenditures per customer and campaign period as our outcome of interest. For estimating the causal effect of coupon provision on the buying behavior, we pool the customer-specific purchases across campaign periods, yielding 33 observations per customer.

Table 5.3.1 provides some descriptive statistics for our data, namely on observed customer characteristics, the share of coupons redeemed and daily in-store spending (descriptive statistics on the composition of daily expenditures by product type are provided in Table 5.A.3 in the appendix). The table reports the mean of these variables in the total sample of 50,624 observations, as well as among observations that received a coupon and among those that did not. Further, it contains the mean difference in these variables between coupon receivers and non-receivers, as well as the p-value of a two-sample t-test. In some 30% of the observations, customers received at least one coupon. Furthermore, customers who received a coupon had on average higher expenditures in the retail store than customers who did not, suggesting that the retailer did not target its previous campaigns to re-activate dormant customers. We also see that the retailer does not have information on the socio-economic characteristics of all customers in the registry, but only for about half of them, as the corresponding variable values are unknown for many observations (see the coding 'unknown'). Such a high rate of non-response in measuring variables can entail selection bias when estimating the effects of interest. For this reason, we will conduct several robustness checks in the empirical analysis to follow further below (see Section 5.6.5). The descriptive statistics also reveal that some socio-economic characteristics as well as their observability are correlated with the reception of coupons. For example, customers aged 70 years or older are less likely to be targeted by a coupon campaign. The main difference

variable N	Overall 50,624	Coupon Receivers 15,327	Non-Receivers 35,297	Diff	p-val
daily expenditures	202	245	184	61	0
age:					
18-25	0.028	0.031	0.027	0	0.02
26-35	0.082	0.102	0.074	0.03	0
36-45	0.118	0.141	0.108	0.03	0
46-55	0.171	0.191	0.163	0.03	0
56-70	0.037	0.045	0.034	0.01	0
70+	0.043	0.039	0.045	-0.01	0
unknown	0.52	0.451	0.55	-0.1	0
family size:					
1	0.157	0.171	0.15	0.02	0
2	0.192	0.213	0.182	0.03	0
3	0.066	0.079	0.06	0.02	0
4	0.03	0.04	0.026	0.01	0
5+	0.036	0.047	0.031	0.02	0
unknown	0.52	0.451	0.55	-0.1	0
marital status:					
married	0.2	0.234	0.186	0.05	0
unmarried	0.072	0.084	0.067	0.02	0
unknown	0.728	0.682	0.747	-0.07	0
dwelling type:					
rented	0.026	0.033	0.023	0.01	0
owned	0.454	0.516	0.428	0.09	0
unknown	0.52	0.451	0.55	-0.1	0
income group:					
1	0.037	0.042	0.035	0.01	0
2	0.043	0.051	0.04	0.01	0
3	0.044	0.049	0.042	0.01	0
4	0.104	0.113	0.1	0.01	0
5	0.118	0.137	0.11	0.03	0
6	0.056	0.061	0.053	0.01	0
7	0.02	0.023	0.019	0	0.01
8	0.023	0.03	0.021	0.01	0
9	0.018	0.024	0.016	0.01	0
10	0.006	0.006	0.006	0	0.89
11	0.003	0.003	0.003	0	0.43
12	0.006	0.01	0.005	0	0
unknown	0.52	0.451	0.55	-0.1	0
coupons redeemed		0.030			

Table 5.3.1: Mean of the variables in the total sample ('Overall'), among coupon receivers and non-receivers as well as the mean difference across treatment states and the p-value of a two-sample t-test.

in the likelihood of receiving a coupon seems to be between customers whose socioeconomic characteristics are not available and those whose characteristics are known, with the former less likely to receive a coupon.

As is noted in several studies (e.g. Danaher, Smith, Ranasinghe, and Danaher (2015), Spiekermann, Rothensee, and Klafft (2011)) coupon redemption rates are typically low, not exceeding 1 to 3% on average. This is also the case in our data, as only in 3% of the observations of coupon recipients did they actually redeem a coupon. However, as mentioned further above, coupons may not only influence customer behavior when redeemed, but may also serve as an advertising tool which attracts customers to the store even without them redeeming the coupon.

5.4 Identification

5.4.1 Causal Effect

We are interested in estimating the causal effect of a specific intervention, commonly referred to as 'treatment' in causal analysis and henceforth denoted by D, on an outcome of interest, denoted by Y.[3] In our context, D reflects the reception or non-reception of coupons and Y the purchasing behavior, measured as the average per-day expenditures during the coupon validity period. In the simplest treatment definition, D is binary and takes the value 1 when the respective customer is provided with a coupon and 0 if this is not the case. Mathematically speaking, the value d which treatment D can take satisfies $d \in \{0, 1\}$. The set of observations with $d = 1$ is commonly referred to as the treatment group, those for which $d = 0$ are called control group. Our subsequent discussion of causal effects and the statistical assumptions required for their measurement will focus on this binary treatment case for the sake of simplicity. However, our empirical analysis will also separately consider the effects of receiving coupons for five product categories, by running separate estimations for the comparison of each category to not receiving any coupons. This implies that the assumptions introduced in Section 5.4.2 need to hold for each of these categories. For discussions of multi-valued treatments, see e.g. Imbens (2000) and Lechner (2001).

For defining the causal effect of coupon provision, we rely on the potential outcome framework

[3] Throughout this paper, capital letters denote random variables and small letters specific values of random variables.

pioneered by Neyman (1923) and Rubin (1974). Let $Y(d)$ denote the potential (rather than observed) outcome under a specific treatment value $d \in \{0,1\}$. That is, $Y(1)$ corresponds to a customer's potential purchasing behavior if she received a coupon, while $Y(0)$ is the behavior without a coupon. The causal effect of the coupon thus corresponds to the difference in the purchasing behavior with and without coupon, $Y(1)-Y(0)$, but can unfortunately not be directly assessed for any customer. This is due to the impossibility of observing customers at the same point in time under two mutually exclusive coupon assignments (1 vs. 0), which is known as the 'fundamental problem of causal inference', see Holland (1986). This follows from the fact that the outcome Y which is observed in the data corresponds to the potential purchasing behavior under the coupon assignment actually received, namely $Y = Y(1)$ for those receiving a coupon ($d = 1$), and $Y(0) = Y$ for those who do not ($d = 0$). For coupon recipients, however, $Y(0)$ cannot be observed in the data, while for customers without a coupon $Y(1)$ remains unknown.

Even though causal effects are fundamentally unidentifiable at the individual level, we may, under the assumptions outlined further below, evaluate them at more aggregate levels, i.e., based on groups of treated and nontreated individuals. One causal parameter which is typically of crucial interest is the average causal effect, also known as average treatment effect (ATE), i.e., the average effect of coupon assignment D on purchasing behavior Y among the total of customers. Formally, the ATE, which we henceforth denote by Δ, corresponds to the difference in the average potential outcomes $Y(1)$ and $Y(0)$:

$$\Delta = E[Y(1) - Y(0)], \tag{5.1}$$

where '$E[...]'$ stands for 'expectation', which is simply the average in the population.

5.4.2 Identifying Assumptions

In order to identify the ATE defined in the previous section, we need to impose several identifying assumptions, which are outlined in this section. We note that in the subsequent discussion, '$\perp$' stands for statistical independence. Further, X denotes the set of covariates that should not be affected by treatment D and therefore be observed before or at, but not after, treatment.

Assumption 1 (conditional independence of the treatment):

$Y(d) \perp D | X$ for all $d \in \{0,1\}$. Assumption 1 states that the treatment is conditionally independent of the outcome when controlling for the covariates, and is known as 'selection on observables', 'unconfoundedness' or 'ignorable treatment assignment', see e.g. Rosenbaum and Rubin (1983). The assumption implies that there are no unobservables jointly affecting the treatment assignment and the outcomes conditional on the covariates. This condition is satisfied if the coupons are quasi-randomly distributed among observations with the same values in X. The retailer may therefore base the distributing of coupons on customer or market characteristics observed in the data, however, not on unobserved characteristics that affect purchasing behavior even after controlling for the observed ones.

We control for the variables in Table 5.3.1, period fixed effects, the customers' average daily pre-campaign spending by product category, as well as for the coupons she received and redeemed in the period prior to the campaign. When evaluating the effect of specific coupon categories, we also include dummies that indicate whether a customer received coupons from another category at the moment of treatment assignment. This is because the availability of other coupons influences purchase behavior and is likely to be correlated with the probability of receiving coupons of the category under study. The reason for including period fixed effects is that there is no information available on holidays or weekdays on which the store is closed or has shortened opening hours, that is, circumstances that may affect purchasing behavior. Also, the retailer is likely to distribute coupons differently across campaign periods. Including pre-campaign expenditures allows controlling for general differences in purchasing behavior between customers that might be correlated with the likelihood of receiving coupons, since the retailer presumably bases decisions about whom to allocate which coupon(s) on past purchasing behavior.

The covariates considered in our estimation are similar to those included in studies on the effect of coupon campaigns that rely on traditional causal inference approaches, see, e.g., Xing, Zou, Yin, Wang, and Li (2020) and Hsieh, Shimizutani, and Hori (2010), both of which control for some demographic characteristics as well as for a proxy for the customers' economic situation and their purchasing behavior before the coupon campaign under study. Unlike the methods used in these studies, however, the causal ML approach we apply in this study allows covariates to enter into the estimation in a flexible, possibly non-linear way, and does not require pre-selection of variables based on theoretical considerations.

Studies on predicting coupon redemption by means of ML mostly rely exclusively on observable customer behavior and coupon characteristics as predictors of coupon redemption while not including socio-demographic characteristics of customers, see, e.g., Greenstein-Messica, Rokach, and Shabtai (2017) use and He and Jiang (2017).

In their study on the performance of causal ML in evaluating Facebook ads, Gordon, Moakler, and Zettelmeyer (2022) include users' gender, age and household size but - unlike our data - their data lack information on users' economic situation, such as their income, employment status, or wealth. They also use several Facebook-specific covariates measuring users' activity on Facebook (likes, posts, type of device used and interests explicitely expressed on Facebook). Furthermore, they take into account users' response to earlier ads from other companies, which is comparable to the covariates on pre-campaign purchasing behavior, coupon reception and coupon redemption considered in our analysis. Despite the large differences in the amount of information available in the Facebook study and our analysis, we cannot conclude that the set of covariates in our estimation is insufficient. For one, the algorithms used by Facebook to determine the target audience for ad placement are far more complex and information-hungry than a retailer's coupon strategy; and Facebook users' decision about whether or not to respond to a Facebook ad is likely to be complex and dependent on several of the characteristics considered in the algorithm (which is why they are considered in Facebook's ad placement algorithm). In order to successfully apply causal ML methods, the authors of the Facebook study had to take into account all the information that is incorporated in Facebook's ad placement algorithms, just as we need to consider the information based on which the retailer distributed its coupons, namely the information available in the customer database.

Assumption 2 (common support):

$0 < Pr(D = 1|X) < 1$.

Assumption 2 states that the conditional probability of being treated given X, in the following referred to as the treatment propensity score, is larger than zero and smaller than one. This so-called common support condition implies that for all values the covariates might take, customers have a non-zero chance of being treated and a non-zero chance of not being treated. While this assumption is imposed w.r.t. to the total of a (large) population, meaning that both treated and non-treated customers exist conditional on X, we can and should also verify it in the data. In our sample, common support appears to be satisfied, as there exist no combinations of covariate values for which either only customers with coupons (of a certain category) or no coupons exist. Appendix 5.B shows the distribution of the estimated propensity scores for receiving coupons (of a particular type) among recipients and non-recipients of that particular coupon(s). The distributions overlap (although the overlap is partially thin), i.e., for each observation in one group, observations can be found in the other group that are comparable with respect to the propensity score.

Another condition that needs to be satisfied is the so-called Stable Unit Treatment Value Assumption (SUTVA), see, e.g., Rubin (1980). In our context, SUTVA rules out that the coupons provided to one individual affect the potential outcome of another individual. The assumption that there are no inter-personal spillover effects of coupon campaigns may be problematic in our setting. Customers receiving coupons may induce their peers to make purchases by, for instance, telling peers about the products they bought when redeeming the coupon or by visiting the store together with peers. On the other hand, customers with coupons may also redeem their coupons to buy the coupon-discounted products not only for themselves but also for their peers, thereby reducing the purchases made by their peers. Such scenarios appear particularly likely when there are several members of the same household in the customer base. There is ongoing research on how to deal with such SUTVA violations under certain assumptions like the observability of groups affected by spillovers, see e.g. Sobel (2006), Hong and Raudenbush (2006), Hudgens and Halloran (2008) Tchetgen and VanderWeele (2012), Aronow and Samii (2017), Huber and Steinmayr (2021) and Qu, Xiong, Liu, and Imbens (2021). However, in our dataset, the relationships between customers are not observable, meaning the data does not allow accounting for possible spillovers of providing coupons to one customer on the outcomes of other customers. If such spillovers existed in our case, they could entail an under- or overestimation of the effect of coupons on purchasing behavior, depending on whether the spillovers occur primarily through treated customers inducing non-treated peers to make purchases or through treated customers redeeming coupons to purchase products for their peers, with the former entailing an overestimation of the outcome under non-treatment and the latter leading to an underestimation.

SUTVA also requires that for every individual in the population, there is a single potential outcome value associated with each treatment state, meaning that there are no different versions of the coupons leading to different potential outcomes. In many empirical applications, it appears likely that at least some aspects of SUTVA are violated, and for this reason, there exist several relaxations of this assumption. In our case, the requirement that there be no different treatment versions is particularly problematic given that we group different coupons into broader categories. The treatment of being provided with coupon(s) from one category comprises the receipt of different coupons, each applicable to a distinct set of products from the respective product category. If a customer is not equally interested in all products belonging to that product category, the customer may only redeem a coupon and/or change her purchasing behavior if the coupon is applicable to certain products. For this reason, we are in a setting where there are different treatment versions, each possibly associated with a different potential outcome.

VanderWeele and Hernan (2013) relax the original SUTVA by allowing for the existence of different unobservable versions of the treatment as long as there are no different versions of non-treatment and the treatment versions are assigned randomly conditional on the covariates X. This permits assessing the average effects of certain bundles of coupons (rather than specific coupons as under the original SUTVA) vs. not receiving any coupons. Indeed, the assumption that there is only one version of non-treatment is satisfied in our analysis of the effect of receiving some vs. no coupons, under the assumption that the marketer has not run any undocumented discount campaigns during the study period. Furthermore, when assessing the effects of coupons applicable to specific product categories, we control for all other coupons that each customer received at treatment assignment, which in turn creates non-treatment states that are necessarily equal after controlling for other coupons. Table 5.3.1 and the tables in Appendix 5.A show that the coupons were distributed under consideration of the covariates in the customer registry. We must now assume that the propensity of receiving a coupon (version) differs only depending on observed characteristics, but not on characteristics that are not available to us. This issue can be easily circumvent in practice as long as the information on customers available to marketing campaign planners is also available to those evaluating the campaign.

We note that our assumptions do not rule out inter-temporal spillover effects on customers' purchasing behavior, since in our main analysis we only examine the (short-term) effect of coupon provision on purchasing behavior during the validity period of the coupon rather than longer-term coupon-induced behavioral shifts. Individuals may, therefore, advance their purchases towards campaign periods in which they receive coupons applicable to the products they are interested in. By including pre-campaign coupon reception and redemption as control variables, we aim at accounting for the fact that previous coupons may influence customer behavior in the outcome period.

In order to get an impression of the extent to which coupons induce inter-temporal spillover effects and, on the other hand, longer-term increases in customer retention, we also assess the effect of coupon provision in campaign period t on daily expenditures in subsequent periods, namely in $t+1$ and $t+2$. It should, however, be noted that the estimated effect is the total effect of coupon reception on purchasing behavior in these subsequent periods, that is, it does not only capture the longer-term coupon-induced change in purchases at the store (net of spillovers from advancing purchases in periods in which the customer has applicable coupons). Rather, it also captures how coupon provision in t affects purchasing behavior in $t+1$ and $t+2$ through changing the likelihood of coupon reception in these later periods (e.g., because the customer redeems coupons in t or the coupons incentivize her to increase her purchases in t). Disentangling the direct effect of coupon provision on purchasing behavior in subsequent periods from the indirect effect mediated via increasing the likelihood of coupon provision in these later periods would require estimating dynamic treatment effects of treatment sequences, such as the sequence of coupon reception in t and non-reception in $t+1$ (see Bodory, Huber, and Lafférs (2020) for an approach to estimating dynamic treatment effects by means of DML). Further, some coupons valid in t may still be valid in $t+1$ and even $t+2$. The estimated effect of coupon provision in t on purchasing behavior in later periods therefore also partially captures the treatment effect of coupons during their validity period. A look at the data shows that the likelihood of having a valid coupon in $t+1$ or $t+2$ is highly correlated with that of having a coupon in t (conditional on X), be it due to the effect of coupons on re-provision or because the validity period of coupons exceeds that of the artificially created campaign periods. Part of the estimated longer-term effect is therefore likely attributable to the indirect effect of coupon provision in t on daily expenditures in $t+1$ and $t+2$, via increasing the probability that the customer has valid coupons in these subsequent periods.

5.5 Causal Machine Learning

In the following, let $i \in \{1, ..., n\}$ be an index for the $n = 1,582$ customers in the dataset and $t \in \{1, ..., T\}$ with $T = 32$ an index for the campaign period. Then, $\{Y_{i,t}, D_{i,t}, X_{i,t}\}$ denote the outcome, the treatment and the covariates, respectively, for individual i in campaign period t. Treatment $D_{i,t}$ is a binary indicator measuring exposure to a coupon campaign (of a specific type) and $Y_{i,t}$ denotes the outcome, defined as average per-day expenditures of customer i in period t. The covariates $X_{i,t}$, all measured prior to or at the time of treatment assignment, include socio-economic variables (see Table 5.3.1), the average daily spending by product type in the period prior to the campaign $t-1$, and variables that measure both whether the customer received coupons in $t-1$ and whether he/she redeemed any. For estimating the effect of a particular coupon type, $X_{i,t}$ also contains variables on what other coupon types were provided to the customer in t; in addition, it includes information not only about whether the customer received coupons in $t-1$, but also about what type of coupons.

Under the identifying assumptions outlined in Section 5.4.2, the ATE Δ defined in equation (5.1) corresponds to θ:

$$\theta = E[\mu(D=1,X) - \mu(D=0,X)] \tag{5.2}$$

where $\mu(D=d,X)$ denotes the conditional mean outcome given treatment state $D=d$ and covariates X. As long as the function μ is of known functional form and X is low-dimensional, we can estimate $\hat{\mu}(D,X)$ by regressing Y on D and X and then determine the ATE according to equation (5.2). However, the amount of customer data available to marketers is often extensive, and the functional form of relationships between observable customer characteristics and purchasing behavior is often unknown and complex. It may, therefore, in many cases be preferable to use an approach that integrates ML algorithms into the estimation of the causal effect to take advantage of the functional flexibility and the ability to deal with high-dimensional data inherent in ML algorithms. Put simply, ML algorithms are used to estimate models for predicting Y as a function of D and X ($\mu(D,X)$) and for predicting the probability of being treated conditional on X, which is commonly referred to as the propensity score $p(X) = Pr(D=1|X)$. These predictions are then integrated into the estimation of the treatment effects.

We assess the causal effect of receiving coupons (of a certain category) on average per-day spending using causal forests, a causal ML method developed by Wager and Athey (2018) that draws on the ML technique of random forests. While the causal forest framework primarily aims at estimating treatment effect heterogeneity, i.e., how the effect of coupons is distributed across different clients and time periods (see Section 5.5.1), the estimated causal forests can also be used to estimate the ATE of coupon provision (see Sections 5.5.2). Both the causal forest algorithm for assessing treatment effect heterogeneity and the estimation procedure used for determining the ATE rely on combining effect estimation on so-called Neyman (1959)-orthogonal scores with sample splitting. The purpose of orthogonalization is to ensure the robustness of the estimation of causal effects to regularization bias which accrues when using ML to estimate $\mu(D, X)$ and $p(X)$, in the following referred to as plug-in parameters $\eta = (\mu_D(X), p(X))$. Sample splitting, on the other hand, aims to avoid overfitting in the estimation of treatment effects. In Section 5.5.3, we outline how the estimated causal forest can be utilized for determining the treatment effect in different customer segments as defined by selected covariates. Section 5.5.4, finally, shows how to use the estimated causal forest to determine which customers should optimally be targeted with the different coupon campaigns.

5.5.1 Treatment Effect Heterogeneity

The causal forest approach by Wager and Athey (2018) is a modified version of the random forest aimed at determining splitting rules that maximize the heterogeneity of treatment effects in the resulting subsamples. The causal forest provides individualized treatment effect estimates for every observation in the sample as a function of its covariates X, which are commonly referred to as Conditional Average Treatment Effects (CATEs), and thereby gives an impression of the heterogeneity in the effect of coupon provision across customers and campaign periods.

Causal forests are built from so-called causal trees just as random forests are built from regression/classification trees. In order to generate a causal forest, the algorithm repeatedly (2,000 times in our case) draws random samples with 50% of the observations in the dataset. In each random sample, it estimates a causal tree as follows: first, a randomly selected subset of $min(\sqrt{k} + 20, k)$ covariates is chosen, which in our case amounts for 30 of our $k = 93$ covariates. The algorithm then utilizes these covariates for splitting the sample into two subsamples such that the CATEs in the two resulting subsamples are as heterogeneous as possible. More precisely, the algorithm determines both the covariate and the value at which the sample should be split (e.g. age < 25 vs. age ≥ 25) to maximize effect heterogeneity. Intuitively, the algorithms considers all possible splits on values of the 30 covariates to find the optimal split in terms of effect heterogeneity. The subsamples obtained from this splitting rule are commonly referred to as nodes. These nodes are further split into a larger number of nodes following the same procedure until some stopping rule is reached, e.g., that no further splits are made if they would entail nodes with less than 5 treated or 5 control observations. The causal forest is finally obtained by averaging over the splitting structure of all 2000 causal trees.

The CATE in the subsamples resulting from each potential split is estimated by means of an approach proposed by Robinson (1988) that allows estimating the CATE with $\sqrt{n}$ consistency. The approach builds on first predicting the plug-in parameters $\eta = (\mu_D(X), p(X))$, where the plug-in parameters can be estimated using any predictive ML algorithm as long as the plug-in estimates converge with a convergence rate faster than $n^{-1/4}$, and then using the predicted plug-in parameters for estimating the CATE. In our case, the plug-in parameters are predicted by means of regression forests with out-of-bag prediction.[4] In a second step, the algorithm calculates the residuals $Y_{i,t} - \hat{\mu}_{D_{i,t}}(X_{i,t})$ and $D_{i,t} - \hat{p}(X_{i,t})$ for all observations i, t in the random sample used for learning the causal tree. In order to determine the split that maximizes effect heterogeneity in the resulting subsamples, the algorithm regresses $Y_{i,t} - \hat{\mu}_{D_{i,t}}(X_{i,t})$ on $D_{i,t} - \hat{p}(X_{i,t})$ in each subsample. That is, for every potential node, the algorithm estimates the following function, where $\hat{\theta}(X)$ denotes the estimated CATE:[5]

$$Y - \hat{\mu}_D(X) = (D - \hat{p}(X))\hat{\theta}(X). \tag{5.3}$$

By comparing the estimated CATEs in all potential nodes, the algorithm determines the splitting rule for which the estimated CATEs differ most between the two resulting subsamples. The approach of first predicting the plug-in parameters and then incorporating them into effect estimation ensures that causal effect estimation is more robust to slight approximation errors in the plug-in parameter estimates, which may arise from regularization biases, i.e., from neglecting less important covariates in the splitting procedure.

Furthermore, the causal forest algorithm addresses another source of bias, namely overfitting, i.e., fitting the effect heterogeneity model too strongly to the particularities of the data, such that the procedure picks up not only the actual differences of causal effects across covariates, but

[4] First, the data set is split into two subsamples, each of which is used to learn regression forests for predicting $\mu_D(X)$ and $p(X)$, respectively. Then, in both subsamples, the plug-in parameters are estimated using the forests learnt in the respective other subsample. The final estimate of the plug-in parameters is obtained by averaging over the estimates from both samples.

[5] For computational efficiency, the splitting rules are not determined by estimating the CATEs in all possible subsamples. Rather, the algorithm approximates the between-node effect heterogeneity generated through every potential split by means of a gradient for each observation. Then, the algorithm involves several conditions for formulating splitting rules that aim at avoiding imbalance in the size of the nodes. Explaining these rules in detail would go beyond the scope of this discussion. The manual to the `grf` package, however, provides all the details (see Athey, Friedberg, Hadad, Hirshberg, Miner, Sverdrup, Tibshirani, Wager, and Wright (2022)). In our application, we keep all options of the `causal_forest` function at their default values.

also random noise. In order to prevent such overfitting bias, the random sample used for learning a causal tree is itself randomly split into two subsamples, one for building the tree by following the procedure mentioned above, while the other one is used for estimating the treatment effect in every node of the learnt causal tree. That is, by following the splitting rules learnt in the first subsample, the algorithm populates the nodes of the estimated tree with the observations from the second subsample and calculates the CATE in each node based on the observations that fall into the respective node. Trees that are estimated based on this sample splitting procedure are commonly referred to as 'honest' trees (because they avoid overfitting).

Through averaging over 2000 causal trees, the causal forest provides the final estimates of the CATEs $\hat{\theta}(X)$, i.e., estimations of individualized treatment effects for every point in X. To account for the issue that the behavior of one and the same customer is in general not independent across different campaign periods we cluster standard errors at the customer level. The estimation is performed in the statistical software R (R Core Team (2022)) by means of the `causal_forest` function provided in the `grf` package by Athey, Friedberg, Hadad, Hirshberg, Miner, Sverdrup, Tibshirani, Wager, and Wright (2022).

5.5.2 Average Treatment Effect

The estimated causal forest can further be used to identify the ATE of coupon provision and thus to assess the overall effectiveness of the coupons (and that of selected coupon types). Athey and Wager (2019) propose to estimate the ATE by means of a modified version of the Augmented Inverse Probability Weighting (AIPW) estimator, a doubly robust estimator proposed by Robins, Rotnitzky, and Zhao (1995), that is based on weighting the observations by the inverse of their estimated propensity score. This weighting of observations makes the treatment and the control group comparable in terms of their propensity scores and hence the distribution of relevant covariates X (for more information on the AIPW estimator see e.g. Glynn and Quinn (2010)). Double robustness is achieved by estimating the ATE via an orthogonalized function, i.e., the predicted plug-in parameters are included in the estimation such that small estimation errors in either predictor result in an overall negligible error and hence do not introduce bias in the estimation of the ATE. The formula used for estimating the ATE is as follows:

$$
\begin{aligned}
\hat{\Theta} &= \frac{1}{NT}\sum_{i\in N, t\in T}\Gamma_{i,t} \qquad (5.4)\\
\text{with } \Gamma_{i,t} &= \hat{\theta}(X_{i,t}) + \frac{D_{i,t}-\hat{p}(X_{i,t})}{\hat{p}(X_{i,t})(1-\hat{p}(X_{i,t}))}\Big(Y_{i,t}-\hat{\mu}(X_{i,t})-\Big(D_{i,t}-\hat{p}(X_{i,t})\Big)\hat{\theta}(X_{i,t})\Big)
\end{aligned}
$$

where the plug-in parameters $\hat{\theta}(X), \hat{p}(X)$ and $\hat{\mu}(X)$ for the doubly robust score $\Gamma_{i,t}$ are obtained from the estimated causal forest. As mentioned above, $\hat{p}(X)$ and $\hat{\mu}(X)$ are predicted by means of regression forests with out-of-bag prediction while $\hat{\theta}(X)$ is determined using honest trees, i.e., the plug-in estimators for observation (i,t) are computed based on models learnt in samples that do not contain observation (i,t). This makes the AIPW-based ATE estimator robust to regularization bias. Thus, similarly to how the CATE is estimated for building causal trees, the modified AIPW estimator by Athey and Wager (2019) combines orthogonalization and out-of-sample prediction in order to address the two sources of bias, overfitting and regularization.

The causal-forest based approach for estimating the ATE described above ensures that the ATE can be estimated with $\sqrt{n}$-consistency, i.e., the estimated ATE converges to the true ATE with a convergence rate of $1/\sqrt{n}$, provided that the ML steps satisfy specific regularity conditions (like $n^{-1/4}$-consistency). A look at equation (5.4) reveals that values of $\hat{p}(X_{i,t})$ that are either close to zero or close to one can yield large weights for the respective observations, resulting in unstable performance of the estimator. This issue is commonly addressed by trimming the dataset, i.e., discarding observations with an estimated propensity score that is below or above certain values. A commonly used trimming rule is to remove observations with estimated propensity scores larger than 0.99 or smaller than 0.01, an approach we also employ in this study.

In our application, we estimate the ATE using the `average_treatment_effect` function provided in the `grf` package for R by Athey, Friedberg, Hadad, Hirshberg, Miner, Sverdrup, Tibshirani, Wager, and Wright (2022), with standard errors clustered at the customer level.

5.5.3 Group Average Treatment Effects

In order to assess the impact of coupon provision in different customer groups, we also estimate selected Group Average Treatment Effects (GATEs), that is, the average treatment effects in different subgroups as defined by age, income, family size and pre-campaign expenditures, respectively. The variables used to distinguish these subgroups are the age group and family size variables as defined in the original dataset, a variable for average daily expenditures that divides the sample into four subgroups of similar size, and a variable measuring income in broader categories, each of which combines two of the more fine-grained income groups in the original variable. We estimate a linear model of the doubly robust scores $\hat{\Gamma}_{i,t}$ (see equation (5.4)) as a function of one of the variables that indicate which subgroup each client belongs to, see Semenova and Chernozhukov (2021) for more details. This approach also allows us to assess effect heterogeneity in customer segments defined by more than one variable, by regressing the Neyman (1959)-orthogonal scores $\hat{\Gamma}_{i,t}$ on several identifiers (or dummy variables) for belonging to a specific subgroup defined in terms of covariate values (e.g. an indicator for being younger female or elderly male customer). We estimate the GATEs by means of the `best_linear_projection` function provided in the `grf`-package.

5.5.4 Optimal Policy Learning

The optimal policy learning approach by Athey and Wager (2021) goes one step further, in the sense that it does not only estimate the effect of coupon provision in predefined customer groups. Rather, it exploits the heterogeneity in coupon effects to determine the coupon distribution rule that maximizes the overall effect of the coupon campaign. Based on observed covariates, the coupon distribution rule distinguishes customer segments that are likely to increase their purchasing behavior upon receiving a coupon from those customer groups not anticipated to respond positively to the campaign. More formally, the algorithm considers specific decision (or policy) rules for whether a coupon should be offered to a customer as a function of the covariate values in X, e.g., the customer's age. Let us denote by $\pi(X)$ such a decision rule, which could, for instance, impose that only elderly, but not younger, customers obtain a coupon.

Mathematically speaking, the rule maps a customer's observed characteristics to the binary treatment decision of whether or not to target the customer through the coupon campaign:

$\pi : X \rightarrow 0, 1$. Optimal policy learning consists of learning the optimal rule among an assumably limited set of implementable candidate policies, where we use Π to denote this set. For instance, another possible rule of how to distribute coupons (in addition to the age-based rule) could be to offer them only to customers with a high volume of previous purchases. Then, both the age- and purchase-dependent rule would enter the set of feasible coupon policies provided in Π.

For learning the optimal coupon policy, the algorithm of Athey and Wager (2021) use the doubly robust scores $\hat{\Gamma}_{i,t}$ (see equation (5.4)). These individual- and time-specific treatment effect estimates are plugged into the following objective function, which aims at maximizing the effectiveness of the coupon campaign by selecting the policy rule with the highest average effect among all policies π that are available in the set Π:

$$\pi^* = argmax\left\{ \frac{1}{NT} \sum_{i \in 1,\ldots,N} \sum_{t \in 1,\ldots,T} (2\pi(X_{i,t}) - 1)\hat{\Gamma}_{i,t} : \pi \in \Pi \right\} \tag{5.5}$$

The optimal policy learning approach does not require defining a priori the policies to be considered, but only the number of customer segments between which coupon allocation can differ and the set of covariates that can be considered for determining these customer segments. Thus, the approach identifies the optimal coupon policy in a data-driven way. To determine the optimal coupon distribution strategy, i.e., the one that maximizes the objective function in (5.5), the algorithm applies a tree-based approach that considers all possible covariate-defined sample splits for generating the customer segmentation (according to the pre-defined number of segments) and all possible coupon assignment strategies within these segments. The resulting coupon distribution rule can be represented as a decision (or policy) tree, i.e., a tree-shaped graph indicating at which values of which covariate the sample is split and which of the resulting customer segments shall receive coupons.

We estimate decision trees of depth 3, implying that we distinguish 8 customer segments for defining the optimal distribution of coupons by means of the `policytree` package for R by Sverdrup, Kanodia, Zhou, Athey, and Wager (2020). For determining the customer segments, we use all the customer characteristics available in the dataset, i.e., age and income group, family size, marital status, and dwelling type. We redefine these variables by setting all missing values to -1, which allows us to omit the variables indicating which observations are missing. Then, we also include the customers' pre-campaign purchasing behavior. Since the algorithm performs a sample split at every possible value of each covariate, i.e., at each observed value, continuous variables can cause performance issues by driving up the number of sample splits. We, therefore, round the pre-campaign average daily expenditures to round values, namely to the nearest 100 for values between 0 and 1,000 and to the nearest 200 for values between 1,000 and 2,000. Further, we group all 157 observations with average daily expenditures of 2,000 or more into one category and include dummies that indicate whether a customer purchased items from the different product categories in the period prior to the campaign. This way, we still capture pre-campaign differences in purchasing behavior well, while substantially reducing the number of sample splits that need to be performed.

5.6 Empirical Results

5.6.1 Treatment Effect Heterogeneity

Figure 5.6.1 shows the distribution of the individualized treatment effects (CATEs) as estimated by means of the causal forest algorithm outlined in Section 5.5.1. We can see that the treatment effect of being provided with any coupon is positive for the vast majority of observations and, except for some outliers, ranges between -100 and 200 monetary units. Similarly, provision of drugstore coupons and coupons applicable to other food have a positive effect for the majority of observations. The distribution of coupons applicable to ready-to-eat food as well as meat and seafood, however, seem to be rather centered around zero, with the estimated effect being positive for about half of the observations and negative for the other half. For coupons applicable to other non-food prodcuts, we can even observe a negative effect on daily expenditures for the majority of observations. The plots suggest greater heterogeneity in the treatment effects of the individual coupon categories than when all coupons are analyzed together. It appears that the effects of the different coupon categories cancel each other out to some extent when combined in one analysis, implying that the different coupon categories should best be analyzed separately.

The differences in CATEs as revealed by the causal forest approach suggest not just assessing the ATE, as is done in Section 5.6.2. Rather, it also invites to analyze how the effect of coupons (of certain categories) differs between customer groups as defined by covariates X (Section 5.6.3) and to learn an optimal coupon distribution scheme that maximizes the expected ATE of coupon provision (Section 5.6.4).

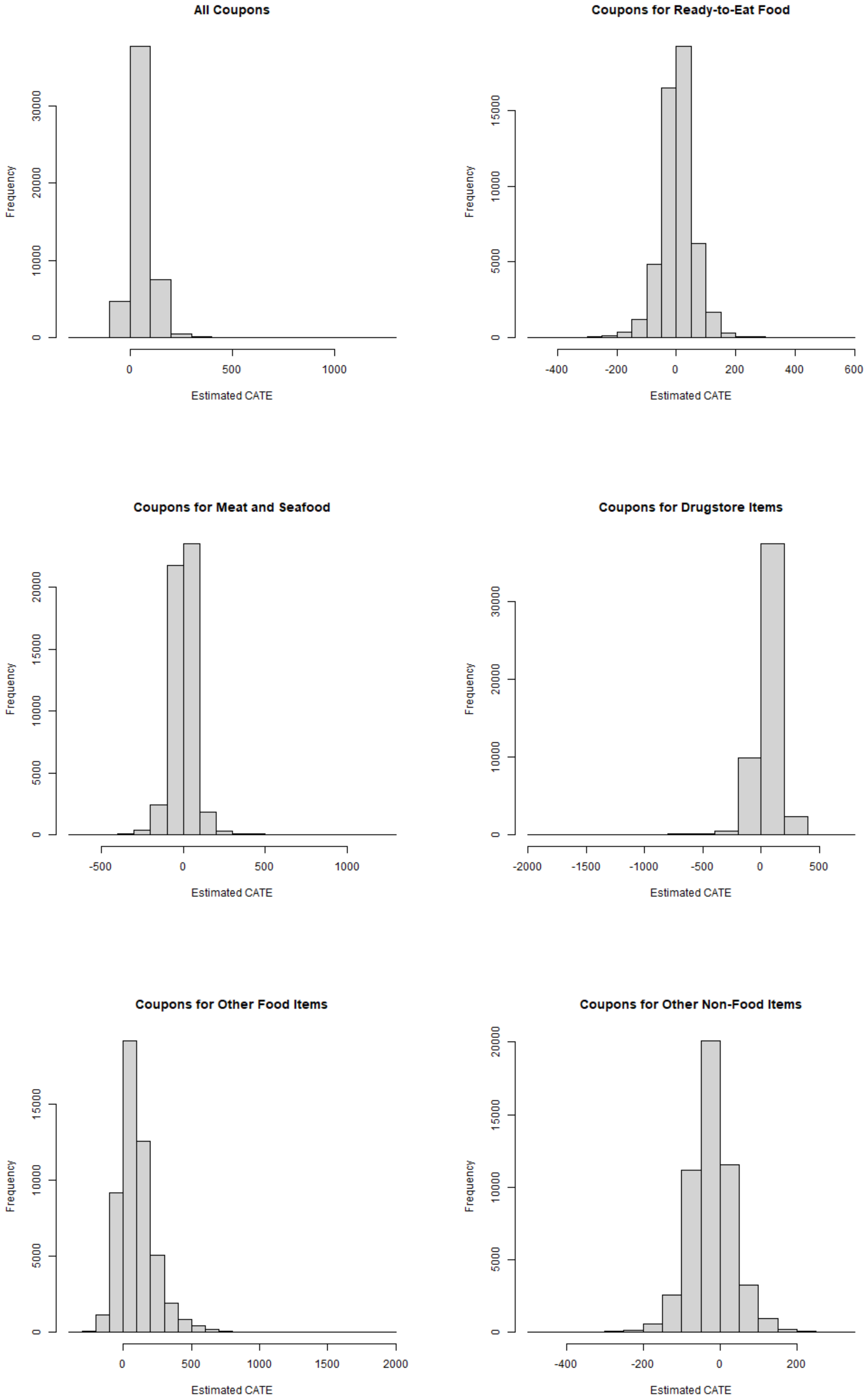


Figure 5.6.1: Distribution of CATE by coupon type.

5.6.2 The Causal Effect of Receiving Coupons

Table 5.6.1 shows the estimated ATE of receiving any coupon on daily expenditures in the campaign period, as well as that of receiving coupons from each of the five coupon categories, based on the AIPW approach outlined in Section 5.5.2. The results show that receiving any coupon has a positive and statistically significant effect on daily expenditures during the campaign period. Providing a customer with a coupon increases her expected daily expenditures by some 63 monetary units. The effect estimates for the different coupon categories provide a more nuanced picture. Provision of coupons for drugstore items and other food has a statistically significant positive effect on daily spending during the campaign period. Receiving coupons that belong to these categories increases expected average daily expenditures during the validity period by some 60 and 75 monetary units, respectively. Handing out coupons applicable to other non-food products, on the other hand, is estimated to decrease a customer's expected average daily expenditures by some 27 monetary units, with this results also being statistically significant. The estimated ATE of providing coupons from the other two categories has no statistically significant effect on the customers' expected daily spending during the campaign period, with the estimated effects being slightly negative. A possible explanation for the insignificant or significantly negative effect of these latter three coupon types is that the receipt of such coupons may not incentivize people to buy, but that such coupons are mainly used for products that the coupon recipient would have purchased anyway.

	Coef.	Standard Error	Sign. Level
ATE: receiving any coupon	63.26	4.553	***
ATE: receiving coupon for ready-to-eat food	-2.90	8.118	
ATE: receiving coupon for meat/seafood	-1.42	6.045	
ATE: receiving coupon for other food	74.74	13.559	***
ATE: receiving coupon for drugstore items	60.07	6.521	***
ATE: receiving coupon for other non-food items	-26.77	6.949	***

Table 5.6.1: ATE of receiving any coupon as well as the ATEs of receiving coupons applicable to specific product categories, each with standard error and significance level. Significance levels: . $p<0.1$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

As discussed in Section 5.4.2, coupon provision may, on the one hand, have longer-term positive effects on purchasing behavior by increasing customer loyalty, and on the other hand, bring about inter-temporal spillovers by inducing customers to advance their purchases to periods when they have coupons applicable to them. We therefore also take a look at the overall effect of coupon reception in t on daily expenditures in the following campaign period $(t+1)$ and the period thereafter $(t+2)$ (see Table 5.6.2). The results suggest that the effect of coupon provision on daily expendidtures is sustainable, i.e., coupon provision in t not only has a short-term effect on purchases in t, but also has a statistically significant, albeit smaller, effect on purchases in subsequent periods. This may be due to a coupon-induced increase in customer retention (but also to indirect effects, see the discussion in Section 5.4.2).

The longer-term effect of drugstore and other food coupons is also positive and statistically significant, with drugstore coupons showing an even larger effect on purchasing behavior in both post-treatment periods than in the short term. Coupons applicable to other non-food products, that in the short run have a statistically significant negative effect, show a statistically signifiificant positive effect on daily spending in the subsequent periods. One possible explanation for this finding is that, while in the short run these coupons were only redeemed for the purchase of products that would have also been purchased without the coupons, in the longer term they may have increased customer loyalty.

The estimated effect of meat and seafood coupons on expenditures in $t+1$ and $t+2$ is not statistically significant, while that of ready-to-eat food coupons is even significantly negative for the outcome in $t+1$, which may indicate spillover effects that are not offset by positive expenditure-increasing effects. For ready-to-eat food and meat/seafood coupons, we can therefore conclude that they do not seem to be an effective marketing tool for increasing customer spending, neither in the short nor in the longer run.

	Effect in $t+1$			Effect in $t+2$		
	coef.	s.e.	sign.	coef.	s.e.	sign.
ATE: receiving any coupon	39.82	3.279	***	34.56	3.87	***
ATE: ready-to-eat food coupons	-28.70	6.113	***	4.18	8.908	
ATE: meat/seafood coupons	6.71	6.171		1.13	6.244	
ATE: other food coupons	52.79	11.506	***	2.46	7.345	
ATE: drugstore coupons	88.39	5.711	***	82.78	5.681	***
ATE: other non-food coupons	28.03	6.204	***	11.64	6.017	.

Table 5.6.2: ATE on daily expenditures in period after coupon campaign $(t+1)$ and the period thereafter $(t+2)$, each with standard error and significance level. Significance levels: . p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

In the following, we will again focus on the short-term effect of coupon provision in t on daily expenditures in t. The next section examines effect heterogeneity with regard to selected customer characteristics. This is because the provision of coupons could significantly increase spending of certain customer groups, despite not having a statistically significant effect on the overall customer base. Similarly, providing coupons applicable to drugstore or other food could have a significant impact on purchasing behavior only among certain subgroups of customers.

5.6.3 Group Average Treatment Effects

In this section, we assess how the provision of coupons affects different customer groups, based on the approach discussed in Section 5.5.3. We illustrate how the effect of providing coupons differs depending on the customers' age, income, family size and pre-campaign expenditures. Further, we also examine the GATEs of those coupon categories with a highly statistically significant ATE, i.e., drugstore coupons and coupons applicable to other food.

Figure 5.6.2 shows the GATEs of receiving any coupon by age, income, family size and pre-campaign expenditures, respectively. The graphs show that providing coupons has a positive effect on purchasing behavior in every customer group that is statistically significant in most subgroups. The effect of providing coupons tends to be particularly large among customers from smaller households and among those who made either no or large purchases in the period prior to the campaign.

The GATE charts in Figure 5.6.3 show that in almost all subgroups considered, the provision of coupons for other food has a positive, and in many cases statistically significant, effect on daily spending. The most pronounced differences in GATEs can be found among customer subgroups defined by average daily spending prior to the campaign period. The effect of these food coupons tends to be high and statistically significant for previously inactive customers, while it is much smaller, though still statistically significant, for customers with high pre-period spending. This may suggest that coupons for other food have the potential to reactivate dormant customers. This hypothesis is also supported by the fact that the provision of coupons applicable to other food has a relatively large statistically significant[6] effect among customers for whom information on socio-economic characteristics is not available. Customers for whom no information is available may be more likely to have low loyalty/retention to the store and to be

[6] the small confidence interval around this GATE estimator can be explained with the large number of observations for which no information on socio-economic characteristics is available

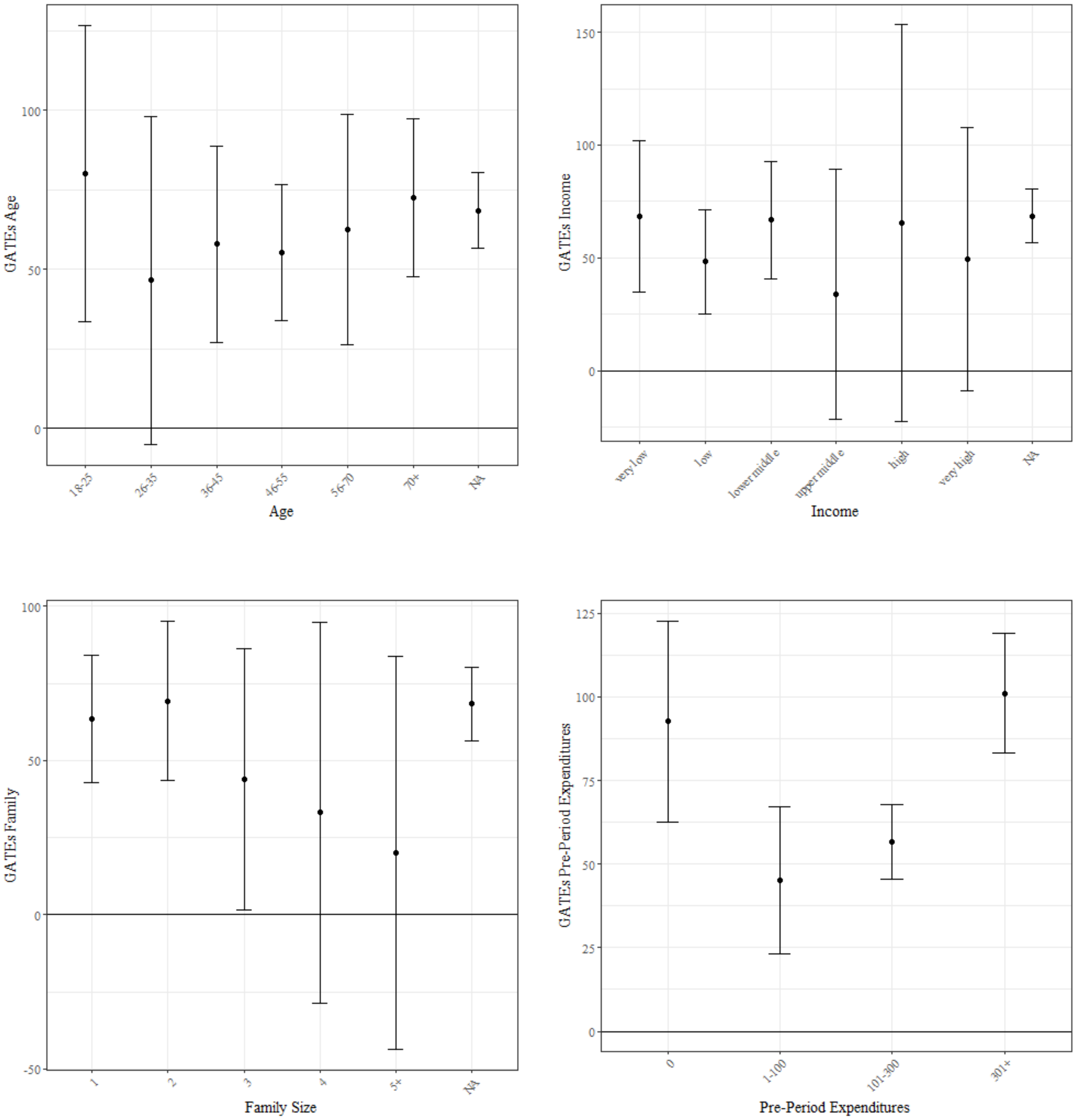


Figure 5.6.2: GATEs of receiving any coupon with 95% confidence interval.

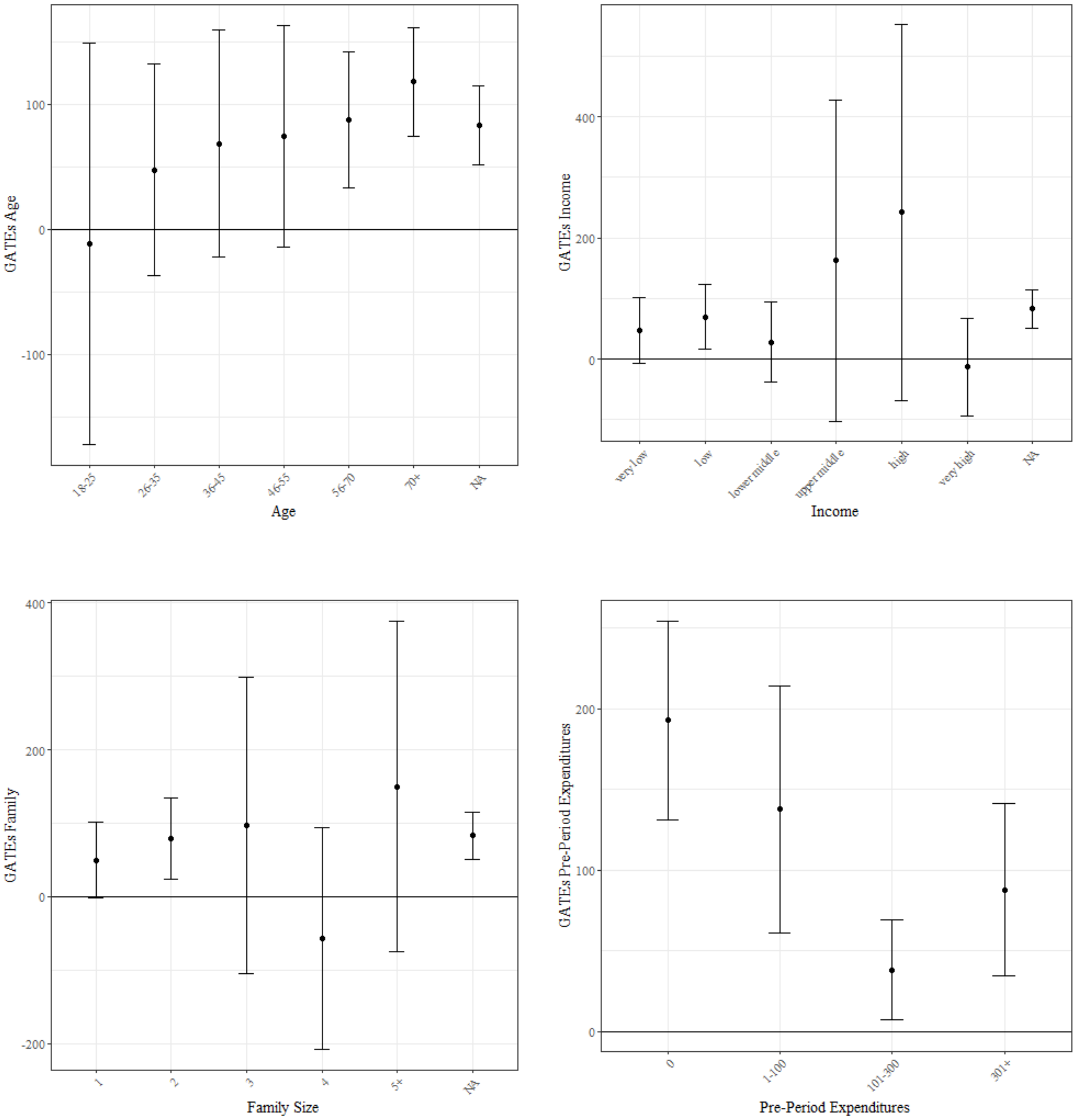


Figure 5.6.3: GATEs of coupons applicable to other food items with 95% confidence interval.

rather inactive customers who can be reactivated by providing them with other food coupons. From these results, we can deduce the hypothesis that coupons applicable to other food are efficient for inactive and non-frequent customers, while they have less impact on the purchasing behavior of frequent shoppers.

Figure 5.6.4 shows that providing drugstore coupons has a positive effect on daily spending for almost all subgroups considered, and that the effect is statistically significant in most cases. Again, the largest difference can be found in the GATE estimates by pre-campaign spending. The effect of drugstore coupons on average per-day expenditures is larger the higher the customer's pre-campaign spending, which is the reverse pattern of what we find for coupons applicable to other food. This suggests that other food coupons are more efficient at reactivating dormant customers and drugstore coupons at retaining frequent shoppers. The GATE plots for the other three coupon categories can be found in Appendix 5.C.

While ML-based estimation of ATEs and GATEs is an excellent tool for evaluating the effect of coupon campaigns, it is not necessarily most appropriate for deriving strategies for later coupon campaigns. For this purpose, the optimal policy learning framework by Athey and Wager (2021) is arguably superior as it determines which customer groups to provide with coupons in order to maximize the ATE.

5.6.4 The Optimal Distribution of Coupons

Figure 5.6.5 shows the optimal distribution rules (or policies) for each coupon category as identified based on the optimal policy learning approach outlined in Section 5.5.4. The optimal distribution rule for ready-to-eat food coupons (decision tree (a)) suggests providing ready-to-eat food coupons to customers with no drugstore purchases in the pre-campaign period if their marital status is unknown[7] and their age is unknown or they are not older than 26, or if their marital status is known and they live in a household of no more than three members. The retailer should further provide ready-to-eat food coupons to customers who purchased drugstore products in the pre-period if their income is in one of the lowest four income groups or unknown, or if their age is unknown and their average daily purchases in the pre-period were less than 50

[7] Please note that for family size, marital status, age group and income group the value -1 denotes that information about these variables is unavailable, see also the description of the methodology in Section 5.5.4

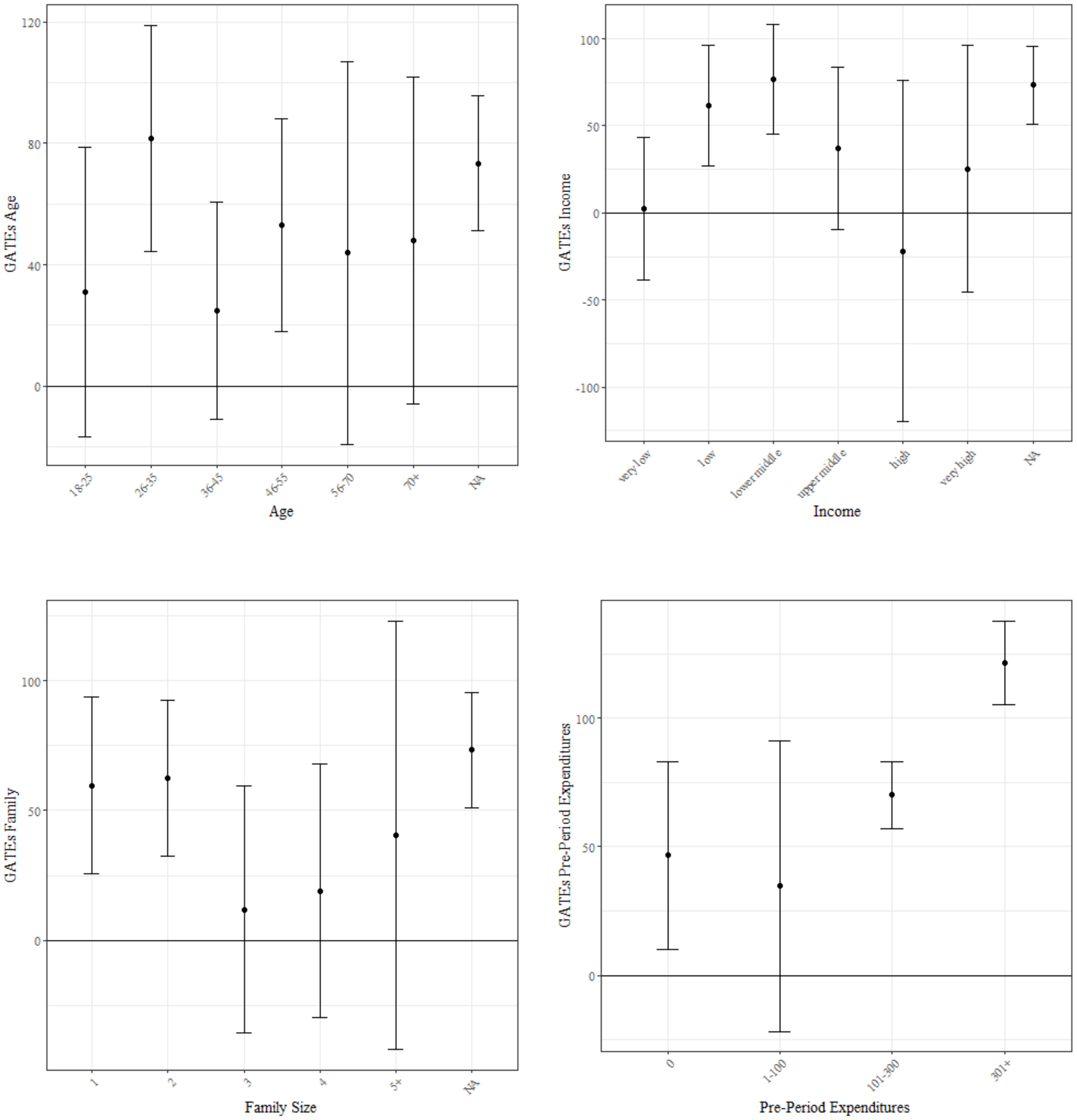


Figure 5.6.4: GATEs of drugstore coupons with 95% confidence interval.

monetary units[8].

The optimal distribution rule for drugstore coupons (decision tree (d)) proposes providing drugstore coupons to those customers with unkown, low, or middle incomes if their daily pre-campaign expenditures did not exceed 600 monetary units. Customers belonging to the high-income group should receive drugstore coupons if their average in-store spending did not exceed 300 monetary units per day in the period before the campaign. In addition, customers whose pre-campaign expenditures exceeded 600 monetary units per day should only be provided with drugstore coupons if they did not purchase any other non-food products in the pre-campaign period and do not belong to the high-income group, or if they purchased non-food products and are either 18-26 years old or of unknown age.

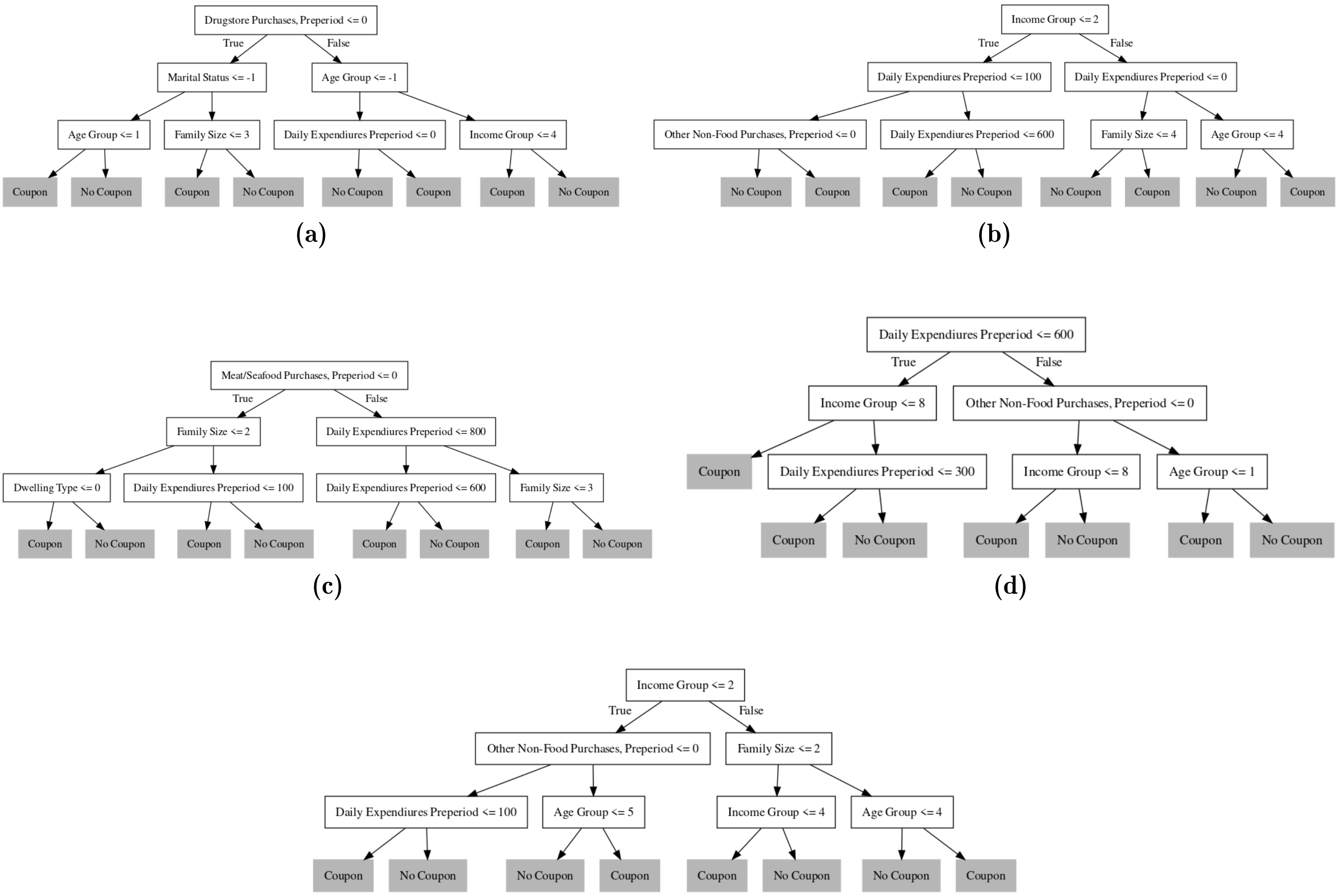


Figure 5.6.5: Depth-3 trees for coupons applicable to (a) ready-to-eat food, (b) meat and seafood, (c) other food, (d) drugstore products as well as (e) other non-food products.

[8] Please note that daily expenditures are rounded for creating policy trees. All customers with daily expenditures below 50 monetary units fulfil the condition ‘Daily Expenditure Preperiod <= 0’

The distribution rules paint a similar picture as the GATE estimates in Section 5.6.1 about which customer groups are likely to be positively impacted by the provision of certain coupon types. In contrast to the assessment of effect heterogeneity across pre-specified broad categories in Section 5.6.1, the optimal policy learning algorithm finds the covariate values at which the sample should optimally be split in order to maximize the ATE and defines groups of coupon recipients and non-recipients based on multiple covariates.

The other decision trees can be interpreted accordingly. A look at the covariates used for sample splitting in those other decision trees shows that each observed customer characteristic is used for distribution rules of at least one coupon type.

5.6.5 Robustness Checks

As described in Section 5.3, our dataset contains a large number of observations with missing socio-economic information. To investigate the robustness of our results with respect to these missing values, we performed the entire analysis on a reduced dataset containing only observations of customers whose socio-economic background is known, i.e., on a dataset with 13,792 observations of the purchasing behavior of $n = 431$ individuals. The estimated ATEs can be found in Table 5.6.3. They are close to the ATE estimates from the full dataset, although the standard errors are of course considerably larger - due to the much smaller number of observations.

	Coef.	Standard Error	Sign. Level
ATE: receiving any coupon	41.63	8.024	***
ATE: receiving coupon for ready-to-eat food	-6.50	10.618	
ATE: receiving coupon for meat and seafood	-26.94	9.374	**
ATE: receiving coupon for other food	96.72	22.897	***
ATE: receiving coupon for drugstore items	41.12	9.493	***
ATE: receiving coupon for other non-food items	-22.72	9.568	*

Table 5.6.3: ATE of receiving any coupon as well as the ATEs of receiving coupons applicable to specific product categories in the reduced dataset (without observations with missing socio-economic information), each with standard error and significance level. Significance levels: . p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

The GATE and policy tree plots are provided in Appendix 5.D. The GATE plots show similar patterns in how different customer groups are affected by each coupon type, although of course they are not exactly identical with the GATEs estimated in the full sample. The policy tree plots show that the splitting rules are based on a similar set of variables with similar cutting

	Effect in $t+1$			Effect in $t+2$		
	coef.	s.e.	sign.	coef.	s.e.	sign.
ATE: receiving any coupon	24.29	7.609	**	20.20	8.545	*
ATE: ready-to-eat food coupons	-19.31	9.878	.	-26.22	9.586	**
ATE: meat/seafood coupons	-7.90	10.305		-8.56	11.434	
ATE: other food coupons	54.77	14.86	***	-4.28	15.827	
ATE: drugstore coupons	68.69	8.398	***	47.31	11.723	***
ATE: other non-food coupons	10.33	7.235		6.96	10.275	

Table 5.6.4: ATE on daily expenditures in period after each coupon campaign ($t+1$) and the period thereafter ($t+2$), estimated in the reduced dataset (without observations with missing socio-economic information), each with standard error and significance level. Significance levels: . p<0.1, * p<0.05, ** p<0.01, *** p<0.001.

points as those estimated in the full dataset. These results suggest that the large number of observations with missing socio-economic information does not introduce systematic bias into the estimation of the treatment effects and the optimal coupon distribution scheme.

5.7 Conclusion

This paper presented different causal ML methods with multiple application possibilities in marketing research and business development. The application of these methods to evaluate a retailer's coupon campaign and optimize the distribution of coupons illustrated their potential in identifying the ATE of coupon provision, effect heterogeneity, as well as optimal coupon distribution rules.

For instance, we found that only coupons belonging to two out of five categories, namely those applicable to drugstore and other food, have a positive and statistically significant overall effect on purchases, while receiving coupons for other non-food products actually significantly reduces customers' daily spending. Additionally, we were able to pinpoint different customer subgroups whose purchasing behavior can be influenced particularly strongly through provision of certain types of coupons and who should therefore be optimally addressed with the corresponding coupon campaigns. This information would enable the retailer to optimally target her coupon campaigns, i.e., such that the overall effect is maximized.

The proposed causal ML methods can further be applied to evaluate and optimize a variety of other marketing and business strategies, requiring only observational data from the context of previous campaigns or business decisions, and utilizing all available (Big) data, whether structured or unstructured. Other potential applications for the proposed causal ML methods are the evaluation and optimization of targetable (online) marketing campaigns, loyalty programs and campaigns for dealing with customer attrition, but also the assessment of different employee benefit plans, designs of job postings, or in-house training programs. The potential use cases of causal ML in business and marketing are manifold.

Appendix

5.A Descriptive Statistics

	ready-to-eat food coupons		meat/seafood coupons		other food coupons	
variable	received	not received	received	not received	received	not received
daily expenditures	237	257	234	264.37	241	335
age: 18-25	0.03	0.032	0.029	0.034	0.031	0.033
26-35	0.101	0.103	0.092	0.12	0.099	0.159
36-45	0.14	0.143	0.138	0.147	0.142	0.132
46-55	0.191	0.192	0.191	0.192	0.191	0.208
56-70	0.051	0.036	0.044	0.047	0.046	0.029
70+	0.042	0.034	0.042	0.032	0.04	0.021
unknown	0.444	0.461	0.463	0.427	0.452	0.418
family size: 1	0.174	0.167	0.173	0.169	0.172	0.164
2	0.222	0.199	0.214	0.212	0.215	0.175
3	0.079	0.078	0.077	0.083	0.078	0.092
4	0.036	0.045	0.034	0.05	0.038	0.073
5+	0.044	0.05	0.039	0.06	0.045	0.078
unknown	0.444	0.461	0.463	0.427	0.452	0.418
marital status: married	0.239	0.227	0.227	0.246	0.235	0.207
unmarried	0.082	0.088	0.081	0.09	0.084	0.097
unknown	0.679	0.686	0.691	0.664	0.681	0.696
dwelling type: rented	0.031	0.036	0.029	0.041	0.032	0.06
owned	0.525	0.503	0.507	0.532	0.516	0.521
unknown	0.444	0.461	0.463	0.427	0.452	0.418
income group: 1	0.045	0.038	0.041	0.044	0.043	0.022
2	0.049	0.054	0.049	0.053	0.05	0.057
3	0.052	0.044	0.048	0.05	0.049	0.041
4	0.115	0.11	0.109	0.121	0.111	0.164
5	0.14	0.133	0.138	0.135	0.137	0.14
6	0.059	0.066	0.061	0.063	0.062	0.052
7	0.023	0.023	0.025	0.02	0.023	0.016
8	0.028	0.033	0.025	0.038	0.03	0.038
9	0.024	0.024	0.021	0.029	0.023	0.043
10	0.008	0.003	0.006	0.008	0.006	0.008
11	0.003	0.003	0.003	0.002	0.003	0
12	0.011	0.008	0.01	0.01	0.01	0
unknown	0.444	0.461	0.463	0.427	0.452	0.418
coupons redeemed	0.037	0.018	0.036	0.018	0.031	0.006

Table 5.A.1: Mean of the variables among the treated who received a coupon of a certain category and of those who did not

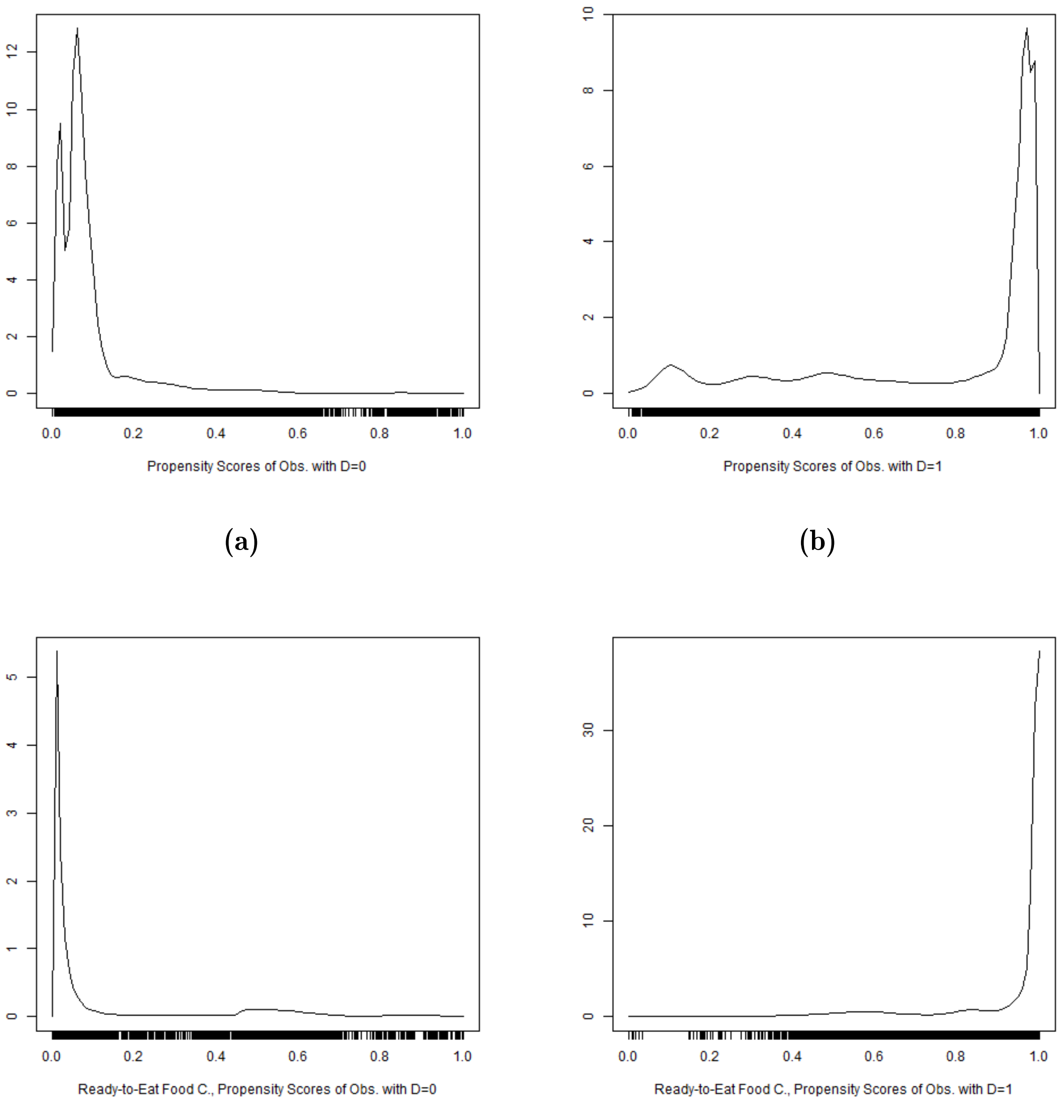
	drugstore coupons		other non-food coupons	
variable	received	not received	received	not received
daily expenditures	243	272	246	243.54
age: 18-25	0.031	0.033	0.036	0.028
26-35	0.104	0.07	0.107	0.099
36-45	0.143	0.117	0.149	0.137
46-55	0.191	0.198	0.2	0.187
56-70	0.046	0.035	0.047	0.044
70+	0.039	0.033	0.04	0.038
unknown	0.447	0.514	0.422	0.466
family size: 1	0.172	0.162	0.177	0.168
2	0.215	0.176	0.224	0.207
3	0.079	0.076	0.079	0.079
4	0.04	0.031	0.043	0.038
5+	0.047	0.04	0.055	0.042
unknown	0.447	0.514	0.422	0.466
marital status: married	0.236	0.205	0.257	0.221
unmarried	0.085	0.074	0.075	0.09
unknown	0.68	0.721	0.668	0.689
dwelling type: rented	0.033	0.047	0.032	0.034
owned	0.521	0.438	0.547	0.499
unknown	0.447	0.514	0.422	0.466
income group: 1	0.041	0.05	0.055	0.035
2	0.051	0.047	0.05	0.051
3	0.05	0.037	0.046	0.051
4	0.114	0.1	0.115	0.112
5	0.139	0.114	0.149	0.131
6	0.062	0.057	0.065	0.059
7	0.023	0.018	0.022	0.023
8	0.031	0.017	0.031	0.029
9	0.024	0.027	0.022	0.025
10	0.006	0.006	0.008	0.005
11	0.003	0.005	0.003	0.003
12	0.01	0.008	0.01	0.009
unknown	0.447	0.514	0.422	0.466
coupons redeemed	0.031	0.005	0.043	0.022

Table 5.A.2: Mean of the variables among the treated who received a coupon of a certain category and of those who did not

variable N	Overall 50,624	Coupon Receivers 15,327	Non-Receivers 35,297	Diff	p-val
By Brand Type:					
Established Brands	221	255	206	48	0
Local Brands	59.17	74.01	52.73	21.28	0
By Product Type:					
Alcohol	0.657	0.975	0.519	0.46	0
Bakery	78.79	85.1	76.05	9.05	0
Dairy, Juices & Snacks	4.72	6.08	4.13	1.96	0
Flowers & Plants	0.699	0.846	0.635	0.21	0.02
Fuel	94.35	107.2	88.77	18.42	0
Garden	1.93	2.61	1.63	0.98	0
Grocery	114	136	105	31	0
Meat	82.04	88.45	79.25	9.2	0
Miscellaneous	4.7	5.47	4.37	1.1	0
Natural Products	6	7.11	5.51	1.6	0
Packaged Meat	87.86	95.4	84.59	10.81	0
Pharmaceutical	31.8	38.88	28.72	10.16	0
Prepared Food	2.48	3.05	2.24	0.81	0
Restaurant	76.08	81.94	73.54	8.41	0
Salads	1.71	2.16	1.52	0.64	0
Seafood	1.86	2.08	1.76	0.32	0.01
Skin & Hair Care	76.94	82.95	74.32	8.63	0
Travel	1.7	2.14	1.5	0.64	0
Vegetables (cut)	0.017	0.026	0.014	0.01	0.03

Table 5.A.3: Mean of daily expenditures by brand and product type in the total sample ('Overall'), among coupon receivers and non-receivers as well as the mean difference across treatment states and the p-value of a two-sample t-test.

5.B Propensity Score Distribution



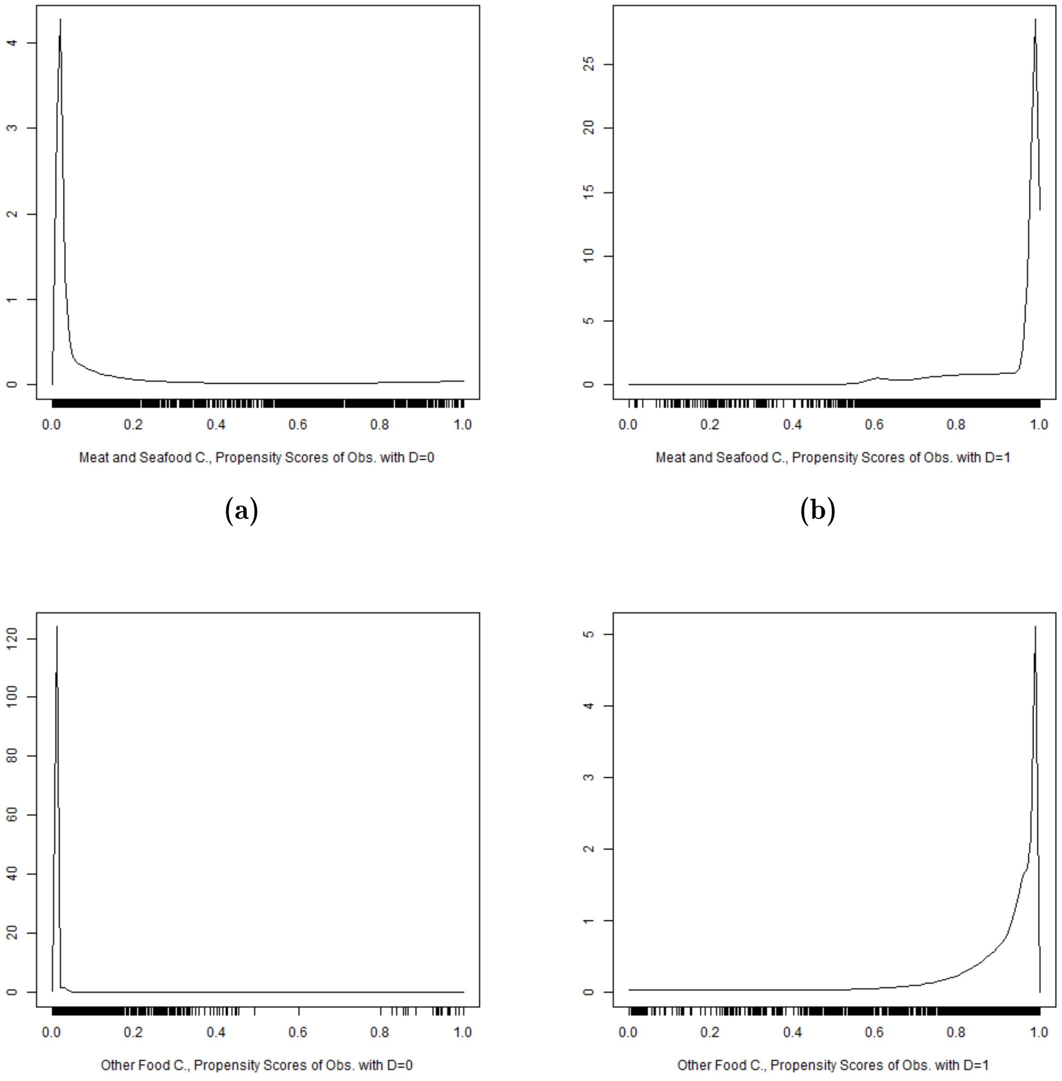
(a)

(b)

(c)

(d)

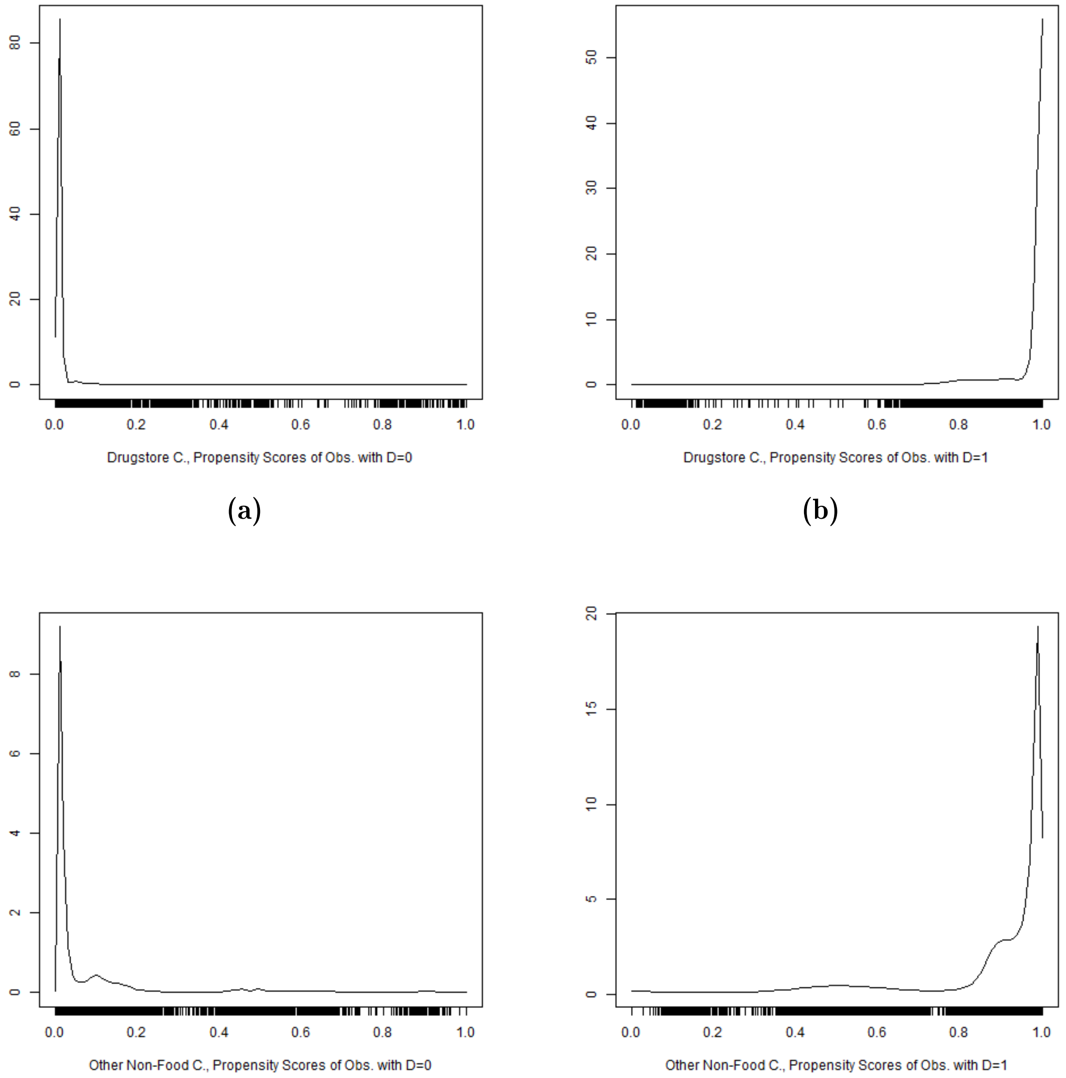
Figure 5.B.1: Distribution of propensity scores of receiving any coupon among observations that received (a) any coupon and (b) no coupon, as well as that of the propensity scores of receiving ready-to-eat food coupons among observations that (c) did and (d) did not received ready-to-eat food coupons. The plots are produced with the `logspline` command in R with the lower and upper bounds of the support of the propensity scores are set to 0 and 1.



(a) **(b)**

(c) **(d)**

Figure 5.B.2: Distribution of propensity scores of receiving meat/seafood coupons among observations that (c) did and (d) did not received meat/seafood coupons, as well as that of the propensity scores of receiving other food coupons among observations that (c) did and (d) did not received other food coupons. The plots are produced with the `logspline` command in R with the lower and upper bounds of the support of the propensity scores are set to 0 and 1.



(a) **(b)**

(c) **(d)**

Figure 5.B.3: Distribution of propensity scores of receiving drugstore coupons among observations that (c) did and (d) did not received drugstore coupons, as well as that of the propensity scores of receiving other non-food coupons among observations that (c) did and (d) did not received other non-food coupons. The plots are produced with the `logspline` command in R with the lower and upper bounds of the support of the propensity scores are set to 0 and 1.

5.C GATE Estimates for Coupons Applicable to Plants, Drugstore Items and Other Products

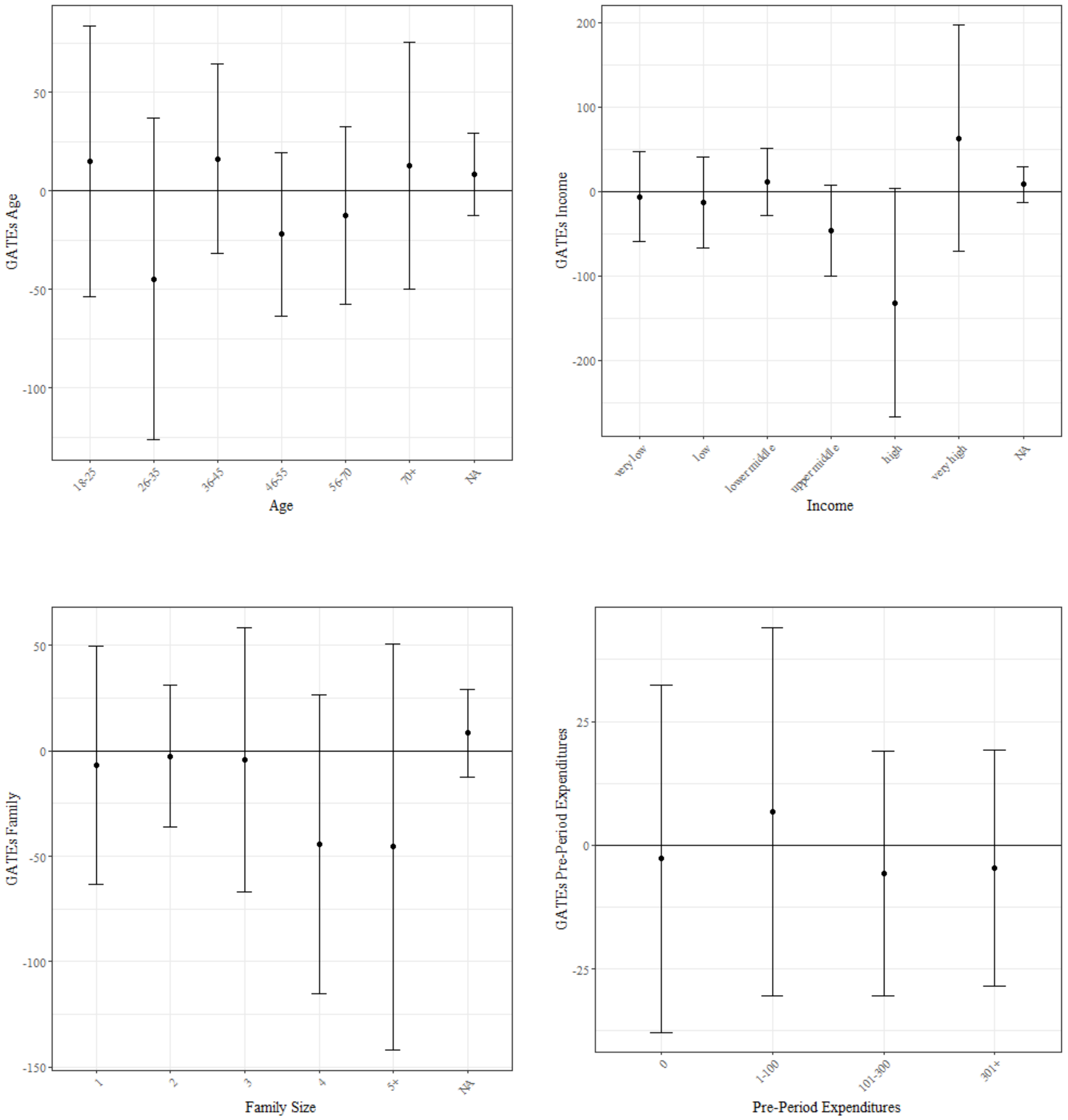


Figure 5.C.1: GATEs of ready-to-eat food coupons with 95% confidence interval.

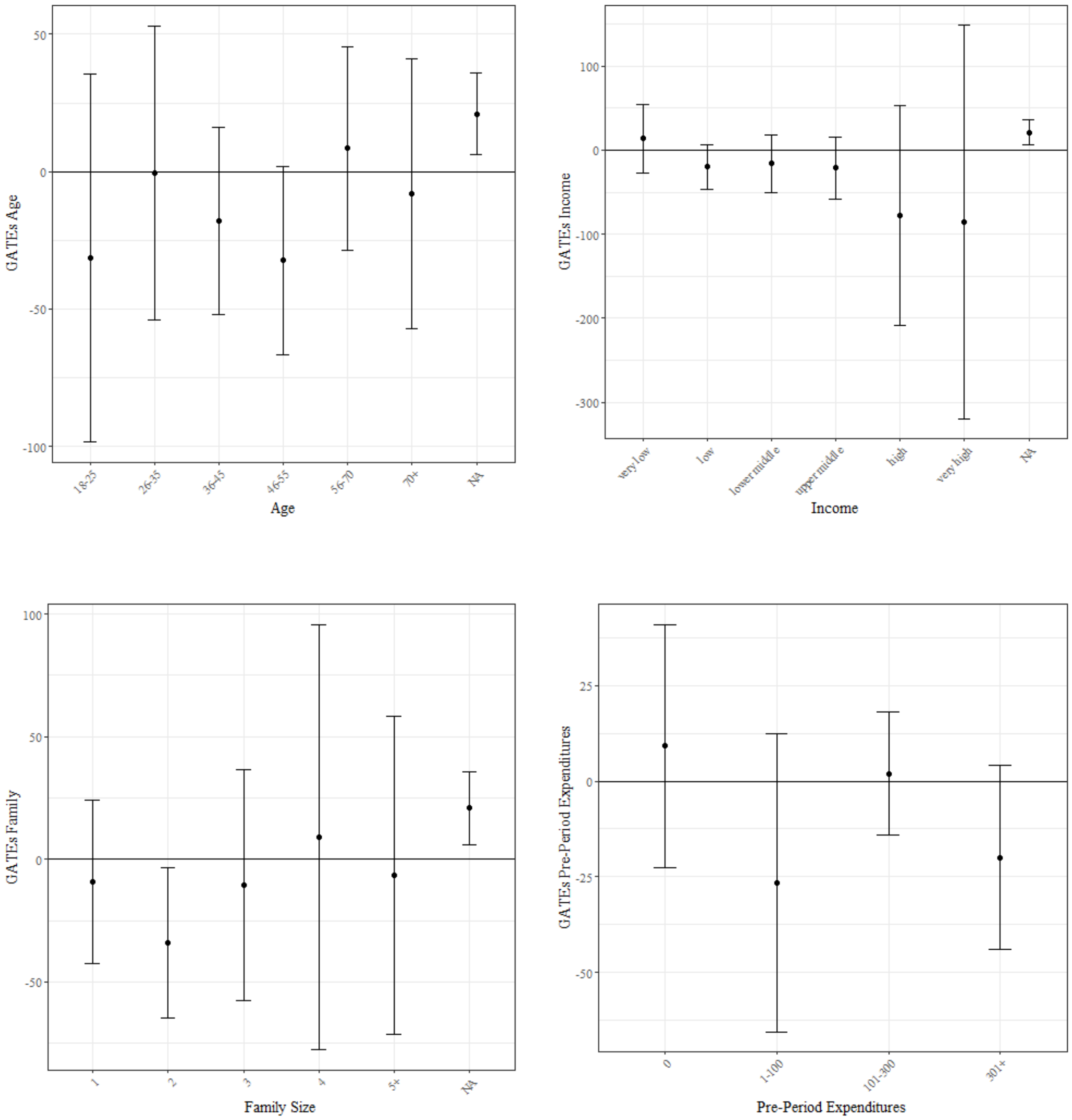


Figure 5.C.2: GATEs of meat/seafood coupons with 95% confidence interval.

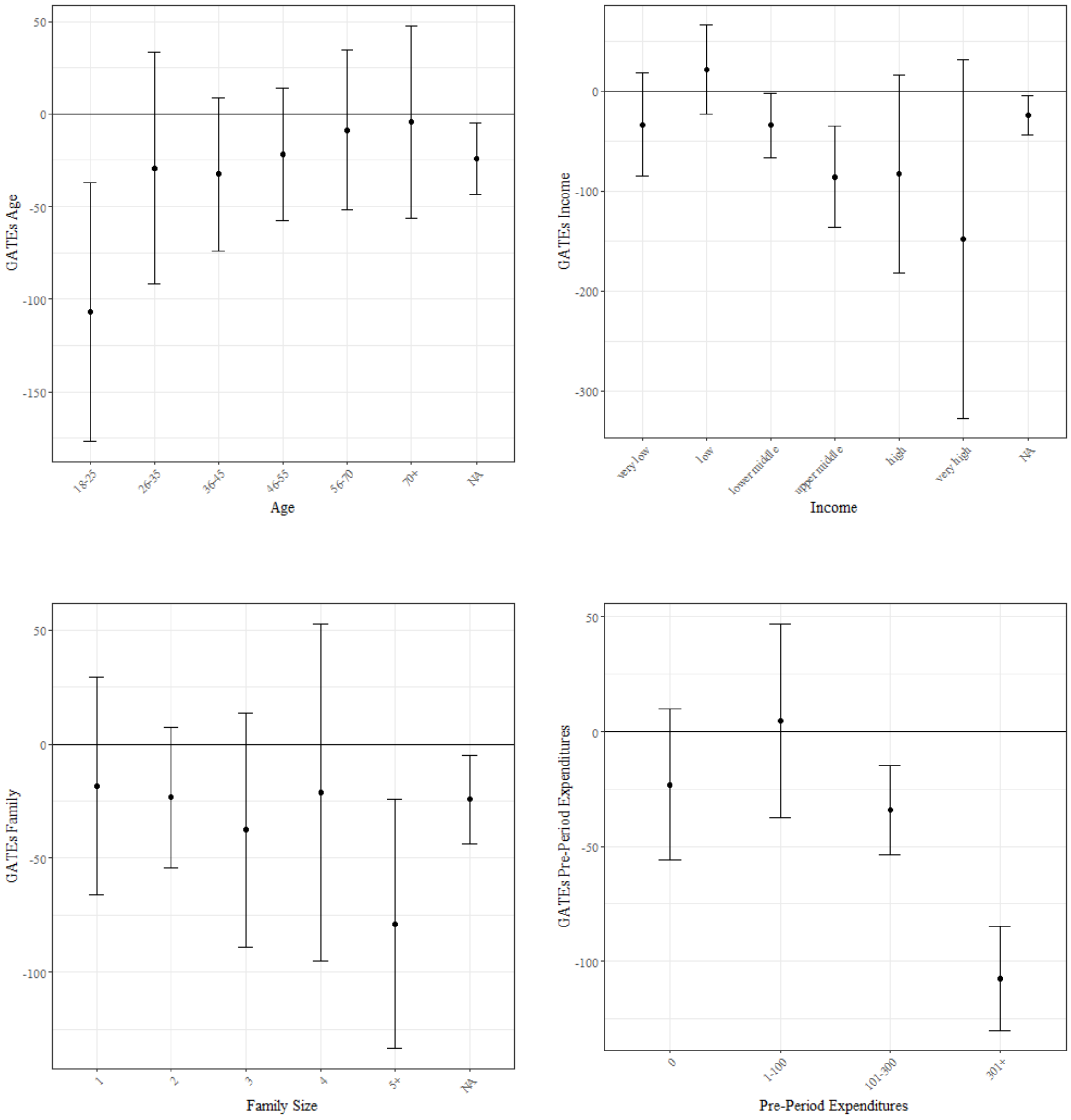


Figure 5.C.3: GATEs of coupons applicable to other non-food products with 95% confidence interval.

5.D Robustness Checks

5.D.1 Reduced Dataset: GATE Estimates

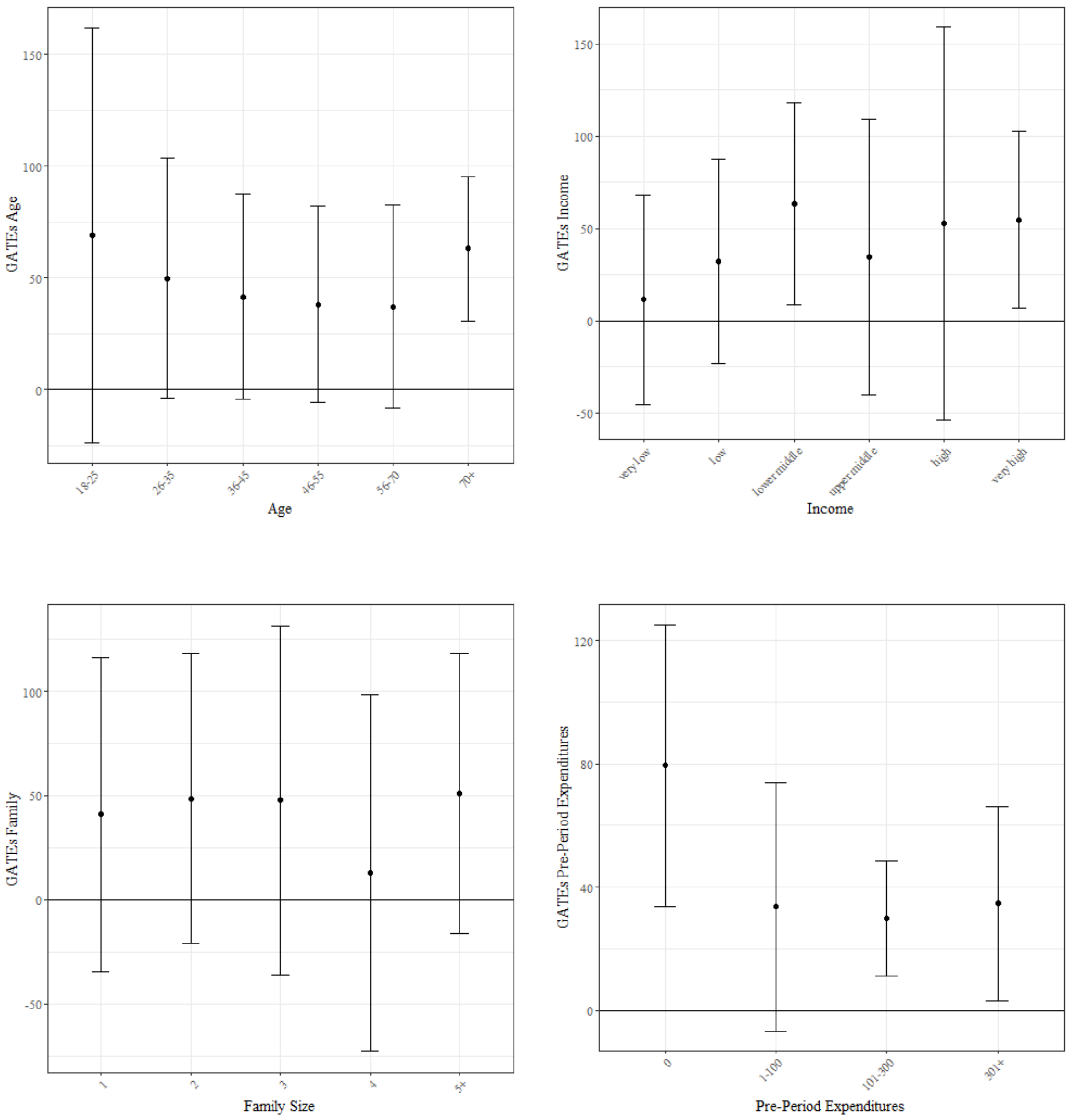


Figure 5.D.1: GATEs of receiving any coupon with 95% confidence interval, estimated in reduced dataset.

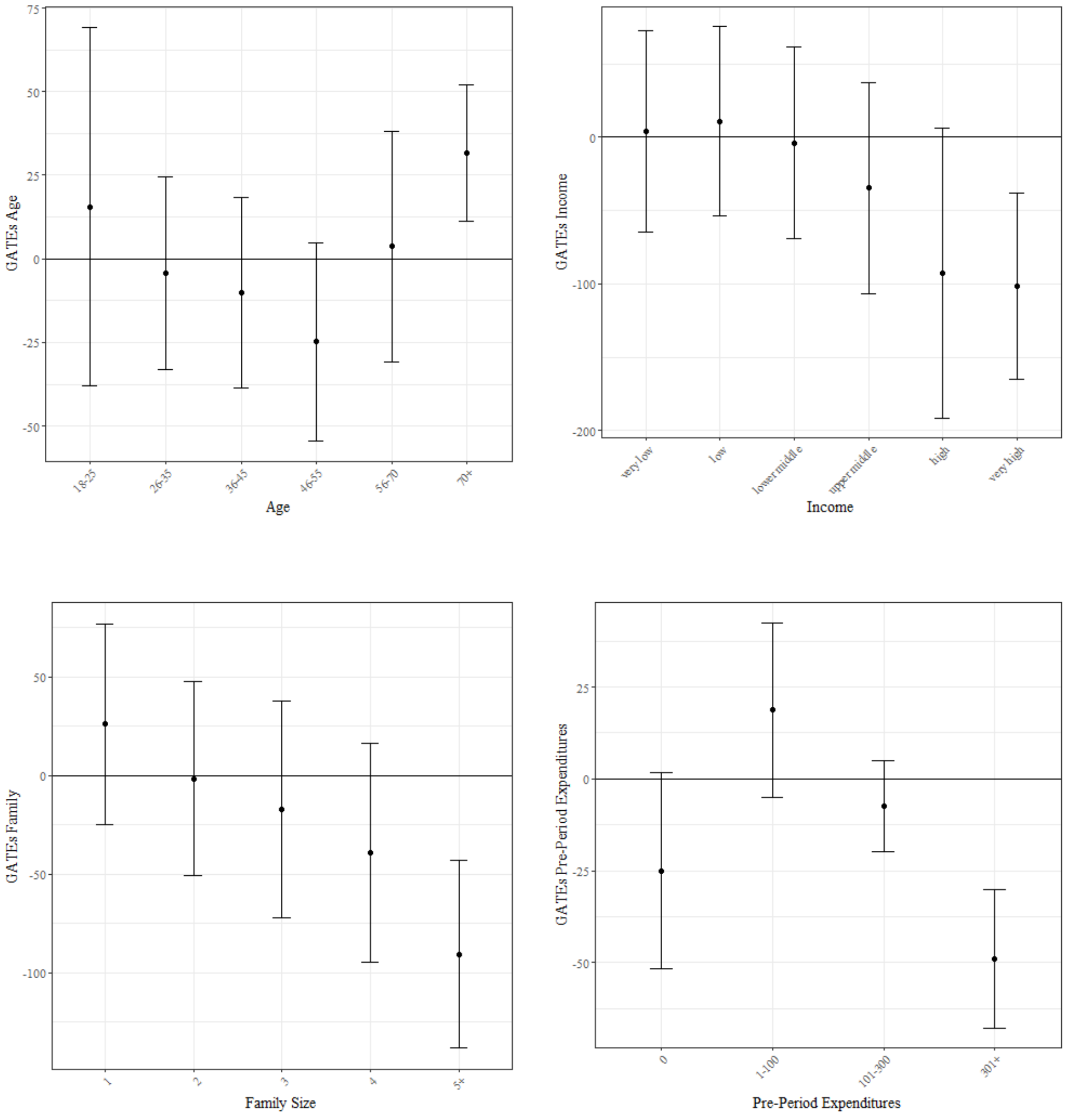


Figure 5.D.2: GATEs of ready-to-eat food coupons with 95% confidence interval, estimated in reduced dataset.

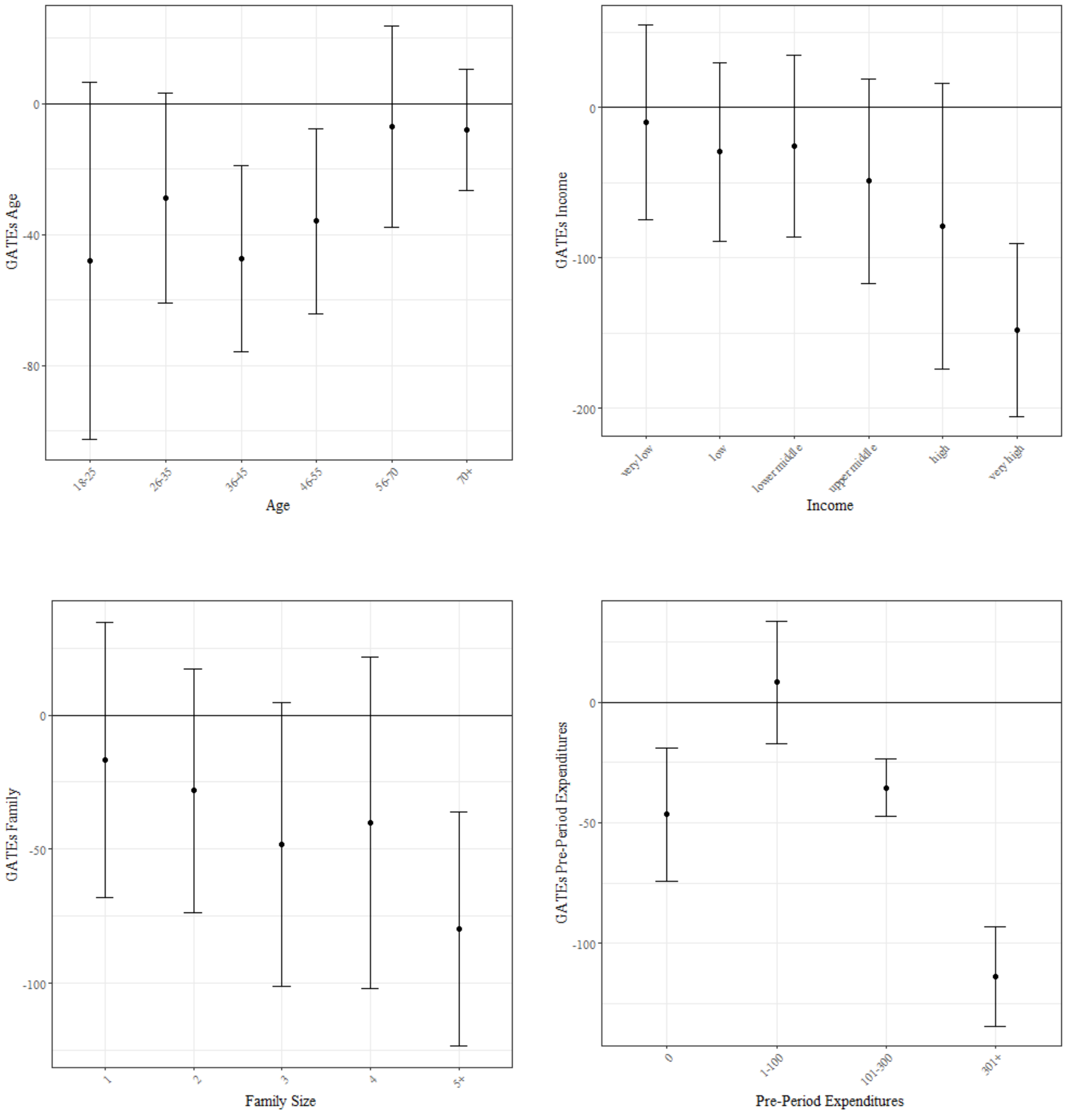


Figure 5.D.3: GATEs of meat and seafood coupons with 95% confidence interval, estimated in reduced dataset.

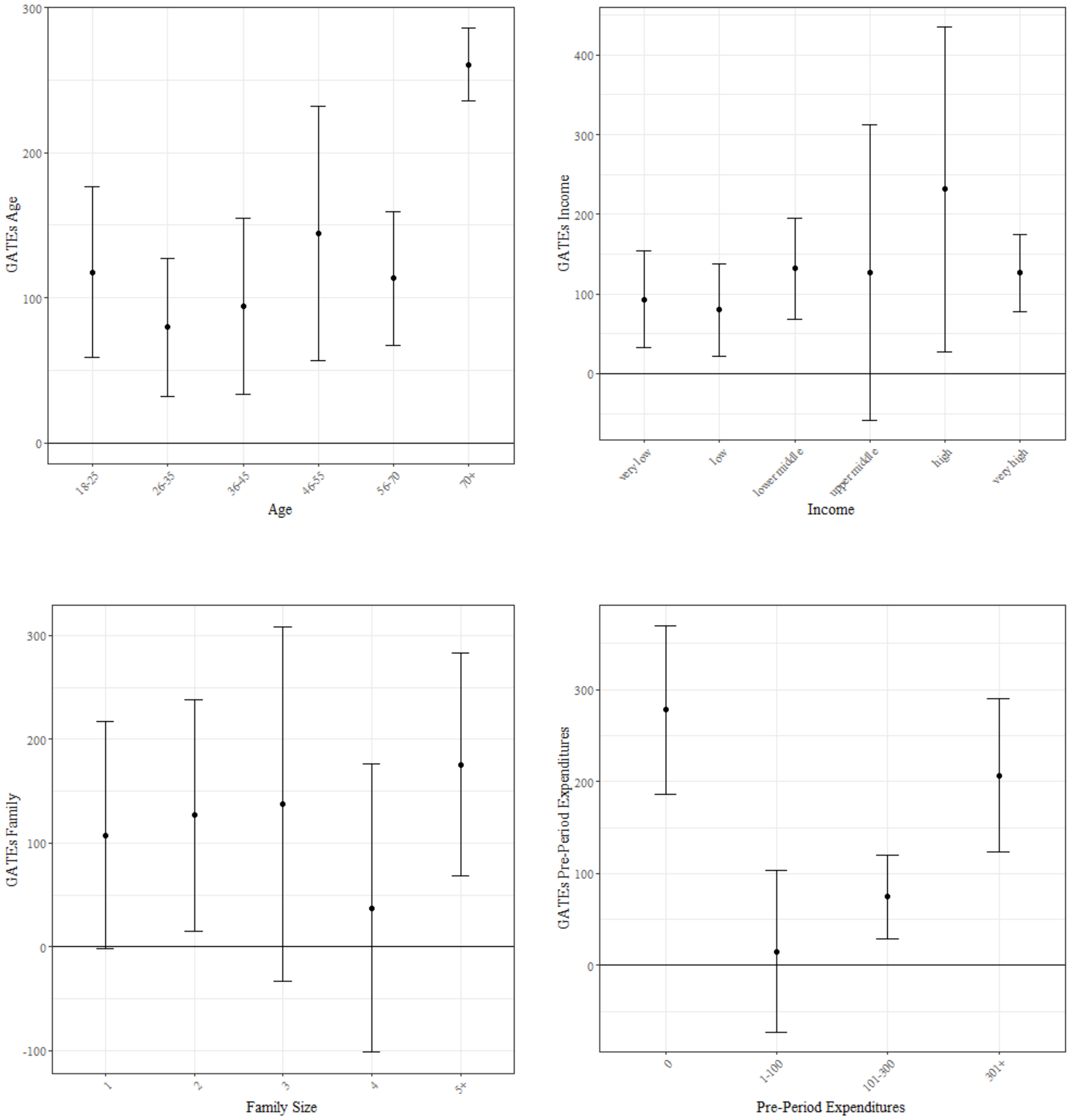


Figure 5.D.4: GATEs of coupons applicable to other food items with 95% confidence interval, estimated in reduced dataset.

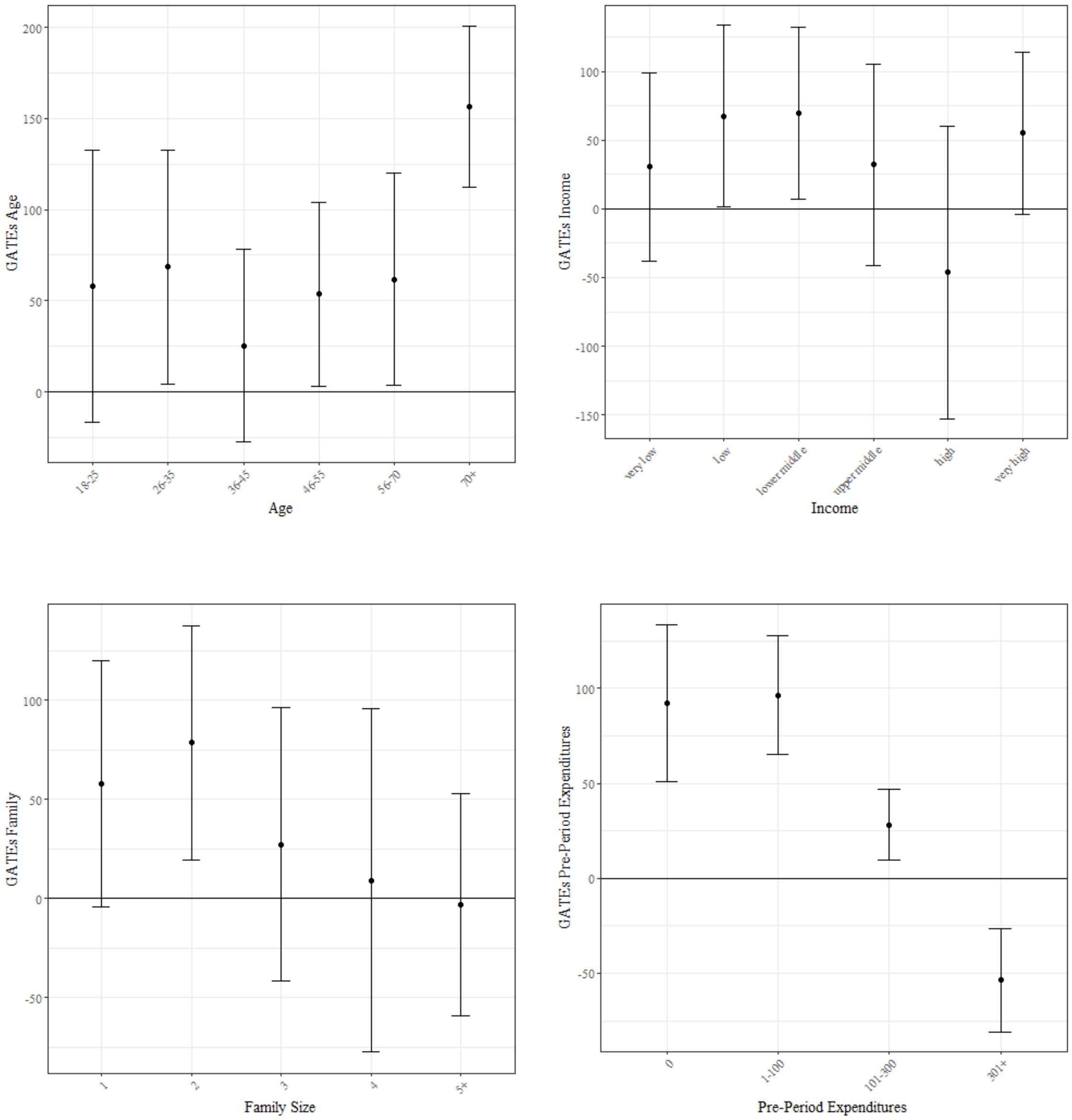


Figure 5.D.5: GATEs of drugstore coupons with 95% confidence interval, estimated in reduced dataset.

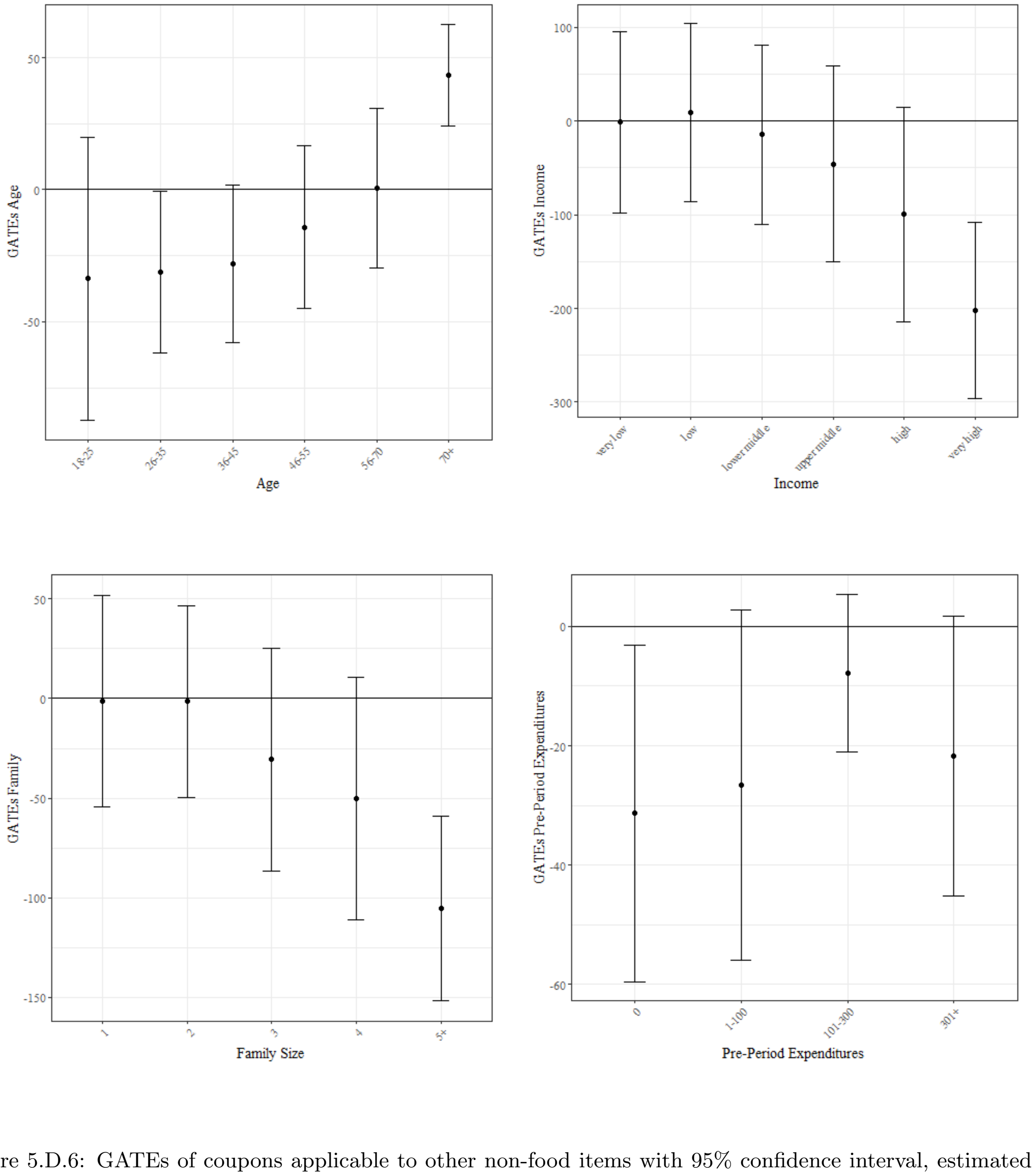
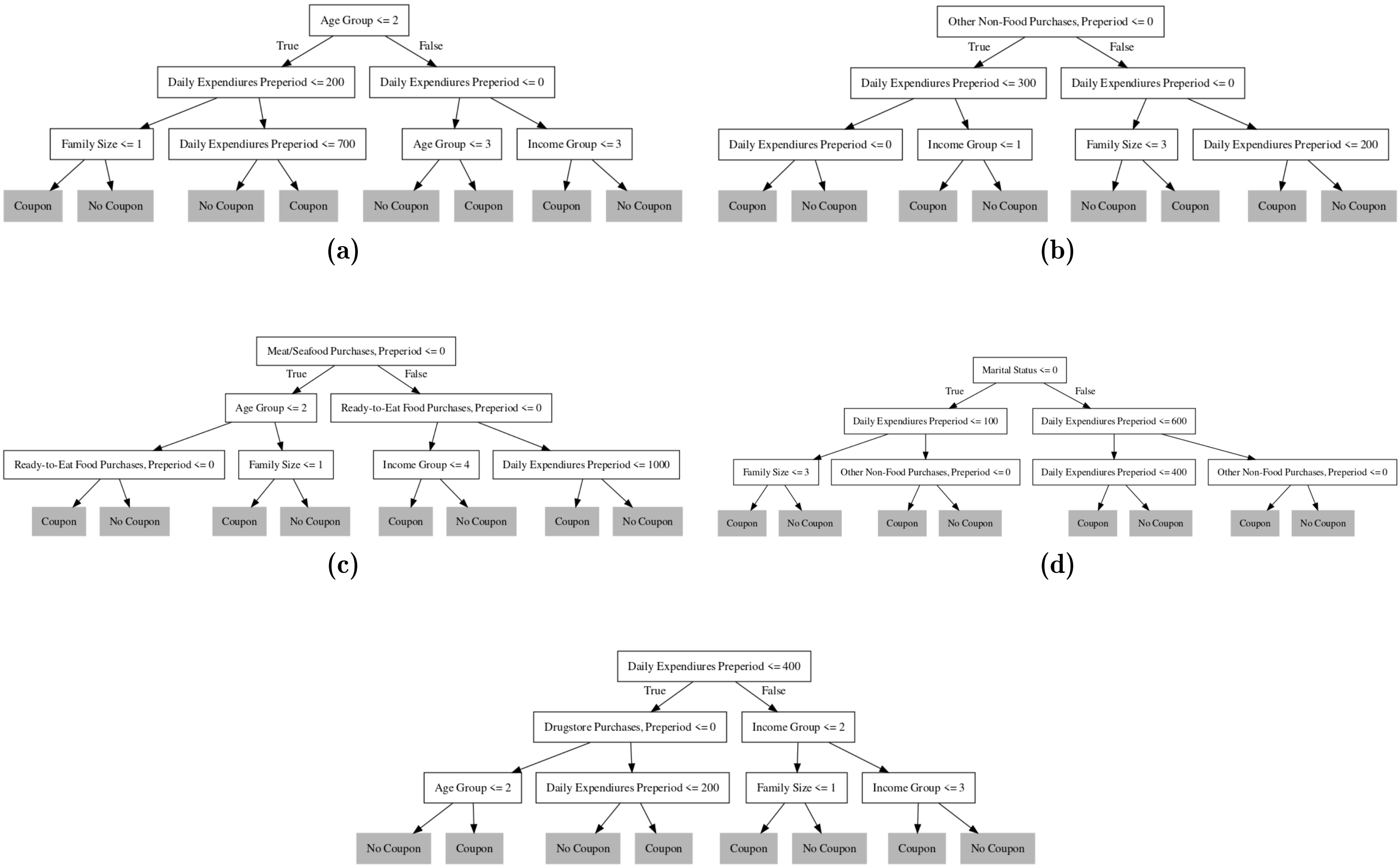


Figure 5.D.6: GATEs of coupons applicable to other non-food items with 95% confidence interval, estimated in reduced dataset.

5.D.2 Reduced Dataset: Policy Trees


Age Group <= 2
True
False
Daily Expendiures Preperiod <= 200
Daily Expendiures Preperiod <= 0
Family Size <= 1
Daily Expendiures Preperiod <= 700
Age Group <= 3
Income Group <= 3
Coupon
No Coupon
No Coupon
Coupon
No Coupon
Coupon
Coupon
No Coupon
Other Non-Food Purchases, Preperiod <= 0
True
False
Daily Expendiures Preperiod <= 300
Daily Expendiures Preperiod <= 0
Daily Expendiures Preperiod <= 0
Income Group <= 1
Family Size <= 3
Daily Expendiures Preperiod <= 200
Coupon
No Coupon
Coupon
No Coupon
No Coupon
Coupon
Coupon
No Coupon
Meat/Seafood Purchases, Preperiod <= 0
True
False
Age Group <= 2
Ready-to-Eat Food Purchases, Preperiod <= 0
Ready-to-Eat Food Purchases, Preperiod <= 0
Family Size <= 1
Income Group <= 4
Daily Expendiures Preperiod <= 1000
Coupon
No Coupon
Coupon
No Coupon
Coupon
No Coupon
Coupon
No Coupon
Marital Status <= 0
True
False
Daily Expendiures Preperiod <= 100
Daily Expendiures Preperiod <= 600
Family Size <= 3
Other Non-Food Purchases, Preperiod <= 0
Daily Expendiures Preperiod <= 400
Other Non-Food Purchases, Preperiod <= 0
Coupon
No Coupon
Coupon
No Coupon
Coupon
No Coupon
Coupon
No Coupon
Daily Expendiures Preperiod <= 400
True
False
Drugstore Purchases, Preperiod <= 0
Income Group <= 2
Age Group <= 2
Daily Expendiures Preperiod <= 200
Family Size <= 1
Income Group <= 3
No Coupon
Coupon
No Coupon
Coupon
Coupon
No Coupon
Coupon
No Coupon

(a)

(b)

(c)

(d)

(e)

Figure 5.D.7: Depth-3 trees for coupons applicable to (a) ready-to-eat food, (b) meat and seafood, (c) other food, (d) drugstore products and (e) other non-food products, estimated in reduced dataset.

Bibliography

ABADIE, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72, 1–19.

ABADIE, A., A. DIAMOND, AND J. HAINMUELLER (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505,.

ABADIE, A., A. J. DIAMOND, AND J. HAINMUELLER (2011): "Synth: An R Package for Synthetic Control Methods in Comparative Case Studies," *Journal of Statistical Software*, 42, 1–17.

ABADIE, A., AND J. GARDEAZABAL (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 1–22.

ABERCROMBIE, G., AND R. T. BATISTA-NAVARRO (2018): "'Aye'or 'no'? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

ACEMOGLU, D., AND J. D. ANGRIST (2001): "Consequences of employment protection? The case of the Americans with Disabilities Act," *Journal of Political Economy*, 109(5), 915–957.

AGGARWAL, C. C., A. HINNEBURG, AND D. A. KEIM (2001): "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*, pp. 420–434. Springer.

AIT BIHI OUALI, L., AND D. J. GRAHAM (2021): "The impact of the MeToo scandal on women's perceptions of security," *Transportation research part A: policy and practice*, 147, 269–283.

ALBERT, J. M. (2008): "Mediation analysis via potential outcomes models," *Statistics in Medicine*, 27, 1282–1304.

ALBERT, J. M., AND S. NELSON (2011): "Generalized causal mediation analysis," *Biometrics*, 67, 1028–1038.

Alekseev, A., A. Katasev, A. Kirillov, A. Khassianov, and D. Zuev (2019): "Prototype of Classifier for the Decision Support System of Legal Documents.," in *SSI*, pp. 328–335.

American Bar Association (2019): "How Courts Works," *https://www.americanbar.org/groups/public_education/resources/law_related_education_network/how_courts_work/appeals/. Retrieved August 15, 2021.*

Amuedo-Dorantes, C., C. Borra, N. R. Garrido, and A. Sevilla (2020): "Timing is Everything when Fighting a Pandemic: COVID-19 Mortality in Spain," *IZA Discussion Paper Series.*

Anderson, C. (2008): "The end of theory: The data deluge makes the scientific method obsolete," *Wired magazine*, 16(7), 16–07.

Anderson, E. T., and D. I. Simester (2004): "Long-run effects of promotion depth on new versus established customers: three field studies," *Marketing Science*, 23(1), 4–20.

Anderson, K. B., H. Cooper, and L. Okamura (1997): "Individual differences and attitudes toward rape: A meta-analytic review," *Personality and Social Psychology Bulletin*, 23(3), 295–315.

Anderson, M., and S. Toor (2018): "How social media users have discussed sexual harassment since #MeToo went viral," *Pew Research Center. url: https://www.pewresearch.org/fact-tank/2018/10/11/how-social-media-users-have-discussed-sexual-harassment-since-metoo-went-viral/.*

Andrews, M., X. Luo, Z. Fang, and A. Ghose (2016): "Mobile ad effectiveness: Hyper-contextual targeting with crowdedness," *Marketing Science*, 35(2), 218–233.

Anitha, J., and M. Kalaiarasu (2021): "Optimized machine learning based collaborative filtering (OMLCF) recommendation system in e-commerce," *Journal of Ambient Intelligence and Humanized Computing*, 12(6), 6387–6398.

Arevalillo, J. M. (2021): "Ensemble learning from model based trees with application to differential price sensitivity assessment," *Information Sciences*, 557, 16–33.

Aronow, P. M., and C. Samii (2017): "Estimating average causal effects under general interference, with application to a social network experiment," *The Annals of Applied Statistics*, 11, 1912–1947.

Askitas, N., K. Tatsiramos, and B. Verheyden (2020): "Lockdown Strategies, Mobility Patterns and COVID-19," *IZA Discussion Paper No. 13293.*

Athey, S., R. Friedberg, V. Hadad, D. Hirshberg, L. Miner, E. Sverdrup, J. Tibshirani, S. Wager, and M. Wright (2022): "generalized random forests (grf 2.1.0)," *https://grf-labs.github.io/grf/index.html.*

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized random forests," *The Annals of Statistics*, 47, 1148–1178.

Athey, S., and S. Wager (2019): "Estimating treatment effects with causal forests: An application," *Observational Studies*, 5(2), 37–51.

——— (2021): "Policy learning with observational data," *Econometrica*, 89(1), 133–161.

Atwater, L. E., A. M. Tringale, R. E. Sturm, S. N. Taylor, and P. W. Braddy (2019): "Looking Ahead: How What We Know About Sexual Harassment Now Informs Us of the Future," *Organizational Dynamics*, 48(4), 100677.

Baicker, K., S. L. Taubman, H. L. Allen, M. Bernstein, J. H. Gruber, J. P. Newhouse, E. C. Schneider, B. J. Wright, A. M. Zaslavsky, and A. N. Finkelstein (2013): "The Oregon experiment—effects of Medicaid on clinical outcomes," *New England Journal of Medicine*, 368(18), 1713–1722.

Banholzer, N., E. van Weenen, B. Kratzwald, A. Seeliger, D. Tschernutter, P. Bottrighi, A. Cenedese, J. P. Salles, S. Feuerriegel, and W. Vach (2020): "Estimating the impact of non-pharmaceutical interventions on documented infections with COVID-19: A cross-country analysis," *medRxiv.*

Baron, R. M., and D. A. Kenny (1986): "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology*, 51, 1173–1182.

Bayer, C., and M. Kuhn (2020): "Intergenerational Ties and Case Fatality Rates: A Cross-Country Analysis," *IZA Discussion Paper Series.*

Bellani, L., and M. Bia (2018): "The long-run effect of childhood poverty and the mediating role of education," *forthcoming in the Journal of the Royal Statistical Society: Series A (Statistics in Society).*

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 85, 233–298.

Bertotti, C., and D. Maxfield (2018): "Most People Are Supportive of #MeToo. But Will Workplaces Actually Change?," *Harvard Business Review, July 10, 2018.*

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How much should we trust differences-in-differences estimates?," *The Quarterly journal of economics*, 119(1), 249–275.

Bicher, M. R., C. Rippinger, C. Urach, D. Brunmeir, and N. Popper (2020): "Agent-Based Simulation for Evaluation of Contact-Tracing Policies Against the Spread of SARS-CoV-2," *medRxiv.*

Bieneck, S., and B. Krahé (2011): "Blaming the victim and exonerating the perpetrator in cases of rape and robbery: Is there a double standard?," *Journal of interpersonal violence*, 26(9), 1785–1797.

Bijwaard, G. E., and A. M. Jones (2018): "An IPW estimator for mediation effects in hazard models: with an application to schooling, cognitive ability and mortality," *Empirical Economics*, pp. 1–47.

Biswas, A., S. Bhowmick, A. Guha, and D. Grewal (2013): "Consumer evaluations of sale prices: role of the subtraction principle," *Journal of Marketing*, 77(4), 49–66.

Bodory, H., and M. Huber (2018): "The causalweight package for causal inference in R," *SES Working Paper 493, University of Fribourg.*

Bodory, H., M. Huber, and L. Lafférs (2020): "Evaluating (weighted) dynamic treatment effects by double machine learning," *arXiv preprint arXiv:2012.00370.*

Bohner, G. (2001): "Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim," *British Journal of Social Psychology*, 40(4), 515–529.

Bonardi, J.-P., Q. Gallea, D. Kalanoski, and R. Lalive (2020): "Fast and local: How did lockdown policies affect the spread and severity of the COVID-19?," *working paper, University of Lausanne.*

Boux, H. J. (2016): *Sexual assault jurisprudence: Rape myth usage in state appellate courts.* Georgetown University.
Bureau of Labor Statistics at the U.S. Department of Labor

Bureau of Labor Statistics at the U.S. Department of Labor (2019): "National Longitudinal Survey of Youth 1997 cohort, 1997-2017 (rounds 1-18).," *Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. url: https://www.nlsinfo.org/investigator/pages/search.*

Burstin, H. R., K. Swartz, A. C. O'Neil, E. J. Orav, and T. A. Brennan (1998): "The effect of change of health insurance on access to care," *Inquiry*, pp. 389–397.

Busso, M., J. DiNardo, and J. McCrary (2009): "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *IZA Discussion Paper No. 3998.*

Cagala, T., U. Glogowsky, J. Rincke, and A. Strittmatter (2021): "Optimal Targeting in Fundraising: A Causal Machine-Learning Approach," *arXiv preprint arXiv:2103.10251.*

Caputi, T. L., A. L. Nobles, and J. W. Ayers (2019): "Internet Searches for Sexual Harassment and Assault, Reporting, and Training Since the #MeToo Movement," *JAMA Internal Medicine*, 179(2), 258–259.

Card, D., and A. B. Krueger (1994): "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *The American Economic Review*, 84(4), 772–793.

Cardella, E., and B. Depew (2014): "The effect of health insurance coverage on the reported health of young adults," *Economics Letters*, 124(3), 406–410.

CAREERARC (2020): "Survey: 76% of Employed Americans Say #MeToo Positively Impacted How Sexual Harassment Is Addressed In The Workplace, While More Than 2 In 5 Say It Has Damaged Trust Between HR and Employees," *Press Release, October 3, 2020.*

CASTLE, J. J., S. JENKINS, C. D. ORTBALS, L. POLONI-STAUDINGER, AND J. C. STRACHAN (2020): "The Effect of the #MeToo Movement on Political Engagement and Ambition in 2018," *Political Research Quarterly*, 73(4).

CHANDRASEKHARAN, E., U. PAVALANATHAN, A. SRINIVASAN, A. GLYNN, J. EISENSTEIN, AND E. GILBERT (2017): "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 1–22.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

CHERNOZHUKOV, V., H. KASAHA, AND P. SCHRIMPF (2020): "Causal impact of masks, policies, behavior on early COVID-19 pandemic in the US," *medRxiv.*

CHOI, P., AND K. S. COULTER (2012): "It's not all relative: the effects of mental and physical positioning of comparative prices on absolute versus relative discount assessment," *Journal of Retailing*, 88(4), 512–527.

COCHRAN, W. G. (1957): "Analysis of Covariance: Its Nature and Uses," *Biometrics*, 13, 261–281.

CONTI, G., J. J. HECKMAN, AND R. PINTO (2016): "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviour," *The Economic Journal*, 126, F28–F65.

COWLS, J., AND R. SCHROEDER (2015): "Causation, correlation, and big data in social science research," *Policy & Internet*, 7(4), 447–472.

DANAHER, P. J., M. S. SMITH, K. RANASINGHE, AND T. S. DANAHER (2015): "Where, when, and how long: Factors that influence the redemption of mobile phone coupons," *Journal of Marketing Research*, 52(5), 710–725.

DAVE, D., A. I. FRIEDSON, K. MATSUZAWA, AND J. J. SABIA (2020): "When Do Shelter-In-Place Orders Fight COVID-19 Best? Policy Heterogeneity across States and Adoption Time," *IZA Discussion Paper No. 13190.*

DAVE, D., A. I. FRIEDSON, K. MATSUZAWA, J. J. SABIA, AND S. SAFFORD (2020): "Were Urban Cowboys Enough to Control COVID-19? Local Shelter-in-Place Orders and Coronavirus Case Growth," .

DAVIES, M., P. ROGERS, AND L. WHITELEGG (2009): "Effects of victim gender, victim sexual orientation, victim response and respondent gender on judgements of blame in a hypothetical adolescent rape," *Legal and Criminological Psychology*, 14(2), 331–338.

DEJMANEE, T., Z. ZAHER, R. SAMANTHA, AND M. J. PAPA (2020): "# MeToo;# HimToo: Popular Feminism and Hashtag Activism in the Kavanaugh Hearings," *International Journal of Communication.*

DESHPANDE, S., Z. LI, AND V. KULESHOV (2022): "Multi-Modal Causal Inference with Deep Structural Equation Models," *arXiv preprint arXiv:2203.09672.*

DÍAZ, I. (2020): "Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning," *Biostatistics*, 21(2), 353–358.

DÍAZ, I., AND N. S. HEJAZI (2020): "Causal mediation analysis for stochastic interventions," 82(3), 661–683.

DONNELLY, R., F. J. RUIZ, D. BLEI, AND S. ATHEY (2021): "Counterfactual inference for consumer choice across many product categories," *Quantitative Marketing and Economics*, pp. 1–39.

DONSIMONI, J. R., R. GLAWION, B. PLACHTER, C. WEISER, AND K. WÄLDE (2020): "Should Contact Bans Be Lifted in Germany? A Quantitative Prediction of Its Effects," *IZA Discussion Paper Series.*

EGAMI, N., C. J. FONG, J. GRIMMER, M. E. ROBERTS, AND B. M. STEWART (2018): "How to make causal inferences using texts," *arXiv preprint arXiv:1802.02163.*

Ehrlich, S. (2012): "Perpetuating—and resisting—rape myths in trial discourse," *Sexual Assault in Canada*, pp. 389–408.

Ennis, E., and L. Wolfe (2018): "Media and #MeToo: How a movement affected press coverage of sexual assault," *Women's Media Center Report.*

Erevelles, S., N. Fukawa, and L. Swayne (2016): "Big Data consumer analytics and the transformation of marketing," *Journal of business research*, 69(2), 897–904.

Evans, A. (2018): "# MeToo: A study on sexual assault as reported in the New York Times," *Occam's Razor*, 8(1), 3.

Farbmacher, H., M. Huber, L. Lafférs, H. Langen, and M. Spindler (2022): "Causal mediation analysis with double machine learning," *The Econometrics Journal*, 25(2), 277–300.

Farrow, R. (2017): "From Aggressive Overtures to Sexual Assault: Harvey Weinstein's Accusers Tell Their Stories," *The New Yorker October 10, 2017.*

Faulkner, L., and H. Schauffler (1997): "The Effect of Health Insurance Coverage on the Appropriate Use f Recommended Clinical Preventive Services," *American Journal of Preventive Medicine*, 13(6), 453–458.

Feuerriegel, S., S. F. Heitzmann, and D. Neumann (2015): "Do investors read too much into news? How news sentiment causes price formation," in *48th Hawaii International Conference on System Sciences*, pp. 4803–4812. IEEE.

Field, A., C. Y. Park, and Y. Tsvetkov (2020): "Controlled Analyses of Social Biases in Wikipedia Bios," *arXiv preprint arXiv:2101.00078.*

Fileborn, B., and R. Loney-Howes (2019): "Introduction: Mapping the emergence of #MeToo," in *#MeToo and the Politics of Social Change*, pp. 1–18. Springer.

Filtz, E., S. Kirrane, A. Polleres, and G. Wohlgenannt (2019): "Exploiting eurovoc's hierarchical structure for classifying legal documents," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 164–181. Springer.

Flores, C. A., and A. Flores-Lagunes (2009): "Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness," *IZA DP No. 4237.*

FONG, C., AND J. GRIMMER (2016): "Discovery of treatments from text corpora," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1600–1609.

FOWLER, J. H., S. J. HILL, N. OBRADOVICH, AND R. LEVIN (2020): "The Effect of Stay-at-Home Orders on COVID-19 Cases and Fatalities in the United States," *medRxiv preprint*.

FOWLER-BROWN, A., G. CORBIE-SMITH, J. GARRETT, AND N. LURIE (2007): "Risk of cardiovascular events and death—does insurance matter?," *Journal of General Internal Medicine*, 22(4), 502–507.

FRANIUK, R., J. L. SEEFELT, S. L. CEPRESS, AND J. A. VANDELLO (2008): "Prevalence and effects of rape myths in print journalism: The Kobe Bryant case," *Violence against women*, 14(3), 287–309.

FRENCH, M. T., K. MORTENSEN, AND A. R. TIMMING (2021): "A multivariate analysis of workplace mentoring and socializing in the wake of #MeToo," *Applied Economics*, pp. 1–19.

FRERMANN, L., AND M. LAPATA (2016): "A bayesian model of diachronic meaning change," *Transactions of the Association for Computational Linguistics*, 4, 31–45.

FRIEDSON, A., D. MCNICHOLS, J. J. SABIA, AND D. DAVE (2020): "Did California's Shelter-in-Place Order Work? Early Coronavirus-Related Public Health Effects.," *NBER Working Paper No. 26992*.

GHAG, K. V., AND K. SHAH (2015): "Comparative analysis of effect of stopwords removal on sentiment classification," in *2015 international conference on computer, communication and control (IC4)*, pp. 1–6. IEEE.

GLYNN, A. N., AND K. M. QUINN (2010): "An introduction to the augmented inverse propensity weighted estimator," *Political analysis*, 18(1), 36–56.

GOLDER, S. A., AND M. W. MACY (2014): "Digital footprints: Opportunities and challenges for online social research," *Annual Review of Sociology*, 40, 129–152.

GOPALAKRISHNAN, A., AND Y.-H. PARK (2021): "The impact of coupons on the visit-to-purchase funnel," *Marketing Science*, 40(1), 48–61.

Gordini, N., and V. Veglio (2017): "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry," *Industrial Marketing Management*, 62, 100–107.

Gordon, B. R., R. Moakler, and F. Zettelmeyer (2022): "Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement," *arXiv preprint arXiv:2201.07055*.

Gravelin, C. R., M. Biernat, and C. E. Bucher (2019): "Blaming the victim of acquaintance rape: Individual, situational, and sociocultural factors," *Frontiers in psychology*, 9, 2422.

Greenfield, R. (2018): "Powerful Men Have Changed Their Behavior at Work Since #MeToo," *Blomberg, April 10, 2018*.

Greenstein-Messica, A., L. Rokach, and A. Shabtai (2017): "Personal-discount sensitivity prediction for mobile coupon conversion optimization," *Journal of the Association for Information Science and Technology*, 68(8), 1940–1952.

Grubb, A., and E. Turner (2012): "Attribution of blame in rape cases: A review of the impact of rape myth acceptance, gender role conformity and substance use on victim blaming," *Aggression and violent behavior*, 17(5), 443–452.

Guo, T., S. Sriram, and P. Manchanda (2021): "The effect of information disclosure on industry payments to physicians," *Journal of Marketing Research*, 58(1), 115–140.

HaCohen-Kerner, Y., D. Miller, and Y. Yigal (2020): "The influence of preprocessing on text classification using a bag-of-words representation," *PloS one*, 15(5), e0232525.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Hair Jr, J. F., and M. Sarstedt (2021): "Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing," *Journal of Marketing Theory and Practice*, 29(1), 65–77.

Halvorsen, A., H. N. Koutsopoulos, S. Lau, T. Au, and J. Zhao (2016): "Reducing subway crowding: analysis of an off-peak discount experiment in Hong Kong," *Transportation Research Record*, 2544(1), 38–46.

Hamilton, W. L., J. Leskovec, and D. Jurafsky (2016): "Cultural shift or linguistic drift? comparing two computational measures of semantic change," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016, p. 2116. NIH Public Access.

Harper, J. (2016): "Remember the Common Law," *CATO Insitute - Policy report.*

Hausladen, C. I., M. H. Schubert, and E. Ash (2020): "Text classification of ideological direction in judicial opinions," *International Review of Law and Economics*, 62, 105903.

He, J., and W. Jiang (2017): "Understanding Users' Coupon Usage Behaviors in E-Commerce Environments," in *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pp. 1047–1053. IEEE.

Heckman, J., R. Pinto, and P. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 2052–2086.

Hellrich, J., S. Buechel, and U. Hahn (2018): "Modeling word emotion in historical language: Quantity beats supposed stability in seed word selection," *arXiv preprint arXiv:1806.08115.*

Henley, N. M., M. Miller, and J. A. Beazley (1995): "Syntax, semantics, and sexual violence: Agency and the passive voice," *Journal of Language and Social Psychology*, 14(1-2), 60–84.

Hirano, K., and J. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77, 1683–1701.

Holland, P. W. (1986): "Statistics and causal inference," *Journal of the American statistical Association*, 81(396), 945–960.

Hong, G. (2010): “Ratio of mediator probability weighting for estimating natural direct and indirect effects,” in *Proceedings of the American Statistical Association, Biometrics Section*, p. 2401–2415. Alexandria, VA: American Statistical Association.

Hong, G., and S. W. Raudenbush (2006): “Evaluating Kindergarten Retention Policy,” *Journal of the American Statistical Association*, 101, 901–910.

Hsieh, C.-T., S. Shimizutani, and M. Hori (2010): “Did Japan’s shopping coupon program increase spending?,” *Journal of Public Economics*, 94(7-8), 523–529.

Hu, M., C. Dang, and P. K. Chintagunta (2019): “Search and learning at a daily deals website,” *Marketing Science*, 38(4), 609–642.

Huber, M. (2014): “Identifying causal mechanisms (primarily) based on inverse probability weighting,” *Journal of Applied Econometrics*, 29, 920–943.

Huber, M. (2015): “Causal pitfalls in the decomposition of wage gaps,” *Journal of Business and Economic Statistics*, 33, 179–191.

Huber, M., and H. Langen (2020): “Timing matters: the impact of response measures on COVID-19-related hospitalization and death rates in Germany and Switzerland,” *Swiss Journal of Economics and Statistics*, 156(1), 1–19.

Huber, M., M. Lechner, and G. Mellace (2017): “Why Do Tougher Caseworkers Increase Employment? The Role of Program Assignment as a Causal Mechanism,” *The Review of Economics and Statistics*, 99, 180–183.

Huber, M., M. Lechner, and A. Strittmatter (2018): “Direct and indirect effects of training vouchers for the unemployed,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 441–463.

Huber, M., J. Meier, and H. Wallimann (2021): “Business analytics meets artificial intelligence: Assessing the demand effects of discounts on Swiss train tickets,” *arXiv preprint arXiv:2105.01426*.

Huber, M., and A. Steinmayr (2021): “A Framework for Separating Individual-Level Treatment Effects From Spillover Effects,” *Journal of Business & Economic Statistics*, 39, 422–436.

Hudgens, M. G., and M. E. Halloran (2008): "Toward Causal Inference With Interference," *Journal of the American Statistical Association*, 103, 832–842.

Hünermund, P., J. Kaminski, and C. Schmitt (2021): "Causal Machine Learning and Business Decision Making," *Available at SSRN 3867326.*

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imai, K., and T. Yamamoto (2013): "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments," *Political Analysis*, 21, 141–171.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

——— (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Inman, J. J., and L. McAlister (1994): "Do coupon expiration dates affect consumer behavior?," *Journal of Marketing Research*, 31(3), 423–428.

Jackson, C., and M. Newall (2018): "The #MeToo Movement: One Year Later," *Ipsos.*

Jatowt, A., and K. Duh (2014): "A framework for analyzing semantic change of words across time," in *IEEE/ACM Joint Conference on Digital Libraries*, pp. 229–238. IEEE.

Jernberg, F., A. Lindbäck, and A. Roos (2020): "A new male entrepreneur? Media representation of male entrepreneurs before and after #metoo," *Gender in Management: An International Journal.*

Jia, H., S. Yang, X. Lu, and C. W. Park (2018): "Do consumers always spend more when coupon face value is larger? The inverted U-shaped effect of coupon face value on consumer spending level," *Journal of Marketing*, 82(4), 70–85.

Johnson, A., R. Sekaran, and S. Gombar (2020): "2020 Progress Update: MeToo workplace reforms in the states," .

Judd, C. M., and D. A. Kenny (1981): "Process Analysis: Estimating Mediation in Treatment Evaluations," *Evaluation Review*, 5, 602–619.

Juranek, S., and F. T. Zoutman (2020): "The Effect of Social Distancing Measures on Intensive Care Occupancy: Evidence on COVID-19 in Scandinavia," *FOR Discussion Paper 2/20, NHH Norwegion School of Economics.*

Kagal, N., L. Cowan, and H. Jawad (2019): "Beyond the Bright Lights: Are Minoritized Women Outside the Spotlight Able to Say #MeToo?," in *# MeToo and the Politics of Social Change*, pp. 133–149. Springer.

Kaggle (2019): "Predicting Coupon Redemption," *https://www.kaggle.com/vasudeva009/predicting-coupon-redemption.*

Kanamori, T., T. Suzuki, and M. Sugiyama (2012): "Statistical analysis of kernel-based least-squares density-ratio estimation," *Machine Learning*, 86(3), 335–367.

Kantor, J., and M. Twohey (2017): "Harvey Weinstein Paid Off Sexual Harassment Accusers for Decades," *The New York Times, October 5, 2017.*

Kaufman, J. S., R. F. MacLehose, and S. Kaufman (2004): "A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation," *Epidemiologic Perspectives & Innovations*, 1, 4.

Keele, L., D. Tingley, and T. Yamamoto (2015): "Identifying mechanisms behind policy interventions via causal mediation analysis," *Journal of Policy Analysis and Management*, 34, 937–963.

Keith, K. A., D. Jensen, and B. O'Connor (2020): "Text and causal inference: A review of using text to remove confounding from causal estimates," *arXiv preprint arXiv:2005.00649.*

Khalid, O., and P. Srinivasan (2020): "Style matters! Investigating linguistic style in online communities," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 360–369.

KHARDE, V., P. SONAWANE, ET AL. (2016): “Sentiment analysis of twitter data: a survey of techniques,” *arXiv preprint arXiv:1601.06971*.

KING, G., E. GAKIDOU, K. IMAI, J. LAKIN, R. T. MOORE, C. NALL, N. RAVISHANKAR, M. VARGAS, M. M. TELLEZ-ROJO, J. E. H. AVILA, ET AL. (2009): “Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme,” *The Lancet*, 373(9673), 1447–1454.

KITAGAWA, T., AND A. TETENOV (2018): “Who should be treated? Empirical welfare maximization methods for treatment choice,” *Econometrica*, 86, 591–616.

KLAR, S., AND A. MCCOY (2021): “The #MeToo movement and attitudes toward President Trump in the wake of a sexual misconduct allegation,” *Politics, Groups, and Identities*, 0(0), 1–10.

KOEHN, D., S. LESSMANN, AND M. SCHAAL (2020): “Predicting online shopping behaviour from clickstream data using deep learning,” *Expert Systems with Applications*, 150, 113342.

KOO, COOK, PARK, SUN, SUN, LIM, TAM, AND DICKENS (2020): “Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study,” in *The Lancet Infectious Diseases*, ed. by Elsevier.

KRISHNA, A., AND Z. J. ZHANG (1999): “Short-or long-duration coupons: The effect of the expiration date on the profitability of coupon promotions,” *Management Science*, 45(8), 1041–1056.

KSENIA KEPLINGER, STEFANIE K. JOHNSON, J. F. K., AND L. Y. BARNES (2019): “Women at work: Changes in sexual harassment between September 2016 and September 2018,” *PloS one*, 14(7).

KULKARNI, V., R. AL-RFOU, B. PEROZZI, AND S. SKIENA (2015): “Statistically significant detection of linguistic change,” in *Proceedings of the 24th International Conference on World Wide Web*, pp. 625–635.

LAMBERT, A. J., AND K. RAICHLE (2000): “The role of political ideology in mediating judgments of blame in rape victims and their assailants: A test of the just world, personal responsibility, and legitimization hypotheses," *Personality and Social Psychology Bulletin*, 26(7), 853–863.

Lechner, M. (2001): "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner, and F. Pfeiffer. Heidelberg: Physica.

Leone, R. P., and S. S. Srinivasan (1996): "Coupon face value: Its impact on coupon redemptions, brand sales, and brand profitability," *Journal of retailing*, 72(3), 273–289.

Levy, R., and M. Mattsson (2021): "The effects of social movements: Evidence from# MeToo," *Available at SSRN 3496903.*

Luk, C., K. Choy, and H. Lam (2019): "Design of an intelligent customer identification model in e-Commerce logistics industry," in *MATEC Web of Conferences*, vol. 255, p. 04003. EDP Sciences.

Luo, Y., and M. Spindler (2016): "High-Dimensional L_2Boosting: Rate of Convergence," .

Lycett, M. (2013): "'Datafication': making sense of (big) data in a complex world," .

Ma, L., and B. Sun (2020): "Machine learning and AI in marketing–Connecting computing power to human insights," *International Journal of Research in Marketing*, 37(3), 481–504.

Maciosek, M. V., A. B. Coffield, T. J. Flottemesch, N. M. Edwards, and L. I. Solberg (2010): "Greater use of preventive services in US health care could save lives at little or no cost," *Health Affairs*, 29(9), 1656–1660.

Manski, C. F. (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

Mariani, M. M., R. Perez-Vega, and J. Wirtz (2021): "AI in marketing, consumer research and psychology: A systematic literature review and research agenda," *Psychology & Marketing.*

McCoy, V. L. (2004): *The effect of language used in newspaper report of rape: Measuring readers' judgments of victim blame.* The University of Tulsa.

McGregor, J. (2019): "#MeToo backlash: More male managers avoid mentoring women or meeting alone with them," *Washington Post, May 17, 2019.*

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013): "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119.

Miles, C. H., I. Shpitser, Kanki, S. Meloni, and E. J. Tchetgen Tchetgen (2020): "On semiparametric estimation of a path-specific effect in the presence of mediator-outcome confounding," *Biometrika*, 107(1), 159–172.

Mitze, T., R. Kosfeld, J. Rode, and K. Wälde (2020): "Face Masks Considerably Reduce COVID-19 Cases in Germany: A Synthetic Control Method Approach," .

Moricz, S. (2019): "Using Artificial Intelligence to Recapture Norms: Did# metoo Change Gender Norms in Sweden?," *arXiv preprint arXiv:1903.00690*.

Mozer, R., L. Miratrix, A. R. Kaufman, and L. J. Anastasopoulos (2020): "Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality," *Political Analysis*, 28(4), 445–468.

Mueller, A., Z. Wood-Doughty, S. Amir, M. Dredze, and A. L. Nobles (2021): "Demographic representation and collective storytelling in the me too Twitter hashtag activism movement," *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–28.

Mustak, M., J. Salminen, L. Plé, and J. Wirtz (2021): "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda," *Journal of Business Research*, 124, 389–404.

Myers, J. (2020): "The Policy Implications of Social Movement: How #MeToo can bring change," *Sociological Viewpoints*, 34(1), 138–156.

Nagelkerke, N. J. D. (1991): "A note on a general definition of the coefficient of determination," *Biometrika*, 78, 691–692.

Nakanishi, N., K. Tatara, and H. Fujiwara (1996): "Do preventive health services reduce eventual demand for medical care?," *Social Science & Medicine*, 43(6), 999–1005.

Narang, U., V. Shankar, and S. Narayanan (2019): "The Impact of Mobile App Failures on Purchases in Online and Offline Channels," Discussion paper, Working Paper.

National Conference on State Legislature (2019): “Sexual Harassment Policies in State Legislatures,” *https://www.ncsl.org/research/about-state-legislatures/2018-legislative-sexual-harassment-legislation.aspx. Retrieved August 15, 2021.*

NBC News and Wall Street Journal (2017): “Study #17409, Question 22d,” .

Neyman, J. (1923): “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles.,” *Statistical Science*, Reprint, 5, 463–480.

——— (1959): “Optimal asymptotic tests of composite hypotheses,” *Probability and statsitics*, pp. 213–234.

Nguyen, D., and C. Rose (2011): “Language use as a reflection of socialization in online communities,” in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pp. 76–85.

Niemi, L., and L. Young (2016): “When and why we see victims as responsible: The impact of ideology on attitudes toward victims,” *Personality and social psychology bulletin*, 42(9), 1227–1242.

Northcutt Bohmert, M., K. Allison, and C. Ducate (2019): ““A rape was reported”: construction of crime in a university newspaper,” *Feminist Media Studies*, 19(6), 873–889.

Ornaghi, A., E. Ash, and D. L. Chen (2019): “Stereotypes in High-Stakes Decisions: Evidence from US Circuit Courts,” *Center for Law & Economics Working Paper Series*, 2.

Pagán, J. A., A. Puig, and B. J. Soldo (2007): “Health insurance coverage and the use of preventive services by Mexican adults,” *Health Economics*, 16(12), 1359–1369.

Palmer, J. E., E. R. Fissel, J. Hoxmeier, and E. Williams (2021): “# MeToo for whom? Sexual assault disclosures before and after# MeToo,” *American journal of criminal justice*, 46(1), 68–106.

Pavalanathan, U., X. Han, and J. Eisenstein (2018): “Mind Your POV: Convergence of Articles and Editors Towards Wikipedia’s Neutrality Norm,” *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–23.

Pearl, J. (2001): “Direct and indirect effects,” in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, San Francisco. Morgan Kaufman.

Petersen, M. L., S. E. Sinisi, and M. J. van der Laan (2006): “Estimation of Direct Causal Effects,” *Epidemiology*, 17, 276–284.

Pew Reasearch Center (2018): “American Trends Panel - Wave 35, May 29 – July 11, 2018,” *https://www.pewresearch.org/internet/dataset/american-trends-panel-wave-35/.*

Pinciotti, C. M., and H. K. Orcutt (2021): “Understanding gender differences in rape victim blaming: The power of social influence and just world beliefs,” *Journal of interpersonal violence*, 36(1-2), 255–275.

Press, R. (2014): “Insurance Coverage and Preventive Care Among Adults,” .

Pryzant, R., D. Card, D. Jurafsky, V. Veitch, and D. Sridhar (2020): “Causal effects of linguistic properties,” *arXiv preprint arXiv:2010.12919.*

Pusztová, L., and F. Babič (2020): “Performance Assessment of Different Classification Methods for Coupon Marketing in E-Commerce,” *Acta Electrotechnica et Informatica*, 20(3), 11–16.

Qiu, Y., X. Chen, and W. Shi (2020): “Impacts of Social and Economic Factors on the Transmission of Coronavirus Disease 2019 (COVID-19) in China,” *IZA Discussion Paper Series.*

Qu, Z., R. Xiong, J. Liu, and G. Imbens (2021): “Efficient Treatment Effect Estimation in Observational Studies under Heterogeneous Partial Interference,” *arXiv preprint arXiv:2107.12420.*

R Core Team (2020): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria.

——— (2022): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria.

Raju, J. S., S. K. Dhar, and D. G. Morrison (1994): “The effect of package coupons on brand choice,” *Marketing Science*, 13(2), 145–164.

Ramzan, B., I. S. Bajwa, N. Jamil, R. U. Amin, S. Ramzan, F. Mirza, and N. Sarwar (2019): “An intelligent data analysis for recommendation systems using machine learning,” *Scientific Programming*, 2019.

Rasmussen, S. R., J. L. Thomsen, J. Kilsmark, A. Hvenegaard, M. Engberg, T. Lauritzen, and J. Sogaard (2007): “Preventive health screenings and health consultations in primary care increase life expectancy without increasing costs,” *Scandinavian Journal of Public Health*, 35(4), 365–372.

Reich, C. M., G. A. Pegel, and A. B. Johnson (2021): “Are survivors of sexual assault blamed more than victims of other crimes?,” *Journal of interpersonal violence*, p. 08862605211037423.

Reimers, I., and C. Xie (2019): “Do coupons expand or cannibalize revenue? Evidence from an e-Market,” *Management Science*, 65(1), 286–300.

Ren, X., J. Cao, X. Xu, et al. (2021): “A two-stage model for forecasting consumers’ intention to purchase with e-coupons,” *Journal of Retailing and Consumer Services*, 59, 102289.

Rho, E. H. R., G. Mark, and M. Mazmanian (2018): “Fostering civil discourse online: Linguistic behavior in comments of# metoo articles across political perspectives,” *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 1–28.

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2020): “Adjusting for confounding with text matching,” *American Journal of Political Science*, 64(4), 887–903.

Robins, J., A. Rotnitzky, and L. Zhao (1995): “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data,” *Journal of American Statistical Association*, 90, 106–121.

Robins, J. M. (2003): “Semantics of causal DAG models and the identification of direct and indirect effects,” in *In Highly Structured Stochastic Systems*, ed. by P. Green, N. Hjort, and S. Richardson, pp. 70–81, Oxford. Oxford University Press.

Robins, J. M., and S. Greenland (1992): “Identifiability and Exchangeability for Direct and Indirect Effects,” *Epidemiology*, 3, 143–155.

Robins, J. M., and A. Rotnitzky (1995): “Semiparametric Efficiency in Multivariate Regression Models with Missing Data,” *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994): “Estimation of Regression Coefficients When Some Regressors Are not Always Observed,” *Journal of the American Statistical Association*, 90, 846–866.

Robinson, P. M. (1988): “Root-N-consistent semiparametric regression,” *Econometrica: Journal of the Econometric Society*, pp. 931–954.

Rosenbaum, P. R., and D. B. Rubin (1983): “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2), 212–218.

Rubin, D. B. (1974): “Estimating causal effects of treatments in randomized and nonrandomized studies.,” *Journal of educational Psychology*, 66(5), 688.

——— (1980): “Randomization analysis of experimental data: The Fisher randomization test comment,” *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B., and R. P. Waterman (2006): “Estimating the causal effects of marketing interventions using propensity score methodology,” *Statistical Science*, pp. 206–222.

Russell, K. J., and C. J. Hand (2017): “Rape myth acceptance, victim blame attribution and Just World Beliefs: A rapid evidence assessment,” *Aggression and Violent Behavior*, 37, 153–160.

Sacks, M., A. R. Ackerman, and A. Shlosberg (2018): “Rape myths in the media: A content analysis of local newspaper reporting in the United States,” *Deviant Behavior*, 39(9), 1237–1246.

Sallin, A. (2021): “Estimating returns to special education: combining machine learning and text analysis to address confounding,” *arXiv preprint arXiv:2110.08807*.

Santiago, C., and D. Criss (2017): “An activist, a little girl and the heartbreaking origin of ’Me too’,” *CNN, October 17, 2017*.

Schneider, L. J., L. T. Mori, P. L. Lambert, and A. O. Wong (2009): “The role of gender and ethnicity in perceptions of rape and its aftereffects,” *Sex Roles*, 60(5), 410–421.

Semenova, V., and V. Chernozhukov (2021): “Debiased machine learning of conditional average treatment effects and other causal functions,” *The Econometrics Journal*, 24(2), 264–289.

Simon, K., A. Soni, and J. Cawley (2017): “The impact of health insurance on preventive care and health behaviors: evidence from the first two years of the ACA Medicaid expansions,” *Journal of Policy Analysis and Management*, 36(2), 390–417.

Smith, A. N., S. Seiler, and I. Aggarwal (2021): “Optimal Price Targeting,” *Available at SSRN 3975957.*

Smith, O., and T. Skinner (2012): “Observing court responses to victims of rape and sexual assault,” *Feminist Criminology*, 7(4), 298–326.

Snow, J. (1856): “On the mode of communication of cholera,” *Edinburgh medical journal*, 1(7), 668.

Sobel, M. E. (2006): “What Do Randomized Studies of Housing Mobility Demonstrate?,” *Journal of the American Statistical Association*, 101, 1398–1407.

Sobolev, A. (2018): “How pro-government “trolls” influence online conversations in Russia,” *University of California Los Angeles.[7].*

Sommers, B. D., B. Maylone, R. J. Blendon, E. J. Orav, and A. M. Epstein (2017): “Three-year impacts of the Affordable Care Act: improved medical care and health among low-income adults,” *Health Affairs*, 36(6), 1119–1128.

Spencer, D., A. Dodge, R. Ricciardelli, and D. Ballucci (2018): ““I think it’s re-victimizing victims almost every time”: police perceptions of criminal justice responses to sexual violence,” *Critical criminology*, 26(2), 189–209.

Spiekermann, S., M. Rothensee, and M. Klafft (2011): “Street marketing: how proximity and context drive coupon redemption,” *Journal of Consumer Marketing.*

Stoye, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151, 70–81.

Strickland, B., M. Fisher, F. Keil, and J. Knobe (2014): "Syntax and intentionality: An automatic link between language and theory-of-mind," *Cognition*, 133(1), 249–261.

Suarez, E., and T. M. Gadalla (2010): "Stop blaming the victim: A meta-analysis on rape myths," *Journal of interpersonal violence*, 25(11), 2010–2035.

Sugiyama, M., M. Kawanabe, and P. L. Chui (2010): "Dimensionality Reduction for Density Ratio Estimation in High-dimensional Spaces," *Neural Networks*, 23(1), 44–59.

Suvarna, A., and G. Bhalla (2020): "# NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 328–335.

Suvarna, A., G. Bhalla, S. Kumar, and A. Bhardwaj (2020): "Identifying Victim Blaming Language in Discussions about Sexual Assaults on Twitter," in *International Conference on Social Media and Society*, pp. 156–163.

Sverdrup, E., A. Kanodia, Z. Zhou, S. Athey, and S. Wager (2020): "policytree: Policy learning via doubly robust empirical welfare maximization over trees," *Journal of Open Source Software*, 5(50), 2232.

Tan, C., L. Lee, and B. Pang (2014): "The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter," *arXiv preprint arXiv:1405.1438.*

Taub, A. (2019): "#MeToo Paradox: Movement Topples the Powerful, Not the Ordinary," *The New York Times, February 11, 2019.*

Tchetgen, E. J. T., and T. J. VanderWeele (2012): "On causal inference in the presence of interference," *Statistical methods in medical research*, 21(1), 55–75.

Tchetgen Tchetgen, E. J. (2013): "Inverse Odds Ratio-Weighted Estimation for Causal Mediation Analysis," *Statistics in Medicine*, 32, 4567–4580.

Tchetgen Tchetgen, E. J., and I. Shpitser (2012): "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis," *The Annals of Statistics*, 40, 1816–1845.

Temkin, J. (2000): "Prosecuting and defending rape: Perspectives from the bar," *Journal of Law and Society*, 27(2), 219–248.

Temkin, J., J. M. Gray, and J. Barrett (2018): "Different functions of rape myth use in court: Findings from a trial observation study," *Feminist Criminology*, 13(2), 205–226.

Ten Have, T. R., M. M. Joffe, K. G. Lynch, G. K. Brown, S. A. Maisto, and A. T. Beck (2007): "Causal mediation analyses with rank preserving models," *Biometrics*, 63, 926–934.

The Free Law Project (2020): *RECAP Archive. Accessed November, 2020, https://www.courtlistener.com/recap/.*

Tibshirani, R. (1996): "Regresson shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society*, 58, 267–288.

Time Magazine (2017): "Time Person of the Year 2017 - The Silence Breakers," *https://time.com/time-person-of-the-year-2017-silence-breakers/.*

Undavia, S., A. Meyers, and J. E. Ortega (2018): "A comparative study of classifying legal documents with neural networks," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 515–522. IEEE.

Uysal, A. K., and S. Gunal (2014): "The impact of preprocessing on text classification," *Information processing & management*, 50(1), 104–112.

van der Laan, M., and D. Rubin (2006): "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics*, 2, 1–38.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 6.

VanderWeele, T. (2013): "A three-way decomposition of a total effect into direct, indirect, and interactive effects," *Epidemiology*, 24, 224–232.

VanderWeele, T. J. (2009): "Marginal Structural Models for the Estimation of Direct and Indirect Effects," *Epidemiology*, 20, 18–26.

VanderWeele, T. J., and M. A. Hernan (2013): "Causal inference under multiple versions of treatment," *Journal of causal inference*, 1(1), 1–20.

Vansteelandt, S., M. Bekaert, and T. Lange (2012): "Imputation Strategies for the Estimation of Natural Direct and Indirect Effects," *Epidemiologic Methods*, 1, 129–158.

Veitch, V., D. Sridhar, and D. Blei (2020): "Adapting text embeddings for causal inference," in *Conference on Uncertainty in Artificial Intelligence*, pp. 919–928. PMLR.

Wager, S., and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wang, Z., and A. Culotta (2019): "When Do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception Using Individual Treatment Effect Estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7233–7240.

Weber, E. (2020): "Which measures flattened the curve in Germany?," *CEPR Covid Economics, Vetted and Real-Time Papers*, 24.

Wood-Doughty, Z., I. Shpitser, and M. Dredze (2018): "Challenges of using text classifiers for causal inference," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2018, p. 4586. NIH Public Access.

Xia, F., R. Chatterjee, and J. H. May (2019): "Using conditional restricted Boltzmann machines to model complex consumer shopping patterns," *Marketing Science*, 38(4), 711–727.

Xiao, F., L. Li, W. Xu, J. Zhao, X. Yang, J. Lang, and H. Wang (2021): "DMBGN: Deep Multi-Behavior Graph Networks for Voucher Redemption Rate Prediction," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3786–3794.

Xing, J., E. Zou, Z. Yin, Y. Wang, and Z. Li (2020): ""Quick Response" Economic Stimulus:

The Effect of Small-Value Digital Coupons on Spending," Discussion paper, National Bureau of Economic Research.

Yoon, S., J. S. Choe, Y. M. Han, and S. H. Kim (2020): "#MeToo Hits Online Dating, too: An Empirical Analysis of the Effect of the Me-Too Movement on Online Dating Users," .

Yörük, B. K. (2016): "Health insurance coverage and self-reported health: new estimates from the NLSY97," *International Journal of Health Economics and Management*, 16(3), 285–295.

Yu, B., S. Kaufmann, and D. Diermeier (2008): "Classifying party affiliation from political speech," *Journal of Information Technology & Politics*, 5(1), 33–48.

Zeileis, A. (2004): "Econometric computing with HC and HAC covariance matrix estimators," *Institut für Statistik und Mathematik, WU Vienna University of Economics and ....*

Zetterqvist, J., and A. Sjölander (2015): "Doubly Robust Estimation with the R Package drgee," *Epidemiologic Methods*, 4.

Zhang, D. J., H. Dai, L. Dong, F. Qi, N. Zhang, X. Liu, and Z. Liu (2017): "How does dynamic pricing affect customer behavior on retailing platforms? evidence from a large randomized experiment on alibaba," Discussion paper, Working paper, SSRN.

Zhang, M., and L. Luo (2021): "Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp," .

Zheng, D., Y. Chen, Z. Zhang, and H. Che (2021): "Retail price discount depth and perceived quality uncertainty," *Journal of Retailing*.

Zheng, W., and M. J. van der Laan (2012): "Targeted Maximum Likelihood Estimation of Natural Direct Effects," *The International Journal of Biostatistics*, 8, 1–40.