Comprehensive Causal Machine Learning

Michael Lechner and Jana Mareckova*

Swiss Institute for
Empirical Economic Research

University of St.Gallen

This version: February 2024

Date this version has been printed: ***28 February 2024***

Comments are very welcome.

Abstract: Uncovering the heterogeneity of causal effects at various levels of granularity provides substantial value to decision makers. Comprehensive approaches to causal effect estimation allow to use a single causal machine learning approach for estimation and inference of causal mean effects for all levels of granularity. Focussing on selection-on-observables, the paper provides the theoretical asymptotic guarantees for one such approach, the modified causal forest (*mcf*). It also compares the asymptotic and finite sample properties of the *mcf* to the generalized random forest (*grf*) and double machine learning (*dml*). The findings indicate that *dml*-based methods excel for average treatment effects at the population level (ATE) and group level (GATE) with few groups. However, for finer causal heterogeneity, explicitly outcome-centred forest-based approaches are superior. The *mcf* has three additional benefits: (i) It is the most robust estimator in cases when *dml*-based approaches underperform because of substantial selectivity; (ii) it is the best estimator for GATEs when the number of groups gets larger; and (iii), it is the only estimator that is internally consistent, in the sense that low-dimensional causal ATEs and GATEs are obtained as aggregates of finer-grained causal parameters.

Keywords: Causal machine learning, statistical learning, conditional average treatment effects, individualized treatment effects, multiple treatments, selection-on-observed-variables

JEL classification: C21, C87

Correspondence to: Michael Lechner or Jana Mareckova, Professors of Econometrics, Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Switzerland, Michael.Lechner@unisg.ch, Jana.Mareckova@unisg.ch, www.sew.unisg.ch.

* Michael Lechner is also affiliated with CEPR, London, CESIfo, Munich, IAB, Nuremberg, IZA, Bonn, and RWI, Essen. Results from two research projects that were part of the National Research Programmes "Big Data" (NRP 75, grant number 407540_166999) and "Digital Transformation" (NRP 77, grant number 407740_187301) of the Swiss National Science Foundation (SNSF) were the basis of this paper. Further information on these National Research Programmes can be found at www.nrp75.ch and www.nrp77.ch. The theoretical part on the *mcf* is a revised version of the theoretical part of our unpublished "Modified Causal Forest" paper. We thank GPT-4 for some limited assistance in editing.

1 Introduction

Machine learning (ML) has paved its way into academia and industry, impacting numerous fields from health care and finance to social sciences. At the core of the ML revolution is the ability of flexible model estimation and amazing predictive power of the methods. The growing debate around ML emphasizes that prediction does not imply causation. Going beyond mere predictive associations to identify cause-and-effect relationships is at the centre of most questions concerning the *effects* of policies, medical treatments, marketing campaigns, business decisions, etc. (see, e.g., Athey, 2017). The different focus of causal modelling called for ML approaches that can flexibly and reliably estimate causal effects, establishing causal machine learning (CML).

CML integrates principles of ML and causal inference. The causality literature provides conditions for identification and estimation of *effects* instead of associations. Building on these conditions, counterfactual problems can be transformed into specific prediction problems (see, e.g., Imbens and Wooldridge, 2009) for which ML methods are suitable (see, e.g., Hastie, Tibshirani, and Friedman, 2009 for an overview on ML methods). The flexibility of ML combined with careful use of data leads to reliable treatment effect estimators that are also able to uncover treatment effect heterogeneity (for an overview, see Athey and Imbens, 2017).

In recent years, researchers from different disciplines contributed to method development in CML. Although, such multidisciplinary work on a common subject is wonderful news from a scientific point of view, it raises the question which of these many proposed methods to use in any specific empirical study. In this paper, we focus on this question for a special case, which is popular in the CML literature: (i) when causal effects are plausibly identified under unconfoundedness (i.e., selection-on-observables), and (ii) the researcher is interested in aggregated average causal effects as well as their possibly fine-grained heterogeneity.

A key condition of any causal effect estimator to be attractive in applications is that it has acceptable statistical properties. Such theoretical guarantees and the implied ability to conduct inference are crucial for effect estimation, because, different to prediction settings, realisations from the true effects, the so-called 'ground truth' is unobservable and can thus not be used to evaluate the performance of the specific estimation. Another important factor is to keep the estimation of the many different causal parameters sufficiently simple, for example, by using the same, or the same type of causal machine learners for all of them. This avoids the time intensive tasks of tuning and monitoring many different models. We will call such methods Comprehensive Causal Machine Learners (CCML). Finally, it is advantageous to have internal consistency of the possibly many estimated effects, in the sense that effects at the higher aggregation levels are close to appropriately aggregated lower-level effects.

In the light of these three arguably desirable properties, we analyse three estimation principles that belong to the class of CCML and have the required theoretical guarantees. The methods are Debiased/Double Machine Learning (*dml*; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018), the Generalized Random Forest (*grf*; Athey, Tibshirani, and Wager, 2019), and the Modified Causal Forest (*mcf*; Lechner, 2018). Only the latter fulfils the third condition of internal consistency by construction.

The intended contribution of this paper to the literature is two-fold: The first contribution are the theoretical guarantees for the *mcf* that were missing so far. Additionally, the similarities and differences between the three approaches are documented. The second contribution is large scale-simulation exercise to evaluate the finite-sample performance of all three CCMLs in many different settings to better understand which estimator may have advantages or disadvantages in certain situations. Crucially, these simulations also cover the inference procedures that usually were absent from the small number of existing and far more limited simulation evidence.

The results yield several practical recommendations. Our findings indicate that *dml*-based methods predominantly excel in estimation of average treatment effects (ATE) or group treatment effects (GATE) with fewer groups. However, for finer causal heterogeneity, explicitly outcome-centred forest-based approaches are superior. Additionally, *mcf* offers three benefits: it is the most robust estimator even for the ATE in cases when *dml*-based approaches underperform because of substantial selectivity that needs to be corrected for. Second, it outperforms both *dml* and *grf* in GATE estimation when the number of groups gets larger. Finally, it is the only estimator that is internally consistent.

The following section surveys the contribution of CML methods to the estimation of ATE and more fine-grained conditional average treatment effects (CATEs) under unconfoundedness. Section 3 introduces the notation of the potential outcome models, defines the parameters of interest, and spells out the standard assumptions imposed under unconfoundedness. Section 4 discusses the three CCML approaches and provides theoretical guarantees for the *mcf*. The large-scale simulation study is summarized in Section 5. Section 6 concludes. Appendix A contains the proofs of the theoretical properties of the *mcf*. Implementational details of the simulation study are collected in Appendix B, while Appendix C holds its detailed results.

2 Literature

For causal inference under the unconfoundedness assumption (see Rubin, 1974), the average treatment effect (ATE) represents one of the main parameters of interest capturing the overall impact of a treatment or an intervention. For ATE estimators, such as inverse probability weighting (IPW) and propensity score matching (see, e.g., Imbens, 2004), propensity score estimation plays a central role. Both estimators are consistent when a parametric model for the propensity score is correctly specified. Consequently, one of the first applications of ML methods in ATE estimation involved flexible propensity score estimation. Various ML approaches, such as Neural Networks, Support Vector Machines, Decision Trees (Westreich, Lessler, and

Funk, 2010) or Boosting Models (McCaffrey, Ridgeway, and Morral, 2004; Westreich et al., 2010), have been proposed to improve the estimation of the propensity score.

However, even flexible ML methods may exhibit poor performance if they prioritize fitting the propensity score well instead of explicitly balancing the covariates across treatment groups, which is important to reduce the bias of the estimator for the causal parameter of interest. An estimator of the propensity score that explicitly aims to maximize covariate balancing was introduced in Imai and Ratkovic (2014). Alternatively, Graham, Pinto, and Egel (2012) improve IPW estimation by covariate balancing within the empirical likelihood framework, without explicit modelling of the propensity score.

Cannas and Arpino (2019) investigate the performance of matching and weighting estimators based on propensity scores obtained from Logistic Regression, Decision Tree, Bagging, Boosting, Random Forest, Neural Networks, and Naïve Bayes. Their simulation results indicate that Random Forests consistently outperform other methods, particularly in the context of IPW. Goller, Lechner, Moczall, and Wolff (2020) compare Random Forests, LASSO Logit Regression, and Probit estimation of the propensity score for a matching estimator in an active labour market policy setting, revealing that LASSO may yield more credible results than conventional propensity score estimation methods, but overall results were mixed.

Under the standard unconfoundedness assumptions, to be detailed in Section 3, ATE is alternatively identified as a difference between conditional expectations of the outcome variable for the treated and non-treated populations. This leads to regression-based estimation of the two conditional expectations. Assuming a correct specification of the outcome models, their difference yields a consistent ATE estimator. In the ML literature, tree-based methods such as BART (Hill, 2011) and Ensemble Methods (Austin, 2012) were introduced to obtain regression-based estimators of ATEs. However, issues like regularization bias (Athey, Imbens, and Wager, 2018) and slow convergence of ML methods render the outcome-based approach less popular.

Methods that combine propensity score-based and regression-based estimators to increase robustness to misspecification have gained popularity. Two related approaches prevail in the literature: Double-robust (DR) estimators and Neyman-orthogonal scores. DR estimators utilize two nuisance functions, i.e., the propensity score and outcome models. They are consistent as long as at least one of the nuisance functions is correctly specified. Certain DR estimators were shown to be semi-parametrically efficient if both nuisance functions are correctly specified. Ideas to combine propensity score and outcome modelling in a double-robust way originally emerged in Robins, Rotnitzky, and Zhao (1994, 1995), leading to the Augmented Inverse Probability Weighting (AIPW) estimator for the ATE. Nevertheless, sensitivity to extreme propensity scores, as in the case of the IPW estimator, remains an issue.[1]

Neyman-orthogonality ensures that the moment conditions identifying the target parameters are not locally affected by small perturbances in the nuisance parameter estimates (Chernozhukov et al., 2018). Such an orthogonality property is leveraged, e.g., in Belloni, Chernozhukov, and Hansen (2014) for ATE estimation with binary treatment and in Farrell (2015) for multiple treatments. As Neyman-orthogonal score mitigates first order bias arising from ML estimation of nuisance parameters (Bach, Chernozhukov, Kurz, and Spindler, 2024, p. 9), it emerged as a key element for the Double Machine Learning (*dml*) framework introduced in Chernozhukov et al. (2018). *dml* is a generic framework for obtaining consistent estimators and valid inference for low-dimensional parameters combining Neyman-orthogonal scores[2] with cross-fitting in high-dimensional settings under convergence rate conditions for the estimation of the nuisance parameters that are achievable by many ML methods. Chernozhukov et al. (2018) apply the *dml* framework to ATE estimation and derives its statistical properties.

[1] For potential benefits of trimming extreme propensity scores, see e.g., Huber, Lechner, and Wunsch (2013).

[2] Note that the DR score of Robins and Rotnitzky (1995) is Neyman-orthogonal.

Another approach for ATE estimation is Targeted Minimum Loss Estimation (*tmle*), a non-parametric method that combines ML methods with a targeted updating step to reduce bias and variance of the estimated parameter of interest (van der Laan and Gruber, 2012). *tmle* shares some similarities with *dml*. They both require estimation of nuisance parameters that are combined in a way that yields a double-robust and semi-parametrically efficient estimator of the ATE. Meanwhile *dml* utilizes the nuisance parameters in a Neyman-orthogonal score to mitigate the bias, *tmle* combines them in a two-stage estimation procedure addressing the bias and yielding quantities for each observation that can be averaged into an ATE.

ML methods extend beyond ATE estimation and help uncover heterogeneity of the treatment effects through analysis of Conditional Average Treatment Effects (CATEs). CATEs quantify how the treatment affects an individual unit with specific characteristics. Prominent methods include Causal Forests (Wager and Athey, 2018; Athey et al., 2019; Lechner, 2018), along with *dml*-based estimators yielding estimated components of the DR efficient score that are further regressed on the covariates as first proposed by van der Laan (2006). Alternative approaches include Meta-learners, which use multiple machine learning algorithms as "base learners" to estimate large-dimensional CATEs. The X-learner, introduced by Künzel, Sekhon, Bickel, and Yu (2019), is effective in cases with uneven treatment groups, while the R-learner, developed by Nie and Wager (2021), leverages the Robinson (1988) transformation to estimate heterogeneous causal effects. For the *tmle*, the above-mentioned quantities of its two-stage procedure can be also averaged into a CATE using the corresponding part of the sample. In case of a CATE for the finest granularity level, the *tmle* quantity depends on an observed outcome and cannot serve for a prediction on a new sample containing covariates only.

The influential paper by Chernozhukov et al. (2018), averaging the estimated components of individual DR efficient scores of Robins and Rotnitzky (1995) to estimate ATE, spanned further methodological contributions in *dml*, particularly in discovering heterogeneity along a chosen set of covariates, such as gender or age group. *dml* estimation of CATEs exploits the fact that the conditional expectation of the DR efficient score given such covariates identifies the respective CATEs. For low-dimensional sets of chosen covariates, OLS, Series, or Kernel Regressions of the estimated DR score components on the chosen set of covariates estimate the low-dimensional CATE and standard statistical inference applies, as shown in Semenova and Chernozhukov (2021), Zimmert and Lechner (2019), and Fan, Hsu, Lieli, and Zhang (2022). For a higher dimensionality of the set of the chosen covariates, Kennedy (2023) introduces the DR-learner that takes the components of a classic DR estimator of ATE and regresses them on all covariates to estimate a large-dimensional CATE and derives an upper bound on the DR-learner error relative to an oracle.

The foundations for the heterogeneous treatment effect estimation by Causal Forests were laid in a seminal paper by Athey and Imbens (2016) introducing splitting rules for treatment effect estimation, as well as honest splitting to ensure valid inference for CATEs through tree-based methods. Subsequently, Wager and Athey (2018) develop a nonparametric Causal Forest algorithm based on honest trees and splitting rules maximizing heterogeneity in treatment effects across final leaves. The last step of the estimator of the possibly large-dimensional CATE is obtained by averaging individual tree estimates. The Causal Forest is consistent and asymptotically normal. In a distinct but related approach, Athey et al. (2019) introduce the Generalized Random Forests (*grf*). It is based on an honest forest with a gradient-based approximation of an estimator-specific splitting rule. The forest estimation determines importance weights of each observation for the parameters of interest. In turn, these weights determine a neighbourhood of observations that will contribute to the CATE estimation, instead of estimating the CATE directly as an average across honest trees as in Wager and Athey (2018). The forest weights then enter a set of local moment conditions identifying the CATE. This estimator is consistent and asymptotically Gaussian and generalizes beyond unconfoundedness. Additionally, Lechner (2018) introduces a Modified Causal Forest (*mcf*) estimation procedure for ATEs and CATEs in a multiple treatment framework. He proposes a weights-based inference procedure utilizing the weighted-outcome representation of forest estimates. *mcf* further differs from the other Causal Forests in using the mean squared error of the CATE directly to find the best split. Furthermore, it introduces a two-sample honesty for building the forest and the estimation of the effects. A simulation study of Lechner (2018) shows superior performance of *mcf* to the Causal Forest of Wager and Athey (2018). Bodory, Busshoff, and Lechner (2022) show that the *mcf* works in applications replicating results from several papers in different fields.

The above-mentioned advances provide practitioners with several methods to estimate treatment effects across different levels of granularity. Table 1 provides an overview of commonly used ML approaches for estimation of such treatment effects and evaluates them based on (i) available statistical inference, (ii) internal consistency of estimates (i.e., higher level aggregates are close to the aggregation of lower-level effects) and (iii) their ability to predict large-dimensional CATEs on new data (with covariate information only). The overview reveals that *tmle* and Meta-learners do not have available statistical inference for all levels and that *tmle* cannot predict CATEs at the finest granularity level on a new data set that contains covariates only. Thus, these two methods are not considered to be CCMLs. On the other hand, *dml* and Causal Forests are approaches that fulfil the conditions for comprehensive treatment effect evaluation with available statistical inference and predictions for new data containing covariates only. For Causal Forests, there is a Bayesian alternative which will not be pursued further because its inference procedure is not based on repeated sampling inference.

Focusing on estimation at different levels of granularity via CCML*, dml* operates via the estimated components of the DR efficient scores that are regressed on covariates to yield lower and higher-level aggregates of average treatment effects. Causal Forests, *grf* and *mcf,* provide in their first stage CATE estimates at the highest level of granularity. Using direct aggregation step of these CATEs, *mcf* leverages the weighted-outcome representation of forest predictions, while *grf* uses a variant of the AIPW score to yield higher level aggregations via a regression step.

Table 1: Properties of commonly used ML methods for treatment effect estimation

Approach	**Parameters**			**Statistical inference**	**Internal consistency**	**CATE prediction by covariates**
	Dimension of CATE		*ATE*			
	large	small				
dml	Regr	Regr	Regr	Yes	No	Yes
grf	LocGMM	Regr	Regr	Yes	No	Yes
mcf	MCF	Aggr	Aggr	Yes	Yes	Yes
Meta-learners	ML	Aggr *	Aggr *	No**	Yes	Yes
tmle	TMLE	TMLE	TMLE	Low-dim. (C)ATE	Yes	No
BART	BART	BayAggr	BayAggr	Bayesian	Yes	Yes

Note: Regr: Regression-based estimation; Aggr: Aggregation of estimated IATEs; LocGMM: Local General Method of Moments, MCF: Modified Causal Forest; BART: Bayesian Additive Regression Trees; BayAggr: Bayesian Aggregation; TMLE: Targeted Minimum Loss Estimation.
* Averaging of the finest large-dimensional CATEs for Meta-learners was implemented e.g., in Salditt, Eckes, and Nestler (2023).
** However, see recent developments in constructing valid confidence intervals for Meta-learners based on conformal prediction in Alaa, Ahmad, and van der Laan (2024).

The literature on large scale simulation comparisons of finite sample properties of causal CML methods is very limited. Among notable contributions, Knaus, Lechner, and Strittmatter (2021) look at the finite-sample performance of selected machine learning estimators, covering *grf* and generic approaches that can be combined with any ML method, yielding large-dimensional CATE estimates based on pseudo-outcomes or modified covariates. Higher level aggregates are simple averages of the corresponding CATEs. Using the Empirical Monte Carlo Study (EMCS) approach, many components of their simulations are based on real data. The results show that best-performing estimators at the large-dimensional CATE level produced most reliable estimates at higher aggregation levels in terms of MSE. In general, methods that utilized both outcome and treatment equations are among the best-performing methods, including *grf* with local centering.[3] Caron, Baio, and Manolopulou (2022) evaluate several Meta-learners in an EMCS based on health data. Their results highlight variability in performance that depends on the complexity of the data generating process. In a setting with a complex CATE, multitask

[3] At the time the simulations of Knaus, Lechner, and Strittmatter (2021) were performed, the *mcf* was not yet available.

learners, designed to estimate both outcome equations jointly such as Multitask Gaussian Process introduced in Alaa and Van Der Schaar (2017), perform best in terms of MSE. The X-learner performs the best in a setting with a simple CATE function and slightly unbalanced treatment groups. Another health-data-based EMCS of Wendling, Jung, Callahan, Schuler, Shah, and Gallego (2018) concentrates on large-dimensional CATE estimation in a common healthcare setting when outcomes are binary and rare. The compared methods include BART, *grf* with local centering, Causal Boosting, and Causal Multivariate Adaptive Regression Splines (MARS). The findings show that BART and Causal Boosting perform better across the scenarios, as evaluated by the root MSE in the conditional probability (risk) difference estimates. Except for one scenario involving highly heterogeneous treatment effects, the coverage rate for BART, *grf*, and causal MARS was close to its nominal level.[4]

Empirical applications using ML for treatment estimation span across many fields and topics. The following non-exhaustive list illustrates the variety of applications ranging from active labour market policy (Davies and Heller, 2017; Bertrand, Crépon, Marguerie, and Premand, 2021; Pytka and Gulyas, 2021; Knaus, Lechner, and Strittmatter, 2022; Cockx, Lechner, and Bollens, 2023), education (Knaus, 2021; Farbmacher, Kögler, and Spindler, 2021), social experiments (Athey and Wager, 2019; Strittmatter, 2023), energy use (Knittel and Stolper, 2021), natural resource rents (Hodler, Lechner, and Raschky, 2023), and the dating market (Boller, Lechner, and Okasa, 2021) to medicine (Langenberger, Steinbeck, Schöner, Busse, Pross, and Kuklinski, 2023). For the comprehensive nature of the *dml* framework, Knaus (2022) demonstrates its flexibility in an active labour market programme evaluation setting.

[4] To the best of our knowledge, this is the only study that also investigates the performance of the corresponding inference procedures. The simulation study in Wendling et al. (2018) differs from the simulation in Section 5 by using fixed covariates across replications. Despite this difference, their results are in line with the results in Section 5 confirming that high treatment heterogeneity negatively affects the coverage probability. Meanwhile DGPs with low treatment heterogeneity have coverage probability closer to nominal rates.

3 Causal framework

3.1 The potential outcome model

We use Rubin's (1974) potential outcome language to describe a multiple treatment model under unconfoundedness, selection-on-observables, or conditional independence (Imbens, 2000, Lechner, 2001). Let *D* denote the treatment that may take a known number of *M* different integer values from *0* to *M-1*. The (potential) outcome of interest that realises under treatment *d* is denoted by Y^d. For each observation, we observe only the particular potential outcome that is related to the treatment status the observation is observed to be in, $y_i = \sum_{d=0}^{M-1} \underline{1}(d_i = d) y_i^d$, where $\underline{1}(\cdot)$ denotes the indicator function, which equals one if its argument is true.[5] There are two groups of variables to condition on, $\tilde{X}$ and *Z*. $\tilde{X}$ contains those covariates needed to correct for selection bias (confounders), while *Z* contains variables that define (groups of) population members for which an average causal effect estimate is desired.[6] $\tilde{X}$ and *Z* may be discrete, continuous, or both. They may overlap in any way. Denote the union of the two groups of variables by *X*, $X = (\tilde{X}, Z), \dim(X) = p$.[7]

Below, we investigate the following average causal effects:

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

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

$$ATE(m,l;\Delta) = E(Y^m - Y^l \mid D \in \Delta) = \int IATE(m,l;x,\Delta) f_{X|D\in\Delta}(x) dx \, .$$

---

5 If not obvious otherwise, capital letters denote random variables, and small letters their values. Small values subscripted by '*i*' denote the value of the respective variable of observation '*i*'.

6 In the rest of the paper, we call these two sets of variables 'features' (common in the ML literature) and 'covariates' (common in the econometrics literature) interchangeably, reflecting the integration of ML and causal inference into CML.

7 To avoid complications, we assume *p* to be finite (although it may be very large).

The **I**ndividualized **A**verage **T**reatment **E**ffects (IATEs) measure the mean impact of treatment *m* compared to treatment *l* for units with features *x* that belong to treatment groups Δ, where Δ denotes all treatments of interest.[8] The IATEs represent the causal parameters at the finest aggregation level of the features available. On the other extreme, the **A**verage **T**reatment **E**ffects (ATEs) represent the population averages. If Δ relates to the population with $D=m$, then this is the **A**verage **T**reatment **E**ffect on the **T**reated (ATET) for treatment *m*. However, for example, it might also relate to a combination of two or more treatment populations. ATE and ATET are the classical parameters investigated in many causal studies. The **G**roup **A**verage **T**reatment **E**ffect (GATE) parameters are in-between those two extremes with respect to their aggregation levels.[9] IATEs and GATEs are special cases of the already mentioned **C**onditional **A**verage **T**reatment **E**ffects (CATEs).

3.2 Identifying assumptions

The following set of assumptions identifies the causal effects discussed in the previous section (see Imbens, 2000, Lechner 2001):[10]

$$\{Y^0,...,Y^m,...,Y^{M-1}\} \amalg D \mid X = x, \qquad \forall x \in \chi; \qquad (CIA)$$

$$0 < P(D = d \mid X = x) = p_d(x), \qquad \forall x \in \chi, \forall d \in \{0,...,M-1\}; \quad (CS)$$

$$Y = \sum_{d=0}^{M-1} \underline{1}(D = d) Y^d; \qquad (Observation\ rule)$$

The conditional independence assumption (CIA) implies that there are no features other than *X* that jointly influence treatment and potential outcomes (for the values of *X* that are in

---

8 Under the identifying assumption imposed in the next subsection, *IATE(m, l, x, Δ)* is the same for all treatment groups and does therefore not depend on *Δ*.

9 We presume that the analyst selects the variables *Z* prior to estimation. However, the estimated IATEs may be analysed by methods picking *Z* in a data-driven way to describe their dependence on certain features. See Section 6 in Lechner (2018) for more details. Note that Abrevaya, Hsu, and Lieli (2015) and Lee, Okui, and Whang (2017) introduce similar aggregated parameters that depend on a reduced conditioning set and discuss inference in their specific settings.

10 To simplify the notation, we take the strongest form of these assumptions. Some parameters are identified under weaker conditions as well (for details, Imbens, 2000, 2004, or Lechner, 2001).

the support of interest, χ). The common support (CS) assumption stipulates that for each value in χ , there must be the possibility to observe all treatments. The stable-unit-treatment-value assumption (SUTVA, observation rule, consistency condition) implies that the observed value of the treatment and the outcome does not depend on the treatment allocation of the other population members (ruling out spillover and treatment scale effects). Usually, to have an interesting interpretation of the effects, it is required that *X* is not influenced by the treatment (exogeneity). In addition to these *identifying* assumptions, assume that a large random sample of size *N* from the random variables *Y, D, X,* (y_i, x_i, d_i), *i=1,…,N*, is available and that all necessary moments of these random variables exist.[11]

If these assumptions hold, then all IATEs are identified in the sense that they can be uniquely deduced from expectations of variables that have observable sample realisations (see Hurwicz, 1950):

$$\begin{aligned} IATE(m,l;x,\Delta) &= E(Y^m - Y^l \mid X = x, D \in \Delta) \\ &= E(Y^m - Y^l \mid X = x) \\ &= E(Y^m \mid X = x, D = m) - E(Y^l \mid X = x, D = l) \\ &= E(Y \mid X = x, D = m) - E(Y \mid X = x, D = l) \\ &= IATE(m,l;x); \qquad\qquad \forall x \in \chi, \forall m \neq l \in \{0,...,M-1\}. \end{aligned}$$

Note that the identified IATE does not depend on the conditioning treatment set, Δ. Since the distributions used for aggregation, $f_{\tilde{X}|Z=z,D\in\Delta}(\tilde{x})$ and $f_{X|D\in\Delta}(x)$, relate to observable variables (*X*, *D*) only, they are identified as well (under standard regularity conditions). This in turn implies that the GATE and ATE parameters are identified (their dependence on Δ remains if the distribution of the features depends on Δ).

[11] The identification results will also hold under weights-based and dependent sampling (if the dependence is not too large and certain additional regularity conditions are imposed), but for simplicity we stick to the i.i.d. case. Higher moments are not needed for identification but may be required for some theoretical guarantees.

4 Comprehensive approaches for estimation and inference

In this section, three comprehensive approaches for treatment effect estimation are presented. In the first two subsections, the underlying principles as well as the concrete estimation and inference algorithms for Double Machine Learning (*dml)* and Generalized Random Forest (*grf)* are reviewed. The third subsection focuses on the Modified Causal Forest (*mcf)*. It explains the idea behind this estimator for the IATE as well as the aggregation steps used to estimate GATE and ATE. While the basic principles of the *mcf* are introduced in Lechner (2018), that paper does not contain explicit theoretical guarantees. Therefore, here we are proving consistency and asymptotic Gaussianity for these parameters. This is followed by an implementation of weights-based approximate inference as a computationally convenient tool to conduct inference for all desired aggregation levels.

4.1 Double Machine Learning

dml estimation (Chernozhukov et al., 2018) is based on moment conditions with scores satisfying the identification condition as well as Neyman-orthogonality, yielding estimators that are robust to small estimation errors in nuisance parameters (as in Neyman, 1959). A Neyman-orthogonal score that identifies ATE, GATE, and IATE for treatments m and l under assumptions outlined in Subsection 3.2 is the DR score of Robins and Rotnitzky (1995):

$$\psi_{m,l}^{dml}(O;\theta_{m,l},\eta_{m,l}) = \Gamma_{m,l}^{dml}(O;\eta_{m,l}) - \theta_{m,l}$$

$$\Gamma_{m,l}^{dml}(O;\eta_{m,l}) = \mu_m(X) - \mu_l(X) + \frac{\underline{1}(D=m)(Y-\mu_m(X))}{p_m(X)} - \frac{\underline{1}(D=l)(Y-\mu_l(X))}{p_l(X)},$$

where $\mu_d(x) = E\left(Y \middle| D=d, X=x\right)$, $\eta_{m,l}(X) = (\mu_m(X), \mu_l(X), p_m(X), p_l(X))$ captures the nuisance parameters, θ represents the treatment effect of interest, and O represents all observable variables, i.e., $O = (X, Y, D)$. Noting that the difference of the third and fourth term in this score has expectation zero conditional on *X*, and letting η^0 and θ^0 denote the true value of η and θ respectively, $\psi^{dml}_{m,l}(O;\theta_{m,l},\eta_{m,l})$ identifies the different treatment effects:

$$E\left(\psi^{dml}_{m,l}(O;\theta^0_{m,l},\eta^0_{m,l})\right)=0, \qquad \theta^0_{m,l}=ATE(m,l),$$

$$E\left(\psi^{dml}_{m,l}(O;\theta^0_{m,l}(z),\eta^0_{m,l})\big|Z=z\right)=0, \quad \theta^0_{m,l}(z)=GATE(m,l;z),$$

$$E\left(\psi^{dml}_{m,l}(O;\theta^0_{m,l}(x),\eta^0_{m,l})\big|X=x\right)=0, \quad \theta^0_{m,l}(x)=IATE(m,l;x), \text{ for all } x\in\chi .$$

Effectively, *ATE(m,l)* is estimated as an average of the estimated component of the DR score, $\hat{\Gamma}^{dml}_{m,l}\left(o_i;\hat{\eta}^{-k(i)}_{m,l}(x_i)\right)$, across all observations, i.e.,

$$\widehat{ATE}^{dml}(m,l)=\hat{\theta}^{dml}_{m,l}=\frac{1}{N}\sum_{i=1}^{N}\hat{\Gamma}^{dml}_{m,l}\left(o_i;\hat{\eta}^{-k(i)}_{m,l}(x_i)\right).$$

This process involves the estimation of nuisance parameters, symbolized as $\hat{\eta}^{-k(i)}_{m,l}(X)$, through *K*-fold cross-fitting. This method ensures that for each observation, the corresponding nuisance parameters are estimated without using that specific observation in their training data. Particularly, the elements of the vector $\hat{\eta}^{-k(i)}_{m,l}(X)$ are estimated on *K-1* folds not containing the observation *i,* which resides in the left-out fold *k*. This is indicated by the superscript *-k(i)*.

Neyman-orthogonal scores mitigate regularization bias in the estimation of the ATE for multiple treatments, provided that the product of the estimation errors for the two nuisance parameters within each treatment group diminishes at about a $\sqrt{N}$- rate. This condition, which many ML methods achieve, allows for flexible estimation of nuisance parameters leading to ATE estimates robust to small estimation errors in nuisance parameters. The additional combination with cross-fitting is important to avoid overfitting when estimating the nuisance parameters and to subsequently guarantee $\sqrt{N}$-consistent estimation of the main parameters of interest. For technical details, refer to Chernozhukov et al. (2018). The variance estimator of the *dml* can be computed as the sample variance of the estimated DR scores following Theorem 3.2 in Chernozhukov et al. (2018), and Theorem 1, equation (33), and Remark 1 in Bach et al. (2024). As shown in Theorem 5.1 in Chernozhukov et al. (2018), the estimator reaches the semiparametric efficiency bound of Hahn (1998).

Building on the identification result, *GATE(m,l;z)* can be estimated by regressing $\hat{\Gamma}_{m,l}^{dml}\left(o_i;\hat{\eta}_{m,l}^{-k(i)}(x_i)\right)$ on a low-dimensional vector Z of pre-specified variables. Semenova and Chernozhukov (2021) provide statistical inference for the best linear predictor in this case, a method that is also implemented in Section 5. While for continuous regressors, inference targets the function, for group indicators it targets the parameters. Standard errors are estimated via heteroscedasticity-robust standard errors for pointwise and uniform confidence bands and their asymptotic validity is proven. For continuous regressors, kernel regression offers a viable alternative, enabling the estimation of *GATE(m,l;z)* and facilitating statistical inference, as demonstrated in Zimmert and Lechner (2019) and Fan et al. (2022).

In a similar fashion, a natural approach to obtain an estimator of *IATE(m,l;x)* is by regressing $\hat{\Gamma}_{m,l}^{dml}\left(o_i;\hat{\eta}_{m,l}^{-k(i)}(x_i)\right)$ on the features. Foster and Syrgkanis (2023) and Kennedy (2023) derive error bounds for this two-step procedure. Particularly, the findings in Kennedy (2023) provide a theoretical foundation for the validity of inferential methods used in this context. In practical applications, different techniques have been employed: Goller (2023) utilizes linear regression, while Knaus (2022) opts for a Random Forest approach.

4.2 Generalized Random Forests

grf (Athey et al., 2019) extends the Random Forests into a nonparametric method estimating parameters identifiable by local moment conditions:

$$E\left(\psi^{grf}(O;\theta^0(x),\eta^0(x))\middle|X=x\right)=0 \text{ for all } x\in\chi\,,$$

where $\psi^{grf}(\cdot)$ is a score function identifying the true values of $\theta(x)$. As before, $\eta(x)$ captures nuisance parameters, $\theta^0(x)$ and $\eta^0(x)$ are the true values of $\theta(x)$ and $\eta(x)$, and O represents all observable variables. This extension notably includes the estimation of heterogeneous treatment effects based on $\psi^{grf}(Y,\tilde{D};\theta(x),\eta(x)) = \left(Y - \tilde{D}'\theta(x) - \eta(x)\right)(1 \quad \tilde{D}')'$, where $\tilde{D}$ is an (M-1) x 1 vector containing dummy variables indicating whether treatment $d \in \{1,...,M-1\}$ was received, $\theta(x)$ is a parameter vector representing corresponding treatment effects (treatment *d* vs 0), and $\eta(x)$ captures the intercept and all confounders in a single nuisance parameter.[12] Note that the score function in *dml* is based on a fully nonparametric DR score identifying treatment effects for all pairwise combinations of treatments, while the score function in *grf* stems from a partial linear model in which only treatment effects with respect to a reference treatment can be identified. The reference treatment is the one left out from vector $\tilde{D}$, here the control group with *d = 0*. This is not too restrictive if effects that are conditioned on the treatment status are not of interest (as for IATE estimation), because the point estimates of the other treatment combinations can be obtained as $\theta_{m,l}(x) = \theta_{m,0}(x) - \theta_{l,0}(x)$.

Building on the *local* generalized method of moments, *grf* estimates IATEs, $\theta(x)$, as

$$\left(\hat{\theta}^{grf}(x), \hat{\eta}(x)\right) \in \underset{\theta(x),\eta(x)}{\arg\min} \left\| \sum_{i=1}^{N} w_i^{grf}(x)\psi(o_i;\theta(x),\eta(x)) \right\|_2 ,$$

where the weights, $w_i^{grf}(x)$, are obtained from an honest Random Forest whose gradient-based splitting rule is designed to maximize treatment heterogeneity in the daughter leaves. Honesty involves a data partitioning strategy to prevent overfitting in the context of Random Forests. The subsample drawn for each tree is further split into two halves: one for building the tree structure and the other one for estimating the parameters of interest – in this case, the tree

[12] This part is coined as an *intercept* term *c(x)* in Athey et al. (2019).

weights which are aggregated into forest weights. As all observations alter between the two halves across trees when subsampled, we call this procedure 'one-sample honesty'. The forest weights, summing up to 1, can be seen as measure of relevance of the observation *i* for the estimation of the local parameters $\theta(x)$. Given forest weights $w_i^{grf}(x)$, the solution to the given optimization problem is:

$$\hat{\theta}^{grf}(x) = \left(\sum_{i=1}^{N} w_i^{grf}(x) \left(\tilde{d}_i - \bar{\tilde{d}}_w \right) \left(\tilde{d}_i - \bar{\tilde{d}}_w \right)' \right)^{-1} \left(\sum_{i=1}^{N} w_i^{grf}(x) \left(\tilde{d}_i - \bar{\tilde{d}}_w \right) \left(y_i - \bar{y}_w \right) \right),$$

where $\bar{\tilde{d}}_w = \sum_{i=1}^{N} w_i^{grf}(x) \tilde{d}_i$ and $\bar{y}_w = \sum_{i=1}^{N} w_i^{grf}(x) y_i$ (compare with eq. (19) in Athey et al., 2019)).

As the splits are chosen to maximize treatment heterogeneity across daughter leaves, to mitigate confounding effects at early splits, Athey et al. (2019) recommend to partial out effect of the covariates *X* on the outcome *Y* and treatment assignment *D* and built the forest using locally centred values $y_i^{cent} = y_i - \hat{\mu}^{(-i)}(x_i)$ and $\tilde{d}_i^{cent} = \tilde{d}_i - \hat{p}_d^{(-i)}(x_i)$ for all observations, where $\mu(x) = E\left(Y \middle| X = x\right)$ and the superscript *(-i)* denotes that the prediction was done excluding observation *i*. The best practice is to estimate $\mu(x)$ and $p_d(x)$, e.g., by *K*-fold cross-fitting as mentioned in Athey et al. (2019). The available *grf* package performs local centring by subtracting out-of-bag predictions[13] obtained from regression forests specifically trained to predict the outcome variable *Y* and the probabilities of treatment assignment, given the covariates *X*. Despite its computational attractiveness, this implementation violates the requirement of complete separation of training and prediction data consistent with the theoretical results. Also, the simulation results in Section 5 reveal the necessity of *K*-fold cross-fitting over out-of-bag predictions for local centring to remove bias.

[13] In Athey et al. (2019) out-of-bag predictions are called leave-one-out predictions as the final prediction for observation *i* averages predictions that are obtained from trees that did not use observation *i* for splitting.

The asymptotic results for $\hat{\theta}^{grf}(x)$ are based on their linear approximation motivated by the method of influence functions. Athey et al. (2019) prove that the two are coupled and asymptotic properties of one apply to the other. Since the linear approximation can be interpreted as an estimate of an honest Causal Forest introduced in Wager and Athey (2018), their asymptotic result can be applied to establish asymptotic Gaussianity of $\hat{\theta}^{grf}(x)$, given that $\theta(x)$ and $\eta(x)$ are consistently estimated.[14] The linear approximation is also leveraged to derive an estimator of the point-wise standard errors for $\hat{\theta}^{grf}(x)$. Derivation of the variance of the linear approximation yields two terms. One can be consistently estimated by means of regressions. The other can be seen as an estimate of a regression forest with weights $w_i^{grf}(x)$ and the score function as outcome variable. Athey et al. (2019) propose to estimate this term by a variant of a so-called bootstrap of little bags. This method was introduced in Sexton and Laake (2009). Athey et al. (2019) motivate it by the observation that "*building confidence intervals via half-sampling – whereby evaluating an estimator on random halves of the training data to estimate its sampling error – is closely related to the bootstrap (Efron, 1982)*". A computationally efficient implementation to estimate within one forest, both, the forest weights and the element of the variance by half-sampling requires drawing for each small bag of trees a random half of the sample that is available for further subsampling. Honesty additionally requires the trees to be built on half of the subsample, and the weights $w_i^{grf}(x)$ to be computed on the half that was not used to build the tree. Thus, each $w_i^{grf}(x)$ is estimated on approximately a quarter of all the trees in the forest when the subsampling rate is close to 1.[15] Consistent estimation of the two variance terms yields asymptotically valid confidence intervals for the IATE at point *x*.

[14] This can be achieved, e.g., via honest forests as outlined in Athey et. al (2019) in Theorem 3.

[15] Figure A.1 in Appendix A.2 captures the *grf* procedure graphically.

Similarly to *dml*, *grf* offers comprehensive estimation of treatment effects at all levels of granularity. However, unlike *dml,* which estimates all effects in two steps by estimating the $\hat{\Gamma}^{dml}_{m,l}\left(o_i;\hat{\eta}^{-k(i)}(x_i)\right)$ first, and subsequently regressing them on features to obtain ATE, GATE and IATE estimates, *grf* takes a more direct approach. Specifically, *grf* estimates IATEs "directly", while ATE and GATEs are estimated through additional regressions. The *grf* package implements ATE estimation (treatment *m* vs no treatment) by a variant of the AIPW estimator, plugging in the estimates from IATE estimation for the nuisance parameters:[16]

$$\widehat{ATE^{grf}}(m,0)=\hat{\theta}^{grf}_{m,0}=\frac{1}{N}\sum_{i=1}^{N}\hat{\Gamma}^{grf}_{m,0}(o_i;\hat{\eta}),$$

$$\begin{aligned}\hat{\Gamma}^{grf}_{m,0}(o_i;\hat{\eta})&=\hat{\theta}_{m,0}(x_i)+\frac{1}{\hat{p}_d(x_i)}\underline{1}(d_i=m)(y_i-\hat{\mu}_m(x_i))\\&=\hat{\theta}_{m,0}(x_i)+\frac{1}{\hat{p}_d(x_i)}\underline{1}(d_i=m)\left(y_i-\left(\hat{\mu}(x_i)-\sum_{d>0,d\neq m}\hat{p}_d(x_i)\hat{\theta}_{d,0}(x_i)\right)\right).\end{aligned}$$

This is effectively a *dml* estimator in which the nuisance parameters were estimated by Random Forests, and by a weighted moment condition with forest-based weights. Nuisance parameters should be estimated such that the observation *i* does not influence the training phase. This is guaranteed by honesty of the Causal Forest estimating the forest weights and by using *K*-fold cross-fitting to estimate $\left(\mu(x),p_0(x),...,p_{M-1}(x)\right)$ in the local centring step. To obtain GATE estimates, $\hat{\Gamma}^{grf}_{m,0}(o_i;\hat{\eta})$ are regressed on *Z*, as in *dml*. Inference for ATE and GATE is also conducted as in *dml*.

[16] The equations for estimating the ATE are reconstructed from the GRF R package (version 2.3.1). The derivation involves code implemented in the functions "get_scores.R" and "average_treatment_effect.R".

4.3 Modified Causal Forest

4.3.1 Motivation

Lechner (2018) builds on Wager and Athey (2018) and introduces the Modified Causal Forest estimator, *mcf*. One of the procedures proposed by Wager and Athey (2018) builds trees by taking the outcome Y as dependent variable and finding splits that maximize treatment effect heterogeneity in the daughter leaves under the restriction that each treatment group needs to have at least a certain number of observations. The tree estimator of an IATE at point x is then obtained from the final leaf containing x by differencing the average outcomes of the treatment groups evaluated on the subsample that was not used to build the tree. The forest estimate of an IATE at point x is calculated as average of all tree estimates. Wager and Athey (2018) acknowledge that this procedure works well in an experimental design with heterogeneous treatment effects but may do poorly in the presence of confounding. This observation led to further extensions addressing the confounding issue.

Formally, *grf* introduces local centring inside its Random Forest to remove the confounding effects by a transformation of the data when searching for splits that maximize effect heterogeneity. In contrast, *mcf* takes a different approach. Motivated by the fact that in the presence of selection bias the difference of the outcome means of treated and controls within all leaves will not correspond to the means of the true effects, *mcf* proposes a splitting criterion targeting the minimisation of the MSE of IATE directly. In addition, a penalty is added penalizing squared differences between propensity scores. Taken together, the chosen split is predictive for both Y and the conditional treatment assignment probabilities.

As a motivation for the splitting rule for the *mcf*, remember that the IATE is also identified as the difference of two outcome regressions in the different treatment pools:

$$IATE(m,l;x) = \mu_m(x) - \mu_l(x)\,; \qquad \forall x \in \chi, \forall m \neq l \in \{0,\ldots,M-1\}.$$

This estimation task is different from standard ML problems because the two conditional expectations must be estimated in different, treatment-specific subsamples. Thus, the ML prediction of the difference cannot be directly validated in a holdout sample. Lechner (2018) obtains the following result for the mean square error of an estimator of the IATE at a given point *x*:

$$MSE\left(\widehat{IATE}(m,l;x)\right) = MSE\left(\hat{\mu}_m(x)\right) + MSE\left(\hat{\mu}_l(x)\right) - 2MCE\left(\hat{\mu}_m(x), \hat{\mu}_l(x)\right),$$

where $MCE\left(\hat{\mu}_m(x), \hat{\mu}_l(x)\right) = E\left(\hat{\mu}_m(x) - \mu_m(x)\right)\left(\hat{\mu}_l(x) - \mu_l(x)\right)$. This derivation of the mean square error is instructive and is the basis of the *mcf* as proposed by Lechner (2018). The main idea is to approximate this term directly. The final splitting criterion is based on the sum of expected MSEs at a given point *x* over all unique treatment combinations, i.e., all parameters are of equal importance:

$$\overline{MSE}_x = \sum_{m=0}^{M-1} \sum_{l=m+1}^{M} E\left(MSE\left(\widehat{IATE}(m,l;x)\right)\right).$$

For finding splits based on this criterion, the MSE and MCE elements need to be estimated in the daughter leaves. Let $N^d_{S(x)}$ denote the number of observations with treatment value *d* in a certain stratum (leaf) *S(x)*, defined by the values of the features *x*. Then a 'natural' choice to estimate the MSEs in leaf *S(x)* is:

$$\widehat{MSE}_{S(x)} = \frac{1}{N^d_{S(x)}} \sum_{i=1}^{N} \underline{1}(x_i \in S(x)) \underline{1}(d_i = d) \left(\hat{y}^d_{S(x)} - y_i\right)^2,$$

where $\hat{y}^d_{S(x)}$ is an average of the observed outcomes in treatment *d* in leaf *S(x)*. To compute the MCE, the *mcf* uses the closest neighbour (in terms of similarity w.r.t. *x*) available in the other treatment (which is denoted by $y_{(i,m)}$ below).[17]

[17] Implementational details on finding the closest neighbours for the MCE estimation are in Section B.3.1.

$$\widehat{MCE}_{S(x)} =$$

$$= \frac{1}{N^m_{S(x)} + N^l_{S(x)}} \sum_{i=1}^{N} \underline{1}(x_i \in S(x)) \left(\underline{1}(d_i = m) + \underline{1}(d_i = l) \right) \left(\hat{y}^m_{S(x)} - \tilde{y}_{(i,m)} \right) \left(\hat{y}^l_{S(x)} - \tilde{y}_{(i,l)} \right),$$

$$\tilde{y}_{(i,m)} = \begin{cases} y_i & \text{if } d_i = m \\ y_{(i,m)} & \quad d_i \neq m \end{cases}.$$

Analysing the estimated $\widehat{MSE}$ reveals that its minimization in the daughter leaves favours splits maximizing the differences of all $\hat{y}^d_{S(\cdot)}$ between the daughter leaves. Additionally, the splits also favour large differences between $\hat{y}^m_{S(\cdot)}$ and $\hat{y}^l_{S(\cdot)}$ in the same leaf. In case of binary treatment, this is equivalent to the sole focus on treatment effect heterogeneity of Wager and Athey (2018), and Athey et al. (2019) under no selection into treatment. In general, the *mcf* splitting approach is only asymptotically equivalent to maximization of treatment effect heterogeneity in case of binary treatment. For multiple treatments, the *mcf* asymptotically targets treatment effect heterogeneity across daughter leaves while simultaneously increasing treatment outcome variability within leaves. Overlooking this dual focus may lead to splits that, while maximizing treatment effect heterogeneity, fail to minimise the sample MSE of parameters within leaves. Thus, *mcf* utilizes a splitting strategy that jointly considers all treatment combinations.

Since inference is important in causal analysis, it is problematic if the MSE-minimal estimator has a substantial bias. Although asymptotically confounding should be taken care of by the splitting rule derived above, the *mcf* adds a penalty term to the splitting criterion as an additional safeguard. As before, denote by *S(x')* and *S(x'')* the values of the features in the daughter leaves resulting from splitting some parent leaf. Lechner (2018) proposes to add the following penalty to a combination of the two 'final' MSEs in the daughter leaves:

$$penalty(x', x'') = \lambda \left\{ 1 - \frac{1}{M} \sum_{d=0}^{M-1} \left(P(D = d \mid X \in S(x')) - P(D = d \mid X \in S(x'')) \right)^2 \right\}, \quad \lambda > 0.$$

The probabilities, local estimates of the propensity score, are estimated as relative shares of the respective treatments in the potential daughter leaves. The penalty term is zero if the split leads to a perfect prediction. It reaches its maximum value, λ, when all probabilities are equal. Thus, the algorithm prefers a split that is also predictive for $P(D=d|X=x)$. Of course, the choice of the exact form of this penalty function is arbitrary. Furthermore, there is the issue of how to choose λ (*without* expensive additional computations) which is taken up again in Section 5.2.2 which also discusses further improvements via local centring.

4.3.2 Theoretical guarantees for the IATEs

Next, we present the asymptotic properties of the *mcf* (abstracting from local centring).[18]

The plain-vanilla *mcf* procedure aggregates individual trees T into a forest in the following way: The data are split into the training set, $\Im_{tr}$, and honest set, $\Im_{ho}$, to form two fixed non-overlapping halves of the full data set of sizes $N_1=N_2=N/2$. The sampling rates for each tree are $N_1^{\beta_1}$ and $N_2^{\beta_2}$, $0<\beta_1<\beta_2<1$, for the training and honest set respectively, leading to subsample size $s_1 \sim N_1^{\beta_1} \propto N^{\beta_1}$ for the training set and $s_2 \sim N_2^{\beta_2} \propto N^{\beta_2}$ for the honest set.[19] We name this procedure 'two-sample honesty' as honest data will not become training data and vice versa. This honesty concept underpins the weights-based inference introduced in Lechner (2018), which is covered in Section 4.3.4 below.

Each tree of the form $T(x;\xi,\Im_{tr},\Im_{ho})$, where $\xi \sim \Xi$ is a source of auxiliary randomness in the tree building process (such as random choice of splitting variables), can be used to estimate $IATE(m,l;x)$. When not important for proofs, ξ and/or $\Im_{tr}$ and $\Im_{ho}$ will be suppressed in the notation for better readability. *mcf* is an average over B trees:

[18] More details of the assumptions needed as well as the formal proofs are contained in Appendix A.1.

[19] If N is an odd number, one of the two subsamples will contain one observation more than the other one without any consequences. Note also that the dependence of the subsample sizes s_1 and s_2 on N is suppressed in most of the following notation.

$$F(x;\Im_{tr},\Im_{ho})=\frac{1}{B}\sum_{b=1}^{B}T_b(x;\xi_b,\Im_{tr,b},\Im_{ho,b}),$$

where $\Im_{tr,b}$ and $\Im_{ho,b}$ are drawn without replacement from $\Im_{tr}$ and $\Im_{ho}$ at subsampling rates β_1 and β_2, respectively. ξ_b is a random draw from Ξ. The b subscript will be supressed when it will not lead to any confusion. Like in Wager and Athey (2018), the trees T need to be *honest*, as defined in Definition 1:

DEFINITION 1 A tree grown on a training sample $\Im_{tr,b}$ is *honest* if the tree does not use the responses Y from the honest sample $\Im_{ho,b}$ to place its splits.

Honesty is crucial for inference and to bound the bias of the *mcf*. The splitting and sub-sampling procedure described above guarantees that the trees are honest. The main difference to the definition of honesty of Wager and Athey (2018) and Athey at al. (2019) is the reverse order of steps. Wager and Athey (2018) first subsample and then split the data for their double-sample trees. Here, the data set is split first and then the training and honest sets are subsampled from the given split. This allows to better control bias and variance rates as the size of the training set will codetermine the size of the final leaves translating into bias and the size of the honest set will influence the variance rate.

To guarantee consistency, the final leaves must asymptotically shrink in all dimensions similarly as in Meinshausen (2006) and Wager and Athey (2018), invoking the following definition of random-split tree:

DEFINITION 2 A tree is considered a *random-split* tree if at every splitting step, marginalizing over ξ, there is a guaranteed minimum probability π / p that the next split occurs along the u-th feature for some $0 < \pi \leq 1$, for all $u = 1,..., p$.

There are several options how to obtain a random-split tree. For the strategy implemented in *mcf*, see Appendix B.3.1.

The following Definition 3 controls the shape of the leaves. Definition 4 imposes *symmetry*. Both are required to derive the asymptotic results.

DEFINITION 3 A tree predictor grown by recursive partitioning is (α, v)*-regular* for some $\alpha > 0$ if (1) each split leaves at least a fraction α of the available training examples of each treatment on each side of the split, (2) the leaf containing *x* has at least v observations from each of the *M* treatment groups for some $v \in \mathbb{N}$, and (3) the leaf containing *x* has at least one treatment with less than $2v-1$ observations.

Regarding the role of the splitting rule on (α, v)-regularity, the algorithm first determines splits that do not violate the regularity condition. For these, the splitting criterion is calculated and the split that achieves the minimum value of the objective function is chosen as the best one. By following this procedure, any influence of the splitting criterion (including the penalty) on the regularity of the final leaves can be ruled out.

DEFINITION 4 A predictor is *symmetric* if the output of the predictor does not depend on the order in which the observations are indexed in the training and honest samples.

Consider B trees satisfying Definitions 1-4, that training and honest data are obtained from a two-sample honesty procedure, and let $\hat{\theta}_{m,l}^{mcf}(x)$ denote an *mcf* estimator of *IATE(m,l;x)* obtained as

$$\widehat{IATE^{mcf}}(m,l;x) = \hat{\theta}_{m,l}^{mcf}(x) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N_2} w_{i,b}^{mcf}(d_i, x_i; x, m, l) y_i = \sum_{i=1}^{N_2} w_i^{mcf}(d_i, x_i; x, m, l) y_i \,,$$

where $w_{i,b}^{mcf}(d_i, x_i; x, m, l) = \left(\frac{\underline{1}(d_i = m)}{N_{S_b(x)}^m} - \frac{\underline{1}(d_i = l)}{N_{S_b(x)}^l} \right) \underline{1}(x_i \in S_b(x))$ represent the weights for the IATE estimate in tree *b*. This estimator is differencing the average outcomes of the treatment groups evaluated on the honest subsample in the tree leaf containing point *x*, $S_b(x)$. Averaging across trees allows for a weighted representation of the IATE forest estimator with forest weights $w_i^{mcf}(d_i, x_i; x, m, l) = \frac{1}{B}\sum_{b=1}^{B} w_{i,b}^{mcf}(d_i, x_i; x, m, l)$. Both forest weights $w_i^{grf}(x)$ and $w_i^{mcf}(d_i, x_i; x, m, l)$ represent how important observation *i* is for estimation of any quantity at point *x*. Meanwhile, the *mcf* weights weigh the observed outcome directly, *grf* applies the forest weights in a locally weighted moment function. Thus, a weighted representation of the final *grf* estimate of the IATE at point *x* weighs the observed outcome y_i by weights that combine $w_i^{grf}(x)$ with treatment assignments in a non-linear fashion (see the result in Subsection 4.2).

A further difference between the *grf* and the *mcf* is related to the number of forest weights used in the IATE estimator and how many trees contribute to the calculation of each weight. Both are determined by the corresponding honesty procedures. Subsection 4.2 summarizes that the *grf* forest weights represent how much each observation is relevant for *IATE(m,l;x)*. This means that there will be *N* estimated forest weights because each half-sample is drawn from the full sample. For large enough *B*, each observation has a chance to end up in the honest sample on which the weights are estimated. However, the half-sampling and further honesty split lead every forest weight $w_i^{grf}(x)$ to be based on approximately *B*/4 trees. On the other hand, *mcf* estimates only *N*/2 of forest weights due to the primary half-split of data into training and honest data set. However, since the same *N*/2 observations are used in each tree to estimate the weights for β_2=1, each forest weight is averaged across all *B* trees.[20] For a given *B* and *N*, *grf* has an advantage in smoothing over more observations and *mcf* in the precision of the weights, potentially influencing finite sample properties of both methods.

Given these concepts, we state the main theorems guaranteeing consistency and asymptotic Gaussianity of the *mcf* IATE estimator. Further assumptions necessary for achieving this asymptotic distribution in the case of i.i.d. sampling are (i) Lipschitz continuity of first and

[20] See Figure A.1 in Appendix A.2 for graphical representation.

second order moments of the outcome variable conditional on the features, (ii) using subsampling to obtain the training data for tree building (subsamples should increase with *N*, slower than *N*, but not too slow), and (iii) conditions on the features (independent, continuous with bounded support, *p* is fixed, i.e. low-dimensional). All proofs are collected in the Appendix A.1. Each section in this appendix contains all the necessary proofs and intermediate results for the corresponding main theorem. One of the main differences to the proofs in Wager and Athey (2018) is the different splitting and subsampling approach to achieve honesty. The Causal Forest estimator in Wager and Athey (2018) approximates a U-statistic. Thus, their proofs are based on the corresponding Gaussian theory. Since the *mcf* allows for β_2=1, the estimator cannot be generally interpreted as a U-statistic and so the results in Wager and Athey (2018) do not apply. Instead, the weighted representation of the forest estimate is leveraged to obtain asymptotic properties that also cover the case of β_2=1.

The first result is the bound on the bias of the forest. The proof is like the proof in Wager and Athey (2018). In the first step, we show that the leaves get small in volume as the s_1 gets large. In the second step, we show that the honest observations in the final leaf can be seen as a subset of nearest neighbours around the point x. Thus, their expected distance and Lipschitz continuity help to bound the bias. Lemma 1 in Appendix A.1.1 gives rates at which the Lebesgue measure of final leaves in a regular, random-split tree shrinks under the assumption of the features being independent from each other and uniformly distributed.

THEOREM 1 In addition to the conditions of Lemma 1 (regular, random-split trees and independent uniformly distributed features), suppose that trees T are honest and all $E\left(Y^d \middle| X = x\right)$ are Lipschitz continuous. Then, the absolute bias of the *mcf* IATE estimator at a given value of x is bounded by

$$\left|E\left(\hat{\theta}_{m,l}^{mcf}(x)\right) - \theta_{m,l}^{0}(x)\right| = O\left(s_1^{-\log(1-\alpha)/p\log(\alpha)}\right).$$

Note that the bias rate of the forest is the same as the bias rate of a single tree, as a forest prediction is the average of tree predictions. The bias rate is mainly driven by the shrinkage of the Lebesgue measure of the leaf that is influenced by the choice of parameter α. The upper bound in Theorem 1 resembles the bias rate of nearest-neighbours regression estimators under Lipschitz continuity in a p-dimensional leaf that shrinks at a rate of (1-α). This stems from the fact that the final leaves can be bounded by balls that shrink at the same rate as the final leaves as shown in the proof in the Appendix A.1.1. Note that this rate is rather conservative as it stems from bounding the shallowest leaf with a leaf that would always end up with a (1-α) share of observations at the same level of depth. The rate is faster than in Wager and Athey (2018) because we bound by the expected distance between the nearest neighbours and the point x instead of the longest expected diameter of the leaf.

The second result is the asymptotic distribution of the IATE estimate at point x.

THEOREM 2 Assume that there is a sample of size N containing i.i.d. data $(X_i, Y_i, D_i) \in [0,1]^p \times \mathbb{R} \times \{0,1,...,M-1\}$ for a given value of x. Moreover, features are independently and uniformly distributed $X_i \sim U\left([0,1]^p\right)$. Let T be an honest, regular, and symmetric random-split tree. Further assume that $E(Y^d \mid X = x)$ and $E((Y^d)^2 \mid X = x)$ are Lipschitz continuous and $Var(Y^d \mid X = x) > 0$. Then for $\beta_1 < \beta_2 < \frac{p+2}{p}\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1$,

$$\frac{\hat{\theta}^{mcf}_{m,l}(x) - \theta^0_{m,l}(x)}{\sqrt{Var\left(\hat{\theta}^{mcf}_{m,l}(x)\right)}} \xrightarrow{d} N(0,1).$$

The restrictions on the sampling rates require that (2+p) log(1-α) / (p log(α)) >1. This means that α needs to be set closer and closer to 0.5 for larger p. This is a consequence of the curse of dimensionality as values of α closer to 0.5 make sure that the shallowest final leaves get tighter upper bounds ensuring that the bias vanishes fast enough. Additionally, the relationship between the subsampling rates further ensures that the final leaf does not end up with too many honest observations and the squared bias–variance ratio converges to 0. The convergence of the bias-variance ratio to 0 together with a side result of the variance converging to 0 guarantees the consistency of the *mcf* IATE estimator as captured in the following corollary.

COROLLARY 2 Let all assumptions from Theorem 2 hold. Then, $\hat{\theta}^{mcf}_{m,l}(x) \xrightarrow{p} \theta^{0}_{m,l}(x)$.

4.3.3 GATE and ATE estimation

Estimates for GATEs and ATE are obtained by averaging the IATEs in the respective subsamples defined by *z* (assuming discrete *Z*) and Δ. Although estimating ATEs and GATEs directly instead of aggregating IATEs could lead to more efficient estimators (see Section 4.1 and 4.2), the computational burden would also be higher, in particular if the number of GATEs of interest is large, as is common in many empirical studies. Furthermore, there is no guarantee that the results are internally consistent, i.e., the respective averages of the IATEs are indeed close to their ATE and GATE counterparts. Therefore, letting $\hat{\theta}^{mcf}_{m,l}(x)$ be an estimator of $IATE(m,l;x)$, the *mcf* estimates GATEs and ATEs as appropriate averages of $\hat{\theta}^{mcf}_{m,l}(x)$s:

$$\widehat{GATE}^{mcf}(m,l;z,\Delta) = \hat{\theta}^{mcf}_{m,l}(z,\Delta) = \frac{1}{N_2^{z,\Delta}} \sum_{i=1}^{N_2} \underline{1}(z_i = z, d_i \in \Delta) \hat{\theta}^{mcf}_{m,l}(x_i)$$

$$= \sum_{i=1}^{N_2} w_i^{mcf}(d_i; z, m, l, \Delta) y_i;$$

$$w_i^{mcf}(d_i; z, m, l, \Delta) = \frac{1}{N_2^{z,\Delta}} \sum_{j=1}^{N_2} \underline{1}(z_j = z, d_j \in \Delta) w_i^{mcf}(d_i; x_j, m, l); \qquad N_2^{z,\Delta} = \sum_{i=1}^{N_2} \underline{1}(z_i = z, d_i \in \Delta).$$

$$\widehat{ATE}^{mcf}(m,l;\Delta) = \hat{\theta}^{mcf}_{m,l}(\Delta) = \frac{1}{N_2^{\Delta}} \sum_{i=1}^{N_2} \underline{1}(d_i \in \Delta) \hat{\theta}^{mcf}_{m,l}(x_i)$$

$$= \sum_{i=1}^{N_2} w_i^{mcf}(m, l, \Delta) y_i;$$

$$w_i^{mcf}(m, l, \Delta) = \frac{1}{N_2^{\Delta}} \sum_{j=1}^{N_2} \underline{1}(d_j \in \Delta) w_i^{mcf}(d_i; x_j, m, l); \qquad N_2^{\Delta} = \sum_{i=1}^{N_2} \underline{1}(d_i \in \Delta).$$

These expressions show that ATEs and GATEs have the same weights-based representation as the IATEs. Hence, asymptotic Gaussianity can be established in a similar way.

THEOREM 3 Let all assumptions of Theorem 2 hold. Then,

$$\frac{\hat{\theta}_{m,l}^{mcf}(\Delta)-\theta_{m,l}^{0}(\Delta)}{\sqrt{Var\left(\hat{\theta}_{m,l}^{mcf}(\Delta)\right)}}\xrightarrow{d}N(0,1).$$

Utilizing the same Central Limit Theorem for triangular arrays as in Theorem 3 and applying similar steps as in Theorem 1 to show that the bias-variance ratio converges to 0 yields the following corollary:

COROLLARY 3 Let all assumptions from Theorem 2 hold. Then,

$$\frac{\hat{\theta}_{m,l}^{mcf}(z,\Delta)-\theta_{m,l}^{0}(z,\Delta)}{\sqrt{Var\left(\hat{\theta}_{m,l}^{mcf}(z,\Delta)\right)}}\xrightarrow{d}N(0,1).$$

4.3.4 Inference

There are several suggestions in the literature on how to conduct inference and how to compute standard errors of Random Forest based predictions (e.g., Wager, Hastie, and Efron, 2014; Wager and Athey, 2018; and the references therein). Although these methods can be used to conduct inference on the IATEs, it is yet unexplored how these methods could be readily generalized to take account of the aggregation steps needed for the GATE and ATE parameters.

Therefore, the *mcf* uses an alternative inference method useful for estimators that have a representation as weighted averages of the observed outcome. This perspective is attractive for Random Forest based estimators as they consist of trees that first stratify the data (when building a tree), and subsequently average over these strata (when building the forest). Thus, the *mcf* exploits the weights-based representation explicitly for inference (see also Abadie and Imbens, 2006, for a related approach).

Considering a weights-based estimator with random weights W_i (that are normalized such that all weights add up to N) for $\hat{\theta}$ (i.e., IATEs, GATEs, or ATEs):

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} W_i Y_i; \qquad Var(\hat{\theta}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} W_i Y_i\right).$$

The variance of the estimator can be rewritten as:

$$Var\left(\frac{1}{N}\sum_{i=1}^{N} W_i Y_i\right) = E_W\left(\frac{1}{N^2}\sum_{i=1}^{N} W_i^2 \sigma_{Y|W}^2(W_i)\right) + Var_W\left(\frac{1}{N}\sum_{i=1}^{N} W_i \mu_{Y|W}(W_i)\right)$$

where $\mu_{Y|W}(W_i) = E(Y_i \mid W_i)$ and $\sigma_{Y|W}^2(W_i) = Var(Y_i \mid W_i)$. The derivation exploits the combination of two-sample honesty with an i.i.d sample. Remember that under two-sample honesty, observations can be either used for training or for estimation, without switching the roles.[21] Further recall that the *mcf* forest weights are computed as a function of training data determining the final leaves and value x_i picks the corresponding weight, leading to the following representation: $W_i = w(x_i, \Im_{tr})$. Therefore, under i.i.d. sampling, Y_i and W_j are independent. Thus, the conditioning set $W_1,...,W_N$ can be reduced to W_i for each conditional mean and conditional variance.

This leads to the following expression of the variance of the proposed estimators:[22]

$$Var\left(\frac{1}{N_2}\sum_{i=1}^{N_2} W_i Y_i\right) = E_W\left(\frac{1}{N_2^2}\sum_{i=1}^{N_2} W_i^2 \sigma_{Y|W}^2(W_i)\right) + Var_W\left(\frac{1}{N_2}\sum_{i=1}^{N_2} W_i \mu_{Y|W}(W_i)\right).$$

The above expression suggests using the following estimator:

[21] In contrast, the 'one-sample honesty' in Wager and Athey (2018) is based on continuously switching the role of observations used for tree building and effect estimation in their Causal Forest. Under this splitting procedure, each weight may still depend on many observations, and the conditioning set cannot be reduced as in the case of two-sample honesty.

[22] Note that the weighting estimator uses only the honest data, therefore the weights here sum up to N_2.

$$\widehat{Var(\hat{\theta})} = \frac{1}{N_2^2}\sum_{i=1}^{N_2} w_i^2 \hat{\sigma}^2_{Y|W}(w_i) + \frac{1}{N_2(N_2-1)}\sum_{i=1}^{N_2}\left(w_i \hat{\mu}_{Y|\hat{W}}(w_i) - \frac{1}{N_2}\sum_{i=1}^{N_2} w_i \hat{\mu}_{Y|W}(w_i) \right)^2 .$$

The conditional expectations and variances may be computed by standard non-parametric or ML methods, as this is a one-dimensional problem for which many well-established estimators exist. Bodory, Camponovo, Huber, and Lechner (2020) investigate $\tilde{k}$-nearest neighbour estimators to obtain estimates for these quantities. They found good results in a binary treatment setting for the ATET. The same method is used here.[23] As both, $\mu_{Y|\hat{W}}$ and $\sigma^2_{Y|\hat{W}}$, are bounded and the number of nearest neighbours in their estimation is chosen such that it grows slower than N_2 (as documented in the Appendix A.2), consistency of the conditional expectations is guaranteed, see, e.g., Devroye, Györfi, Krzyzak, and Lugosi (1994). Additionally, for larger N the non-zero weights concentrate at a close neighbourhood of the given point x as the leaves are shrinking towards 0. It is, however, beyond the scope of this paper to analyse rigorously the exact statistical conditions needed for this estimator to lead to valid inference.

5 Monte Carlo study

5.1 Concept

In this section, we compare the finite-sample performance of various advanced causal machine learning methodologies in estimating average effects across three specified aggregation levels. This comparison includes evaluating both the precision of point estimates and the robustness of inference procedures.

There are primarily two methodologies for such comparisons. A prevalent method in the machine learning field involves using established benchmark datasets to assess the performance

[23] They also found a considerable robustness on how exactly to compute the conditional means and variances. Note that since their results relate to aggregate treatment effect parameters, their generalisability to the level of IATE's is unclear.

of different estimators.[24] Although this method has been adapted for causal analysis,[25] it typically falls short in evaluating inference procedures without additional modifications. Consequently, we adopt an alternative strategy, employing a Monte Carlo simulation approach. This involves the repeated generation of synthetic data, allowing for extensive variability in the data generating process (DGP).

In executing this Monte Carlo approach, we meticulously construct artificial datasets, varying numerous elements of the DGP. These elements include the degree of selectivity, the type and quantity of covariates, sample size, diverse functional forms, varying degrees of effect heterogeneity, the influence of covariates on outcomes and heterogeneity, the share of treated within the sample, and the number of treatments.[26] Despite its significant computational demands, this approach offers a comprehensive exploration of various scenarios, surpassing the more limited scope of Monte Carlo studies typically employed in papers introducing a new methodology.

The subsequent subsection provides a concise overview of the simulation designs' components. The fundamental procedure involves initially generating random data, applying the different estimators to this training dataset, predicting effects on a separate dataset of equivalent size derived from the same DGP, recording the outcomes, and repeating these steps R times. After this process, we calculate a range of performance metrics that capture different dimensions of the accuracy and reliability of both estimation and inference processes.

[24] Typical repositories to find such data sets are for example the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/) or Kaggle (www.kaggle.com/datasets).

[25] See for example the data sets of Atlantic Causal Inference Conference 2019 challenge (sites.google.com/view/acic2019datachallenge/).

[26] Table B.2 in Appendix B.2.5 gives the list of the scenarios investigated and Appendix C collects the tables that contain the corresponding results.

5.2 Key features of the simulation study

5.2.1 Data Generating Process

The simulated (i.i.d.) data consist of the covariates, the treatment, and the potential outcomes. The simulation of these components will be discussed in turn (for additional details see Appendix B).

The simulation involves generating a range of 10 to 50 independent covariates (p=10, 20, 50). These covariates are normally (p^N), uniformly (p^U), or binomially (dummy variables, p^D) distributed, or as combinations of the three types ($p= p^N+ p^U + p^D$). Among these p covariates, the first k covariates (where $k = k^N + k^U + k^D$) influence both the selection into treatment and the potential outcomes. The effect of these k covariates is modelled to decrease linearly, ranging from 1 to 1/k. The base specification consists of 20 covariates ($p^N= p^U$=10, p^D=0) with half of them (in each category) relevant ($k^N= k^U$=5).[27]

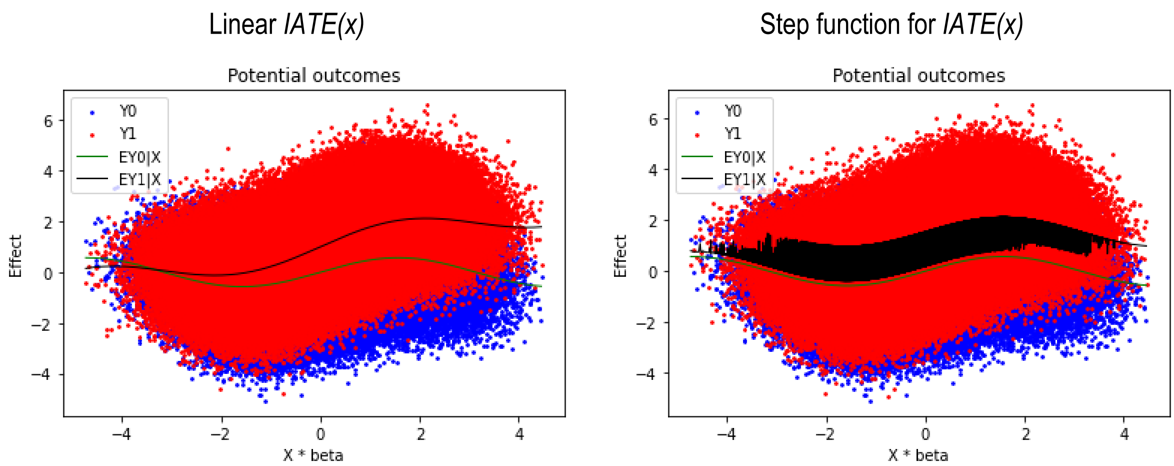
The selection process is based on a linear index function of the k relevant covariates plus noise. The quantiles of this index function are used to generate the treatments. The base specification considers cases with random selection (experiment), medium selection (true R^2 of about 10%), and strong selection (true R^2 of about 42%). Cases of 2 and 4 treatments with equal as well as asymmetric treatment shares are considered. The base specification consists of 2 treatments with equal treatment shares.

The non-treatment potential outcomes are obtained by simulating the expected non-treatment potential outcome as a sine-function of the linear index of the covariates plus noise. The relevance of the sine-function relative to the noise level is varied in the simulations from cases with true R^2 of 0 to 45% (base specification: 10%). The potential treatment outcomes are ob-

[27] When covariates are simulated from both, the normal and uniform distributions, the 1st, 3rd, 5th, … covariates are drawn from the uniform distribution.

tained by adding simulated IATEs plus noise to the expected non-treatment potential outcomes.[28] The IATEs are generated as functions of the linear index or as a step function of the first two, most important, covariates. In the former case, a linear function, a logistic function, and a quadratic function is specified. Finally, the step function approach follows closely the specification of Wager and Athey (2018). In all these cases the ATE is close to one. The case of zero IATEs, and thus zero ATE, is considered as well. The specification of the IATE as step function is used in the base scenario presented in Section 5.3. Figure 1 shows the expected and realised potential outcomes and their relation to the linear index. Thus, it summarizes main properties of the main base specification, as well as of the linear specification of the IATE. Similar plots for the other specifications can be found in Appendix B.2.4.

Figure 1: Shape of potential outcomes for different shapes of IATEs (base scenario)



Note: Figures are based on 1'000'000 observations.

The GATEs depend only on the first covariate, which is uniformly distributed if there are uniformly as well as normally distributed covariates in the DGP. This continuous covariate is split in groups with the same expected sizes and corresponding indicator variables are created. Thus, these dummy variables have no direct effect on the DGP. They are passed to the GATE estimators.

[28] The 2 (2 treatments) or 4 noise terms (4 treatments) used to simulate potential outcomes are independent of each other.

A final aspect of the DGP is the sample size. We mainly consider sample sizes of 2'500 and 10'000 as compromise between computational costs and practical relevance.[29] To keep the noise from the simulations on the performance measures stable (at least for the estimators of ATE and GATEs, which can be expected to show $\sqrt{N}-$convergence), the number of replications, R, declines at the same rate as the sample increases. Thus, the results for N=2'500 are based on R=1'000, while R declines to 250 for N=10'000.

In our simulation study, it's crucial to highlight that we adopt a methodology designed to replicate repeated sampling inference more closely, especially in scenarios involving stochastic covariates. To achieve this, the samples used for computing the out-of-training-sample effects are freshly drawn from the same Data Generating Process (DGP) for each replication. Additionally, these samples are matched in size to the training sample. This approach contrasts with the method used in Knaus et al. (2021), where the sample for calculating effects is drawn once prior to the initial replication and remains constant across subsequent replications, even when similar DGPs are employed.

The methodology chosen for our study, however, comes with certain trade-offs. Notably, for each Individual Average Treatment Effect (IATE) as a function of covariate values (IATE(x)), there exists only a single true and estimated value across all replications. This uniqueness arises because all or some covariates are continuous, leading to a scenario where each specific covariate value appears only once in the simulations. Consequently, this aspect of our methodology precludes the possibility of estimating higher-order moments for the IATEs. This is an important consideration to bear in mind when interpreting the results and the applicability of our findings.

[29] N=40'000 is also considered, but, for computational reasons, only for one specification.

5.2.2 Estimators

In the main part of the paper, we present the results of the Modified Causal Forest and the Generalized Random Forest, both with the outcome variable centred prior to estimation, *mcf-cent* and *grf-cent*, as well as of Double/debiased Machine Learning with normalized weights, *dml-norm*. For IATE estimation, we also include a more efficient version of the centred *mcf*, *mcf-cent-eff*, in the main part. The tables in Appendix C contain additional results for the un-centred *mcf* and *grf*, standard (non-normalized) *dml*, and *OLS*. We describe the implementation of these estimators briefly in this section and refer the interested reader to Appendix B.3 for details.

While Appendix B.3.1 details the implementation of the *mcf* further, at least 2 points merit some more discussion. The first point concerns the penalty parameter, λ. In the simulations, λ is set equal to *Var(Y)*. *Var(Y)* corresponds to the MSE when the effects are estimated by the sample mean without any splits. Thus, it provides some ad-hoc benchmark for plausible values of λ. In small-scale experiments with different values of λ, the MSE shows little sensitivity for values half and twice the size of *Var(Y)*. Generally, decreasing the penalty increases biases and reduces variances, et vice versa. The simulations below will show that biases are more likely to occur when selection is strong. Thus, if a priori knowledge about the importance of selectivity is available, then the researcher might adjust the penalty term accordingly.

Secondly, if inference is not a priority, like when using the IATEs as inputs into the training of an optimal assignment algorithm, the efficiency loss inherent in the *mcf*'s two-sample honesty approach can be avoided by cross-fitting, i.e., by repeating the estimation with exchanged roles of the two samples and averaging the two (or more) estimates. However, in such a case it is unclear how to compute the weights-based inference for the averaged estimator *mcf-cent-eff* as the two components of this average are correlated.[30] A similar efficiency loss can be

[30] For such a cross-fitted estimator (e.g., computed as mean of the single estimators), conservative inference could be obtained by basing inference on normality with a variance taken as average over the variances of the single estimations.

avoided in the *grf* procedure by switching off the half-sampling when inference is not relevant. However, the current implementation in the *grf* package does not support this option.

As mentioned in Section 4.2, *grf* performs local centring inside the Random Forest to remove confounding bias. The default *grf* local centring, labelled here as *uncentered*, refers to local centring that uses out-of-bag predictions of the outcome and treatment assignment from Random Forests trained on the full sample, to calculate the residuals. The simulation study includes an additional local centring strategy, labelled as *centred*. The difference between the centred and the uncentered *mcf* and *grf* is that the former uses a transformed outcome variable. This transformation subtracts a Random Forest prediction of $E(Y|X)$ from the observed Y (obtained with 5-fold-cross-fitting), and thus purges them from much of the influence of X. [31]

The *grf* GATEs and ATE are estimated via linear regression in which a variant of a local AIPW estimator is regressed on group indicators or a constant. The GATE variance estimator exploits the knowledge of a homoscedastic error term.

Compared to the standard *dml*, the normalized *dml* is more robust to extreme values of the estimated propensity score. It is obtained by normalizing the weights that are implicit in the *dml* scores. While it is straightforward to use *dml* for ATE, it is less straightforward to use them for heterogeneity estimation. Here, GATEs and IATEs are obtained by regression-type approaches in which the estimated components of the *dml* scores serve as dependent variable. The GATEs are obtained as OLS-coefficients of a saturated regression model with the indicators for the groups defined by the discrete variable Z as independent variables. IATEs are computed by using X as independent variables either in a regression Random Forest or in an OLS regression. When OLS is used, inference is based on the heteroscedasticity-robust covariance matrix of the corresponding coefficients. No inference is obtained for the Random Forest based IATEs.

[31] Note that the *grf* package does not provide centring of this kind. We added this estimator as the original version used in the *grf* package performed poorly in many of our DGPs for reasons discussed in Section 4.2.

All estimators are not tuned as computational costs would be prohibitive given the already extensive simulations. Instead, default values, as provided in the respective software packages, are used.

5.2.3 Performance measures

The main performance measures are the biases of the effects and their standard errors, the standard deviation of the effects, the mean absolute error, and the root mean squared error (RMSE) of the effects. The coverage probability (CovP) is reported to gauge the quality of the inference. It is presented at the 80% level since the limited number of replications for the larger sample leads to substantial simulation noise for, e.g., the more conventional 95% confidence interval.[32]

Note that, as mentioned above, the standard deviation, and thus the bias of the estimated standard error, cannot be computed for the IATE due to the simulation design.

Whenever several parameters are involved (as for the GATEs and IATEs), the performance measures are computed for each parameter and then averaged.

5.3 Results

In this subsection, we analyse *dml-norm*, *grf-cent*, and *mcf-cent* (when there is no confusion, in this subsection, we will drop the *-norm* and *-cent* ending) for the base Data Generating Processes (DGPs) and sample sizes of *N*=2,500 and *N*=10,000. Additionally, the degree of selectivity within the data is varied to investigate. Concerning the asymptotic ordering of the estimators, *dml* and *grf* are efficient for the ATE and the GATEs (as they are based on a discrete variable with few values). For the IATEs, to the best of our knowledge, no asymptotic ordering is possible (at least when there are continuous covariates).

[32] In addition to CovP for the 95% interval, the tables in the appendix also report the skewness and excess kurtosis of the estimators, which are, however, in a large majority of cases in the 'normal', unproblematic ranges.

Since it turned out that the relative performance of the estimators does not only depend on the strength of selection into treatment, but also on the specific parameter to be estimated, the results for the ATE, the GATEs, and the IATEs are discussed in turn.

5.3.1 *Average treatment effects*

Table 2 shows the results for the ATE. If there is no selectivity, like in an experiment, we obtain the expected results: the efficient estimators, *dml* and *grf*, are similar and outperform the *mcf*, at least when selection is not too strong. All estimators are essentially unbiased, and empirical coverage is close to the nominal 80% level. Furthermore, when the sample size quadruples, the standard deviation and RMSE halve, which is indicative of $\sqrt{N}$ – convergence.

The analysis reveals a notable trend with increasing selectivity: the performance of the normalized double/debiased machine learning (*dml-norm*) estimator deteriorates, particularly in terms of bias. This leads to a large increase in the RMSE. One plausible explanation is that (despite the normalisation) the double-robust score becomes more problematic when propensity score values become more extreme. A similar increase in bias is visible for the *grf*, although it is not as extreme as for *dml*. The resulting bias impacts also their coverage rates, which fall to very low levels. These issues appear to a much lesser extent for the *mcf*.

One summary coming from this table is that *mcf* appears to be more robust to stronger selectivity at the cost of some additional RMSE when selectivity does not matter much. The results in Appendix C.1 show that these performance patterns with respect to the degree of selectivity also appear for linear IATEs and non-linear IATEs. However, in the case of quadratic IATEs it turns out that the uncentered (!) *grf* outperforms the centred *grf* in terms of bias even for strong selectivity.

Fixing selectivity to medium levels and varying the other parameters of the DGP (Appendix C.2), suggests that the patterns observed in Table 2 qualitatively appear almost in all other DGPs as well. The most remarkable case is when covariates are more important for the non-treatment potential outcome (Table C.17, Table C.21): all estimators become biased, which increases the RMSE and leads to coverage far below nominal levels.

Table 2: Simulation results for average treatment effect (ATEs)

			Estimation of effects				Inference	
Estimator	Selec-tivity	Sample size	Bias	Mean absolute error	Std. dev.	RMSE	Bias (SE)	CovP (80) in %
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
dml-norm	None	2'500	-0.002	0.034	0.043	0.043	0.005	84
grf-cent			0.003	0.033	0.042	0.042	0.000	80
mcf-cent			0.003	0.048	0.061	0.061	-0.001	78
dml-norm	Med-ium	2'500	0.020	0.040	0.045	0.050	0.004	78
grf-cent			0.037	0.046	0.042	0.056	-0.000	64
mcf-cent			0.037	0.057	0.061	0.072	0.000	72
dml-norm	Strong	2'500	0.140	0.140	0.056	0.150	0.002	13
grf-cent			0.090	0.091	0.046	0.101	-0.006	20
mcf-cent			0.046	0.060	0.058	0.074	0.008	72
dml-norm	None	10'000	0.000	0.016	0.020	0.020	0.003	85
grf-cent			0.002	0.017	0.021	0.021	0.000	79
mcf-cent			0.004	0.022	0.028	0.028	0.002	81
dml-norm	Med-ium	10'000	0.010	0.019	0.022	0.024	0.003	82
grf-cent			0.015	0.022	0.022	0.027	-0.001	69
mcf-cent			0.014	0.025	0.027	0.030	0.003	82
dml-norm	Strong	10'000	0.079	0.079	0.032	0.085	0.000	12
grf-cent			0.042	0.044	0.027	0.050	-0.006	27
mcf-cent			-0.013	0.025	0.028	0.031	0.007	86

Note: RMSE abbreviates the Root Mean Squared Error. *CovP (80)* denotes the (average) probability that the true value is part of the estimated 80% confidence interval. 1'000 / 250 replications are used for 2'500 / 10'000 observations.

5.3.2 Conditional average treatment effects with a small number of groups (GATEs)

Table 3 shows that the relative performance of the different estimators for the GATEs depends not only on the strength of selectivity, but also on the number of groups for which a GATE is computed for. Table 3 shows the results for the cases of 5 and 40 groups, while Tables C.28 to C.30 in Appendix C.2 show the intermediate cases of 10 and 20 groups as well.

Concerning the point estimate, *mcf* outperforms *grf* and *dml* once there are 10 and more (*N*=2'500) or 20 and more groups (*N*=10'000), independent of the strength of selectivity. It is not surprising that *mcf* performs better for strong selectivity: the previous section showed that for strong selectivity the *mcf* dominates even for a GATE with only one group, i.e., the ATE.

Table 3: Simulation results for group average treatment effects (GATE)

			Estimation of effects				Inference	
Estimator	Selec-tivity	Sample size	Bias	Mean absolute error	Std. dev.	RMSE	Bias (SE)	CovP (80) in %
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				5 Groups				
dml-norm	None	2’500	-0.002	0.075	0.094	0.094	0.002	80
grf-cent			0.004	0.075	0.094	0.094	0.000	80
mcf-cent			0.002	0.090	0.091	0.114	-0.010	64
dml-norm	Med-ium	2’500	0.020	0.081	0.081	0.101	0.001	79
grf-cent			0.037	0.081	0.093	0.102	-0.000	76
mcf-cent			0.037	0.094	0.090	0.121	-0.006	65
dml-norm	Strong	2’500	0.140	0.157	0.123	0.187	-0.005	52
grf-cent			0.090	0.119	0.096	0.148	-0.007	56
mcf-cent			0.046	0.112	0.085	0.146	0.009	62
dml-norm	None	10’000	-0.001	0.037	0.046	0.046	0.001	82
grf-cent			0.002	0.036	0.046	0.046	0.001	80
mcf-cent			0.004	0.045	0.045	0.058	-0.002	68
dml-norm	Med-ium	10’000	0.010	0.041	0.050	0.051	0.000	80
grf-cent			0.016	0.041	0.048	0.051	-0.001	75
mcf-cent			0.013	0.049	0.045	0.063	0.004	69
dml-norm	Strong	10’000	0.078	0.090	0.073	0.108	-0.007	49
grf-cent			0.043	0.063	0.051	0.080	-0.006	56
mcf-cent			-0.014	0.085	0.045	0.100	0.007	41
				40 Groups				
dml-norm	None	2’500	-0.002	0.214	0.268	0.269	-0.003	79
grf-cent			0.004	0.213	0.269	0.269	-0.001	80
mcf-cent			0.002	0.096	0.096	0.120	-0.013	62
dml-norm	Med-ium	2’500	0.020	0.226	0.283	0.284	-0.005	79
grf-cent			0.037	0.213	0.265	0.268	-0.001	79
mcf-cent			0.036	0.099	0.095	0.128	-0.008	64
dml-norm	Strong	2’500	0.140	0.293	0.338	0.367	-0.020	71
grf-cent			0.090	0.223	0.255	0.280	-0.003	75
mcf-cent			0.043	0.119	0.090	0.153	0.007	61
dml-norm	None	10’000	-0.001	0.106	0.132	0.132	0.000	80
grf-cent			0.002	0.104	0.130	0.130	0.002	80
mcf-cent			0.004	0.050	0.050	0.063	-0.005	64
dml-norm	Med-ium	10’000	0.010	0.113	0.140	0.141	-0.001	79
grf-cent			0.016	0.106	0.131	0.132	0.000	79
mcf-cent			0.012	0.055	0.051	0.069	-0.002	64
dml-norm	Strong	10’000	0.079	0.164	0.192	0.209	-0.015	71
grf-cent			0.043	0.115	0.130	0.145	-0.002	74
mcf-cent			-0.016	0.091	0.050	0.104	0.005	37

Note: RMSE abbreviates the Root Mean Squared Error. *CovP (80)* denotes the (average) probability that the true value is part of the estimated 80% confidence interval. 1’000 / 250 replications are used for 2’500 / 10’000 observations.

We observe a differential dependence of the standard deviation on the number of groups: its increase is much slower for the *mcf* than for *dml* and *grf*. The reason is the way the GATE estimators are constructed by the different methods. As mentioned above, *dml* and *grf* are averaging double-robust scores within the cells of the discrete Z. Since the estimators are $\sqrt{N}$ – convergent, we expect that when the sample is reduced to one quarter of the original sample, the standard deviations of such estimators double. As the cells defined by the group variable in the DGPs are of approximately equal size, the number of groups and observations per group are inverse proportionally related. The resulting doubling of the standard deviation of the *grf* and *dml* estimators is exactly what is observed in Tables C.28 to C.30 when comparing results for 5 vs. 20 groups, or 10 vs. 40 groups, respectively.

The *mcf* aggregates IATEs within these cells. However, as the IATEs also use honest data outside of these cells, the standard deviation of the *mcf* increases much slower than for *grf* and *dml*.[33] In fact, we observe that when the number of groups increases further, the RMSE of the *mcf* approaches the one of the *mcf* IATEs from below, while the RMSE of *dml* and *grf* increases substantially and sometimes exceeds the RMSE of the IATEs which are of much higher dimension.

Concerning inference, the findings are a bit more pronounced: *dml* and *grf* have the correct coverage rates for no and medium selectivity even for 40 groups, while the coverage rates for *mcf* are too low. The figures contained at the end of Tables C.28, C.29, and C.30 in Appendix C indicate that this problem comes mainly from a bias of the GATEs with the smallest true values, while for the other GATEs coverage is close to the nominal level. In fact, the good performance in terms of RMSE and the problems with coverage are two sides of the same coin. Due to their aggregation from IATEs that are averaged over trees, *mcf* estimates share the

[33] The aggregation of IATEs that weigh information from observations belonging to other groups into a GATE resembles smoothing across (adjacent) categories in the context of categorical regressors. Heiler and Mareckova (2021) showed that optimal smoothing parameters in a non-parametric regression involving categorical regressors do not vanish asymptotically. Furthermore, they prove that the variance of the smoothed estimator is a weighted sum of the asymptotic variances from other categories. This finding provides a plausible rationale for the observed steadiness in the variance of the *mcf* estimator across varying group sizes.

strengths and the weaknesses of many smoothing methods: the variance is reduced at the cost of an increasing bias. To improve on the inference in finite samples, the simulations indicate that it needs some debiasing. This is however beyond the scope of this paper.

5.3.3 Individualized average treatment effects

Table 4 contains the results for the IATEs, averaged over all *N* IATEs. As inference is usually not the main goal when computing IATEs, for the centred *mcf* the more efficient version is included as well.

For the point estimates the ordering is clear-cut. *grf* and *mcf* are similar and outperform both *dml* versions (*ols*, *rf*). In almost all cases, the efficient centred *mcf* (*mcf -cent-eff*) performs best. This is also confirmed by the variations in the DGP as shown in Appendices C.1 and C.2. The only exception to this rule seems to be, again, the case of a quadratic function for the IATE, in which all methods seem to do (almost) equally bad with large RMSEs and large mean absolute errors. With respect to the other estimators, in a couple of cases OLS outperforms the more sophisticated CML estimators (but substantially underperforms in many other scenarios).

For inference, the message is not good: All estimators have coverage probabilities that are substantially below their nominal levels. However, given the level of granularity of the IATEs, this finding is of course not surprising. At least for the *mcf*, we conjecture that this problem comes from the biases of the individual IATEs. It is less likely that it comes from a too small estimated standard error, because weight-based standard error estimation is performed in a very similar way as for the ATE and the GATEs, in which it turned out to be almost unbiased.

Table 4: *Simulation results for individualized average treatment effects (IATEs)*

			Estimation of effects			Inference
Estimator	Selectivity	Sample size	Bias	Mean absolute error	RMSE	CovP (80) in %
(1)	(2)	(3)	(4)	(5)	(7)	(9)
dml-ols	None	2’500	-0.002	0.229	0.286	15
dml-rf			-0.002	0.270	0.342	-
grf-cent			0.004	0.155	0.187	57
mcf-cent			0.002	0.162	0.199	56
mcf-cent-eff			0.004	0.151	0.184	-
dml-ols	Med-ium	2’500	0.020	0.236	0.295	16
dml-rf			0.013	0.284	0.362	-
grf-cent			0.034	0.175	0.210	51
mcf-cent			0.037	0.174	0.212	53
mcf-cent-eff			0.040	0.163	0.197	-
dml-ols	Strong	2’500	0.140	0.285	0.357	18
dml-rf			0.117	0.349	0.461	-
grf-cent			0.092	0.254	0.297	33
mcf-cent			0.046	0.212	0.252	45
mcf-cent-eff			0.048	0.205	0.240	-
dml-ols	None	10’000	-0.001	0.179	0.219	4
dml-rf			0.000	0.201	0.256	-
grf-cent			0.002	0.082	0.103	76
mcf-cent			0.004	0.096	0.121	63
mcf-cent-eff			0.003	0.088	0.110	-
dml-ols	Med-ium	10’000	0.010	0.181	0.222	5
dml-rf			0.005	0.214	0.274	-
grf-cent			0.016	0.089	0.111	73
mcf-cent			0.013	0.106	0.131	59
mcf-cent-eff			0.013	0.100	0.111	-
dml-ols	Strong	10’000	0.079	0.205	0.256	8
dml-rf			0.058	0.275	0.418	-
grf-cent			0.053	0.123	0.162	66
mcf-cent			-0.014	0.156	0.185	43
mcf-cent-eff			-0.014	0.151	0.151	-

Note: *dml* is the normalized *dml* (denoted as *dml-norm. in Tables 1 and 2).* RMSE abbreviates the Root Mean Squared Error. *CovP (80)* denotes the (average) probability that the true value is part of the estimated 80% confidence interval. 1’000 / 250 replications are used for 2’500 / 10’000 observations.

5.3.4 Summary

dml shines when the target is low dimensional, such as the ATE and GATEs with few groups, and selectivity is not too strong. If selectivity is strong or the number of groups becomes too large, then its performance quickly deteriorates.

grf shows a similar good behaviour for the low dimensional parameters, as well as a similarly bad behaviour for GATEs with many groups. However, it performs well for IATEs. It is also less affected by strong selectivity than *dml*. These results however only hold for the modified version of the *grf* in which the outcome variable in the training data is explicitly centred before it enters the *grf* algorithm. The original uncentered version of *grf* as suggested in Athey et al. (2019) and implemented in their package underperforms in many scenarios due its bias problem (as does the uncentered version of the *mcf*).

Compared to *dml* and *grf*, the *mcf* generally shows a robust and competitive behaviour in many scenarios. The price to pay for this robustness and competitiveness is the somewhat higher standard deviation for the very low dimensional parameters like ATE and GATEs. Depending on the sample and the resulting precision of the estimators, this price may well be worthwhile paying. A further important advantage of the *mcf* is that its estimates are internally consistent over aggregation levels as ATE and GATEs are computed as averages of IATEs. This advantage becomes particularly apparent when the number of groups for the GATEs increases. While *dml* and *grf* have substantially higher RMSEs than for the more fain-grained IATEs, the uncertainty (and bias) in the *mcf* estimators smoothly increases with the number of groups, until it reaches the level of the most fine-grained heterogeneity parameter, the IATE.

6 Conclusion

Estimation of causal effects at different levels of granularity is of great importance for informed decision-making and tailored interventions. Lower levels of granularity capture the effect of a policy or intervention on a large population, guiding decisions on policies that cannot be targeted at individuals but must be deployed universally. Higher levels of granularity capture effects at a group or individual level that can serve for decisions how policies can be tailored more individually or for better understanding of the effects of large-scale policies at more granular levels. The complexity of such interventions requires methods that can estimate fine-grained heterogeneities of causal effects flexibly, such as some Causal Machine Learning (CML) methods.

In this paper, we investigate such CML methods subject to the restrictions that (i) they provide estimators of the causal effects at all aggregations levels, (ii) are essentially non-parametric, and (iii) that they allow for classical repeated sampling inference. Ideally, they are also internally consistent that aggregation of lower-level effects lead to the higher-level effects. Double/debiased Machine Learning (*dml*), the Generalized Random Forest (*grf*), and the Modified Causal Forest (*mcf*) fulfil these criteria and thus belong to the group of estimation methods which we call 'Comprehensive Causal Machine Learners' (CCML).

Here, we describe these estimators and their proven theoretical guarantees. For *dml* and the *grf*, they are already known, but not so for the *mcf*. Therefore, we explicitly provide them. The large-scale simulation study reveals scenarios in which the methods perform well. *dml* with normalized weights performs well in terms of RMSE and coverage probability when the target is low-dimensional, i.e., ATE and GATEs with few groups, and when the selection into treatment is not strong. *grf* shows similar behaviour as *dml*, however its performance for IATEs in terms of RMSE is much better. *mcf* has similar performance as *grf* in case of IATEs and outperforms *dml* and *grf* in scenarios with many GATEs or strong selection.

The results of the simulation study offer several practical recommendations. For low-dimensional targets when selection to treatment is moderate, *dml* is preferred including statistical inference. For IATEs, Causal Forest based methods perform well in terms of point estimation. When inference is not a priority, the more efficient version of *mcf* is recommended to estimate IATEs. For large groups, *mcf* is recommended for point estimation of GATEs. When sample is large enough and slight loss of efficiency is not detrimental, *mcf* can be used to estimate all effects due to its robustness to strong selection into treatment and large GATE groups. When internal consistency of the effects is important, only *mcf* can guarantee it due to its aggregation strategy.

The practical use of these three CCML methods is supported by the availability of well-maintained software packages: *dml* is available as Python and R packages, the g*rf* is available as an R package, and the *mcf* is available as Python package.[34] On a practical note, the results in this paper indicate that users of the *grf* package are encouraged to perform local centring of the outcome variable via *K*-fold cross-fitting to remove any potential bias (the same holds for *mcf*, where this step is already implemented in the package).

The simulation study also points to topics for further research. For example, the sensitivity of the *dml* estimates to strong selectivity opens a topic of how to make them more robust to extreme propensity scores without impairing its efficiency and statistical inference. Due to the smoothing character of Causal Forests, in finite samples *mcf* estimation of GATEs and IATEs balances the bias-variance trade-off yielding low RMSEs but impairing the coverage probability due to a small bias and low variance. Finding a way how to further de-bias the estimates would improve coverage probability in finite samples.

[34] For example, Bach, Chernozhukov, Kurz, and Spindler (2022, 2024), and Knaus (2022) for *dml*; Athey and Wager (2019) for *grf*; and Bodory, Busshoff, and Lechner (2022) for *mcf*.

References

Abadie, A., and G. W. Imbens (2006): Large Sample Properties of Matching Estimators for Average Treatment Effects, *Econometrica*, 74 (1), 235-267.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015): Estimating Conditional Average Treatment Effects, *Journal of Business & Economic Statistics*, 33:4, 485-505.

Alaa, A. M., Z. Ahmad, and M. van der Laan (2024). Conformal meta-learners for predictive inference of individual treatment effects. *Advances in Neural Information Processing Systems*, 36.

Alaa, A. M., and M. van der Schaar (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 30.

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

Athey, S., and G. W. Imbens (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113 (27), 7353–7360.

Athey, S., and G. W. Imbens (2017). The State of Applied Econometrics: Causality and Policy Evaluation, *Journal of Economic Perspectives*, 31 (7). 3-32.

Athey, S., and S. Wager (2019). Estimating Treatment Effects with Causal Forests: An Application, *Observational Studies*, 5, 21-35.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80 (4), 597-623.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized Random Forests. *The Annals of Statistics*, 47, 1148-1178.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behavioral Research*, 47 (1), 115-135.

Bach P., V. Chernozhukov, M. Kurz, & M. Spindler (2022). DoubleML – An Object-Oriented Implementation of Double Machine Learning in Python. *Journal of Machine Learning Research*, 23 (53), 1–6.

Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., & Spindler, M. (2024). Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study. *arXiv:2402.04674*.

Bach, P., V. Chernozhukov, M. S. Kurz, and M. Spindler (2024). DoubleML - An Object-Oriented Implementation of Double Machine Learning in R. *Journal of Statistical Software*, 108 (3), 1-56.

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

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

Bertrand, M., B. Crépon, A. Marguerie, and P. Premand (2021). Do Workfare Programs Live Up to Their Promises? Experimental Evidence from Cote D'Ivoire. NBER working paper.

Billingsley, P. (1968). *Convergence of Probability Measures.* 1st edition, New York: Wiley.

Bodory, H., H. Busshoff, and M. Lechner (2022). High resolution treatment effects estimation: Uncovering effect heterogeneities with the modified causal forest. *Entropy.* 24 (8), 1039.

Bodory, H., L. Camponovo, M. Huber, and M. Lechner (2020). The finite sample performance of inference methods for propensity score matching and weighting estimators, *Journal of Business and Economic Statistics*, 38, 183-200.

Boller, D., M. Lechner, and G. Okasa (2021). The Effects of Sport on Online Dating: Evidence from Causal Machine Learning. *arXiv*: 2104.0460.

Bradley, R. C. (1986). Basic properties of strong mixing conditions, In: Eberlein, E., Taqqu, M. S. (eds) *Dependence in Probability and Statistics. Progress in Probability and Statistics,* 11, Birkhäuser, Boston, MA.

Cannas, M., and B. Arpino (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, 61 (4), 1049-1072.

Caron, A., G. Baio, and I. Manolopoulou (2022). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185 (3), 1115-1149.

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: Double/debiased machine learning. *The Econometrics Journal*, 21 (1), C1-C68.

Cockx, B., M. Lechner, and J. Bollens (2023). Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium. *Labour Economics*, 80, 102306.

Davis, J. M., and S. B. Heller (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107, 546–550.

Devroye, L., L. Györfi, A. Krzyzak, and G. Lugosi (1994). On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates. *The Annals of Statistics,* 22 (3), 1371-1385.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics.

Fan, Q., Y. C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40 (1), 313-327.

Farbmacher, H., H. Kögel, and M. Spindler (2021). Heterogeneous effects of poverty on attention. *Labour Economics*, 71, 102028.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189 (1), 1-23.

Foster, D. J., and V. Syrgkanis (2023). Orthogonal statistical learning. *The Annals of Statistics*, *51*(3), 879-908.

Goller, D. (2023). Analysing a built-in advantage in asymmetric darts contests using causal machine learning. *Annals of Operations Research*, 325 (1), 649-679.

Goller, D., M. Lechner, A. Moczall, and J. Wolff (2020). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, 65, 101855.

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79 (3), 1053-1079.

Györfi, L., M. Kohler., A. Krzyzak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer New York.

Hansen, B. B. (2008). The prognostic analogue of the propensity score, *Biometrika*, 95 (3), 481-488.

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

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition, Springer (10th printing with corrections, 2013).

Heiler, P., and J. Mareckova (2021). Shrinkage for categorical regressors. *Journal of Econometrics*, 223 (1), 161-189.

Hill, J. L. (2011): Bayesian Nonparametric Modelling for Causal Inference, *Journal of Computational and Graphical Statistics*, 20 (1), 217–240.

Hodler, R., M. Lechner, and P. A. Raschky (2023). Institutions and the resource curse: New insights from causal machine learning. *Plos One*, 18 (6), e0284968.

Huber, M., M. Lechner, and C. Wunsch (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, 175 (1), 1-21.

Hurwicz, L. (1950). Generalization of the Concept of Identification, in *Statistical Inference in Dynamic Economic Models*. Cowles Commission Monograph, 10, New York: Wiley.

Imai, K., and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evaluation, *Annals of Applied Statistics*, 7 (1), 443–470.

Imai, K., and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76 (1), 243-263.

Imbens, G. W. (2000). The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*, 87, 706-710.

Imbens, G. W. (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 (1), 5-86.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17 (2), 3008-3049.

Knaus, M. C. (2021). A double machine learning approach to estimate the effects of musical practice on student's skills. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184 (1), 282-300.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25 (3), 602-627.

Knaus, M. C., M. Lechner, and A. Strittmatter (2021). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *The Econometrics Journal*, 24 (1), 134-161.

Knaus, M. C., M. Lechner, and A. Strittmatter (2022). Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach. *Journal of Human Resources*, 57 (2), 597-636.

Knittel, C. R., and S. Stolper (2021). Machine learning about treatment effect heterogeneity: The case of household energy use. In *AEA Papers and Proceedings,* 111, 440-444.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the National Academy of Sciences* (*PNAS)*, 116, 4156-4165.

Langenberger, B., Steinbeck, V., Schöner, L., Busse, R., Pross, C., & Kuklinski, D. (2023). Exploring treatment effect heterogeneity of a PROMs alert intervention in knee and hip arthroplasty patients: A causal forest application. *Computers in Biology and Medicine*, 107118.

Lechner, M. (2001). Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. In: M. Lechner and F. Pfeiffer (eds.), *Econometric Evaluation of Active Labour Market Policies*, 43-58, Heidelberg: Physica.

Lechner, M. (2018). Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*.

Lee, S., R. Okui, and Y.-J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function, *Journal of Applied Econometrics*, 32:1207–1225.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9 (4), 403-425.

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research, 7*, 983-999.

Neumann, M. H. (2013). A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics. *ESAIM: Probability and Statistics*, 17, 120-134.

Neyman, J. (1959). Optimal asymptotic tests of composite hypotheses. *Probability and Statistics*, 213-234.

Nie, X., and S. Wager (2021): Quasi-Oracle Estimation of Heterogeneous Treatment Effects, *Biometrika*, 108 (2), 299-319.

Pytka, K., & Gulyas, A. (2021). *Understanding the Sources of Earnings Losses After Job Displacement: A Machine-Learning Approach*. Verein für Socialpolitik/German Economic Association.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427), 846-866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90 (429), 106-121.

Robins, J. M., and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90 (429), 122-129.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931-954.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66, 688-701.

Salditt, M., T. Eckes, and S. Nestler (2023). A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. *Administration and Policy in Mental Health and Mental Health Services Research*, 1-24.

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.

Sexton, J., and P. Laake (2009). Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, 53 (3), 801-811.

Strittmatter, A. (2023). What is the value added by using causal machine learning methods in a welfare experiment evaluation? *Labour Economics*, 84, 102412.

Van der Laan, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, 2 (1).

Van der Laan, M. J., and S. Gruber (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8 (1).

Wager, S., and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv:1503.06388*.

Wager, S., and S. Athey (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Society*, 113:523, 1228-1242.

Wager, S., T. Hastie, and B. Efron (2014). Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *Journal of Machine Learning Research*, 15, 1625-1651.

Wendling, T., K. Jung, A. Callahan, A. Schuler, N. H. Shah, B. Gallego (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 1–16.

Westreich, D., J. Lessler, and M. J. Funk (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826-833.

Zimmert, M., and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.

Appendix A: More details on the *mcf*

Here, we present detailed proofs and some more details of the Causal Forest algorithms used in the EMCS.

Appendix A.1: Theoretical proofs

The following notation as in Wager and Athey (2018, WA18) is used for the asymptotic scaling: $f(s) \gtrsim g(s)$ means that $\liminf_{s\to\infty} f(s)/g(s) \geq 1$, and $f(s) \lesssim g(s)$ means that $\liminf_{s\to\infty} f(s)/g(s) \leq 1$. Further, $f(s) = \Omega(g(s))$ means that $\liminf_{s\to\infty} |f(s)|/g(s) > 0$, i.e., that $|f(s)|$ is bounded below by $g(s)$ asymptotically.

Appendix A.1.1: Proofs for Theorem 1

In this section, we collect all necessary results for the bias bound. In the first step, we show that the volume of the tree shrinks with larger subsample size s_1. We focus on the volume instead of the diameter in comparison to WA18 as the volume is important to determine the expected value of honest samples in the leaf as the subsampling rates in the training and honest set may differ.

LEMMA 1 (based on Lemma 1 in Wager and Athey, 2018) Let $S(x)$ be a final leaf containing the point x in a regular, random-split tree according to the definitions above and let $\lambda(S(x))$ be its Lebesgue measure. Suppose that $X_1,...,X_{s_1} \sim U\left([0,1]^p\right)$ independently. Then for $\alpha \leq 0.5$, the expected value of the Lebesgue measure of the final leaf has the following bounds

$$E\left[\lambda(S(x))\right] = O\left(s_1^{-\log(1-\alpha)/\log(\alpha)}\right),$$
$$E\left[\lambda(S(x))\right] = \Omega\left(s_1^{-1}\right).$$

Proof: Let $c(x)$ denote number of splits leading to the leaf $S(x)$ and $s_{1,d}$ be the number of observations treated with treatment d in the training subsample. By Wager and Walther (2015), in particular using their Lemma 12, Lemma 13 and Corollary 14, with high probability and simultaneously for all but last $O(\log(\log s_1))$ parent nodes above $S(x)$, the number of training examples in the node divided by s_1 is within a factor $1+o(1)$ of the Lebesgue measure of the node. Therefore, for large enough s_1 with probability greater than $1-1/s_1$ it holds that

$$\lambda(S(x)) \leq (1-\alpha+o(1))^{c(x)}.$$

To further evaluate the upper bound for $\lambda(S(x))$, the smallest number of splits that could lead to a leaf $S(x)$ needs to be determined. Let $s_{\min} = \min_d s_{1,d}$ denote the smallest treatment in the training subsample. By regularity, the following holds $s_{\min}\alpha^{c(x)} \leq 2v-1$. Since $s_{\min} \gtrsim s_1\varepsilon$, then $c(x) \geq \log\left((2v-1)/(s_1\varepsilon)\right)/\log(\alpha)$ for large s_1 and $\lambda(S(x))$ is bounded by

$$\lambda(S(x)) = O\left((1-\alpha)^{\log((2v-1)/(s_1\varepsilon))/\log(\alpha)}\right) = O\left(s_1^{-\log(1-\alpha)/\log(\alpha)}\right).$$

This also translates into the upper bound for $E\left[\lambda(S(x))\right]$.

Let $|S(x)|$ and $|S(x,d)|$ denote the number of observations in leaf $S(x)$ and the number of observations treated with d in leaf $S(x)$, respectively. Regarding the lower bound, by regularity $|S(x,d)| \geq v$ for all treatments. It can be shown that the probability of $|S(x)| \geq vM$ when $\lambda(S(x)) \leq vM/s_1$ decays at least at a rate of $1/s_1$ for uniformly distributed covariates. Therefore, $vM/s_1 \leq \lambda(S(x))$ with probability $1-O(1/s_1)$ yielding $E\left[\lambda(S(x))\right] = \Omega\left(s_1^{-1}\right)$. To guarantee that the upper bound is above the lower bound, $\alpha \leq 1/2$. ■

THEOREM 1 Under the conditions of Lemma 1**,** suppose moreover that trees T are honest and $E\left[Y^d \middle| X = x\right]$ are Lipschitz continuous. Then, the bias of the *mcf* IATE estimator at a given value of x is bounded by

$$\left|E\left[\hat{\theta}_{m,l}^{mcf}(x)\right] - \theta_{m,l}^{0}(x)\right| = O\left(s_1^{-\log(1-\alpha)/p\log(\alpha)}\right).$$

Proof: As forest is an average of trees, the rate of the bias of a tree is also the rate of the bias of the forest. Let $T_b(d;x)$ denote a tree estimate of a potential outcome of treatment d at point x of a form

$$T_b(d;x) = \sum_{j=1}^{s_2} \hat{w}_{bj,b}(d_{bj}, x_{bj}; x, d) y_{bj},$$

where we use the following notation for the weights:

$$\hat{w}_{bj,b}(d_{bj}, x_{bj}; x, d) = \begin{cases} \left|S_b(d,x)\right|^{-1} & if \quad x_{bj} \in S_b(x) \quad and \quad d_{bj} = d \\ 0 & else \end{cases}.$$

The tree estimator of the treatment effect is $T(m,l;x) = T(m;x) - T(l;x)$. As the treatment effect estimator is a difference of two potential outcome estimators, the absolute bias can be bounded by the rate of a bias of the potential outcome estimator as

$$\left|E\left[T(m,l;x)\right] - IATE(m,l;x)\right| \leq \sum_{d\in\{m,l\}} \left|E\left[T(d;x)\right] - E\left[Y^d \middle| X = x\right]\right|.$$

Therefore, in the following the bias of the tree estimator for the potential outcome will be analysed.

The next observation is that if $E\left[Y^d \middle| X = x\right]$ is Lipschitz, then $E\left[Y \middle| X = x, D = d\right]$ is also Lipschitz as the two expectations coincide under the CIA, CS and observation rule. By using Jensen's inequality and Lipschitz continuity, the absolute bias can be bounded by

$$
\begin{aligned}
\left|E\left[T_b(d;x)\right]-E\left[Y^d \middle| X=x\right]\right| &= \left|E\left[\frac{1}{|S(d,x)|}\sum_{i\in S(d,x)} Y_i\right]-E\left[Y^d \middle| X=x\right]\right| \\
&= \left|E\left[E\left[\frac{1}{|S(d,x)|}\sum_{i\in S(d,x)} Y_i \middle| S(d,x)\right]\right]-E\left[Y^d \middle| X=x\right]\right| \\
&= \left|E\left[\frac{1}{|S(d,x)|}\sum_{i\in S(d,x)} E\left[Y_i \middle| S(d,x)\right]\right]-E\left[Y^d \middle| X=x\right]\right| \\
&\le E\left[\frac{1}{|S(d,x)|}\sum_{i\in S(d,x)} \left|E\left[Y_i \middle| S(d,x)\right]-E\left[Y^d \middle| X=x\right]\right|\right] \\
&= E\left[\frac{1}{|S(d,x)|}\sum_{i\in S(d,x)} \left|E\left[Y_i \middle| S(d,x)\right]-E\left[Y_i \middle| X_i=x, D_i=d\right]\right|\right] \\
&\le E\left[\frac{1}{|S(d,x)|}\sum_{i\in S(d,x)} C_d \left\|X_i-x\right\|\right].
\end{aligned}
$$

As the α regularity would yield a very loose bound on the expected distance, we take a different approach here based on the nearest neighbours (NN) of the point x that also contain all the observations in the final leaf. Each final leaf can be bounded by a ball with the centre at x and radius equal to the longest segment of the leaf which we denote as $diam(S(x))$. Then the number of treated observations in the ball, denoted as $|B(S(x),d)|$, follows a binomial distribution for s_2 observations and the success probability being the Lebesgue measure of the ball that can be seen as a function of the $diam(S(x))$ since the features are independent and uniformly distributed. At the same time the number of observations in the final leaf $|S(x,d)|$ follows also a binomial distribution for s_2 observations and the success probability being the Lebesgue measure of the final leaf that can be seen as a function of the $diam(S(x))$ and the angles to the vertices from the origin point using a high-dimensional polar system. As both random variables depend on the $diam(S(x))$, we can conclude that $|B(S(x),d)| \le O\left(|S(x,d)|\right)$ with a constant larger than 1. As the ball contains the final leaf, the following inequality holds $\sum_{i\in S(d,x)} \left\|X_i-x\right\| \le \sum_{i\in B(S(x),d)} \left\|X_i-x\right\|$.

When we randomly split all data points in the honest subsample with treatment d into $\left|B(S(x),d)\right|+1$ segments, the first $\left|B(S(x),d)\right|$ segments will have a length $s_{2,d}/\left|B(S(x),d)\right|$. Denote $\tilde{X}_j^x$ as the first nearest neighbour in the j^{th} segment. Then,

$$\sum_{i\in B(S(x),d)}\left\|X_i-x\right\|\leq\sum_{j=1}^{|B(S(x),d)|}\left\|\tilde{X}_j^x-x\right\|.$$

Therefore, the absolute bias can be further bounded by

$$\begin{aligned}
\left|E\left[T_b(d,x)\right]-E\left[Y^d\left|X=x\right.\right]\right| &\leq C_d E\left[\frac{1}{\left|S(d,x)\right|}\sum_{j=1}^{|B(S(x),d)|}\left\|\tilde{X}_j^x-x\right\|\right]\\
&=C_d E\left[E\left[\frac{1}{\left|S(d,x)\right|}\sum_{j=1}^{|B(S(x),d)|}\left\|\tilde{X}_j^x-x\right\|\middle|\left|S(d,x)\right|\right]\right]\\
&=C_d E\left[E\left[\frac{\left|B(S(x),d)\right|}{\left|S(d,x)\right|}\left\|\tilde{X}_1^x-x\right\|\middle|\left|S(d,x)\right|\right]\right]\\
&\leq C_{d,B}E\left[E\left[\left\|\tilde{X}_1^x-x\right\|\middle|\left|S(d,x)\right|\right]\right]\\
&=C_{d,B}E\left[\left\|X_{\left(1,\left\lfloor s_{2,d}/\left|B(S(x),d)\right|\right\rfloor\right)}-x\right\|\middle|\left|S(d,x)\right|\right],
\end{aligned}$$

where $X_{(1,N)}$ denotes the first nearest neighbour among N observations and $C_{d,b}$ collects the Lipschitz constant and the constant from the ratio of the observations in the ball and the final leaf. For a fixed $\left|S(d,x)\right|$ in a regular, random split tree, we can use results in Györfi, Kohler, Krzyzak and Walk (2002) for nearest neighbour (NN) estimators that also use the expected distance of the first neighbours in their Theorem 6.2 and Lemma 6.4 yielding the bound

$$C_{d,B}E\left[\left\|X_{\left(1,\left\lfloor s_{2,d}/\left|B(S(x),d)\right|\right\rfloor\right)}-x\right\|\middle|\left|S(d,x)\right|\right]\leq\tilde{c}_d E\left[\left(\frac{\left|S(d,x)\right|}{s_{2,d}}\right)^{\frac{1}{p}}\right],$$

where $\tilde{c}_d$ collects the constant $C_{d,B}$ and all constants that emerge in the NN proof. The upper bound for the last expectation can be then found by applying Jensen's inequality,

$$E\left[E\left[\left(\frac{|S(d,x)|}{s_{2,d}}\right)^{\frac{1}{p}}\middle|\lambda(S(x))\right]\right]\leq E\left[\left(\frac{E\left[|S(d,x)|\|\lambda(S(x))\right]}{s_{2,d}}\right)^{\frac{1}{p}}\right]$$

$$\leq\tilde{c}_{\varepsilon}E\left[\left(\lambda(S(x))\right)^{\frac{1}{p}}\right]\leq\tilde{c}_{\varepsilon}\left(E\left[\lambda(S(x))\right]\right)^{\frac{1}{p}}$$

where $\tilde{c}_{\varepsilon}$ collects constants stemming from the common support assumption. Using the results from Lemma 1, the result at the tree level is

$$\left|E\left[T_b(d,x)\right]-E\left[Y^d\left|X=x\right]\right|=O\left(s_1^{-\log(1-\alpha)/p\log(\alpha)}\right).$$

The final constant is a function of the Lipschitz constant, constant from the ratio of the observations in the ball and the final leaf, common support parameter ε and k, the regularity parameter controlling the number of the observations in the final leaf. As the forest is average of trees, the result above holds also for the forest estimate. ■

Appendix A.1.2: Proofs for Theorem 2

The asymptotic Gaussianity proofs build on a central limit theorem for weakly dependent random variables of a form $A_{i,N_2}=\hat{W}_iY_i-E\left[\hat{W}_iY_i\right]$, introduced in Neumann (2013). In this section, we prove that A_{i,N_2} satisfy the conditions of the CLT yielding the first necessary result. As not all Y_i are unbiased estimators of $\mu_d(x)$, the weighted average is not an unbiased estimator. Therefore, in the second step it is necessary to show that the ratio of bias and variance converges to zero as the sample gets larger and the final leaf shrinks asymptotically.

For the analysis of the variance of the $IATE$ forest estimator, we make the following observation.

COROLLARY 1 For any forest estimator F that averages B tree estimators T, the rate of the forest variance $Var(F)$ is bounded from above and below by the rate of the individual tree variance $Var(T)$.

Proof: The variance of a forest in a simplified notation is

$$Var(F) = \frac{1}{B^2}\left[\sum_{b=1}^{B} Var(T_b) + \sum_{b=1}^{B}\sum_{b' \neq b} Cov(T_b, T_{b'})\right].$$

As a forest is an average of trees, the worst upper bound of the variance of the forest is the variance of an individual tree. Lemma 3 shows the upper bound for the tree estimate that converges to zero for $\beta_2 > \beta_1$.

Since any covariance has to go to zero as fast as the variance, the lower bound of the rate of the variance is also determined by the variance of an individual tree. The lower bound is scaled by $1/B$. ■

In the following, we therefore focus on deriving the properties of the tree weights.

LEMMA 2 Let $\hat{W}_{i,b} := \hat{W}_{bi,b}(D,X;d,x)$. Suppose that the assumptions from Lemma 1 hold and the tree is symmetric. Moreover $\beta_2 > \beta_1/2$. Then, the moments of the tree weights have the following rates and bounds:

a) $E\left[\hat{W}_{i,b}\right] \sim \frac{1}{N^{\beta_2}}$,

b) $E\left[\hat{W}_{i,b}^2\right] = \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 - 2\beta_2}\right)$ and $E\left[\hat{W}_{i,b}^2\right] = O\left(N^{\beta_1 - 2\beta_2}\right)$,

c) $Var(\hat{W}_{i,b})$ has the same bounds as $E\left[\hat{W}_{i,b}^2\right]$,

d) $\left|Cov\left(\hat{W}_{i,b}, \hat{W}_{j,b}\right)\right| = \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 - 3\beta_2}\right)$ and $\left|Cov\left(\hat{W}_{i,b}, \hat{W}_{j,b}\right)\right| = O\left(N^{\beta_1 - 3\beta_2}\right)$.

Proof:

a) Due to symmetry, the expected value of the first moment of the tree weights can be expressed as:

$$
\begin{aligned}
E\left[\hat{W}_{i,b}\right] &= E\left[E\left[\hat{W}_{i,b} \| S(d,x)|\right]\right] \\
&= E\left[\frac{s_2 - |S(d,x)|}{s_2} \cdot 0 + \frac{|S(d,x)|}{s_2} \frac{1}{|S(d,x)|}\right] \\
&= \frac{1}{s_2} \sim \frac{1}{N^{\beta_2}}.
\end{aligned}
$$

b) Let $\tilde{p}_d(S(x)) = s_2 p_d(S(x)) / s_{2,d}$ where $p_d(S(x))$ is the propensity score of getting treatment d when on leaf $S(x)$. Due to the common support assumption $\tilde{p}_d(S(x))$ must lie in an interval $\left(\varepsilon/(1-\varepsilon),(1-\varepsilon)/\varepsilon\right)$. Thus, under symmetry, the expected value of the second moment of the tree weights can be expressed as:

$$
\begin{aligned}
E\left[\hat{W}_{i,b}^2\right] &= E\left[E\left[\hat{W}_{i,b}^2 \| S(d,x)|\right]\right] \\
&= E\left[\frac{s_2 - |S(d,x)|}{s_2} \cdot 0 + \frac{|S(d,x)|}{s_2} \frac{1}{|S(d,x)|^2}\right] \\
&= \frac{1}{s_2} E\left[\frac{1}{|S(d,x)|}\right] = \frac{1}{s_2} E\left[\frac{1-(1-\lambda(S(x))\tilde{p}_d(S(x)))^{s_{2,d}}}{s_{2,d}\lambda(S(x))\tilde{p}_d(S(x))}\right],
\end{aligned}
$$

where the last equality uses the fact that $|S(d,x)|$ is a positive binomial random variable. Since $E\left[1/|S(d,x)|\right] > 0$, $E\left[1/s_{2,d}\lambda(S(x))\tilde{p}_d(S(x))\right]$ determines the lower bound. The lower bound can then be derived as follows using Jensen's inequality and results in Lemma 1:

$$
E\left[\hat{W}_{i,b}^2\right] = \Omega\left(\frac{1}{s_2 s_{2,d} E\left[\lambda(S(x))\right]}\right) = \Omega\left(N^{-2\beta_2 + \frac{\log(1-\alpha)}{\log(\alpha)}\beta_1}\right).
$$

The upper bound can be similarly derived for $\beta_2 > \beta_1 / 2$ as

$$E\left[\frac{1}{|S(x,d)|}\right]=E\left[\frac{1-(1-\lambda(S(x))\tilde{p}_d(S(x)))^{s_{2,d}}}{s_{2,d}\lambda(S(x))\tilde{p}_d(S(x))}\right]$$
$$=E\left[\frac{1-(1-\lambda(S(x))\tilde{p}_d(S(x)))^{s_{2,d}}}{s_{2,d}\lambda(S(x))\tilde{p}_d(S(x))}\middle|\lambda(S(x))\leq vM/s_1\right]\Pr\left(\lambda(S(x))\leq vM/s_1\right)$$
$$+E\left[\frac{1-(1-\lambda(S(x))\tilde{p}_d(S(x)))^{s_{2,d}}}{s_{2,d}\lambda(S(x))\tilde{p}_d(S(x))}\middle|\lambda(S(x))>vM/s_1\right]\Pr\left(\lambda(S(x))>vM/s_1\right)$$
$$\leq 1\cdot O(1/s_1)+O(s_1/s_{2,d})\cdot 1=O\left(1/N^{\beta_2-\beta_1}\right).$$

The two results yield

$$E\left[\hat{W}_{i,b}^2\right]=\Omega\left(N^{-2\beta_2+\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1}\right),$$

$$E\left[\hat{W}_{i,b}^2\right]=O\left(N^{-2\beta_2+\beta_1}\right).$$

Both rates have constants that depend on ε, the common support parameter and regularity parameter *v*. The upper bound also depends on the number of treatments as the number of treatments influences the smallest expected Lebesgue measure of the leaf.

c) Since $E\left[\hat{W}_{i,b}^2\right]\rightarrow 0$ for $\beta_2>\beta_1/2$, the bounds for $Var(\hat{W}_{i,b})$ are the same as for $E\left[\hat{W}_{i,b}^2\right]$.

d) The covariance can be expressed as

$$Cov(\hat{W}_{i,b},\hat{W}_{j,b})=Cov\left(\hat{W}_{i,b},1-\sum_{k\neq j}\hat{W}_{k,b}\right)=-Var(\hat{W}_{i,b})-\sum_{k\neq i,j}Cov(\hat{W}_{i,b},\hat{W}_{k,b})$$

yielding

$$Cov(\hat{W}_{i,b},\hat{W}_{j,b})=-\frac{Var(\hat{W}_{i,b})}{s_2-1}.$$

The bounds for the covariance are

$$\left|Cov\left(\hat{W}_{i,b},\hat{W}_{j,b}\right)\right| = \Omega\left(N^{-3\beta_2+\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1}\right),$$

$$\left|Cov\left(\hat{W}_{i,b},\hat{W}_{j,b}\right)\right| = O\left(N^{-3\beta_2+\beta_1}\right).$$

Note that

$$\sum_{k\neq j} Cov(\hat{W}_{i,b},\hat{W}_{k,b}) = \frac{Var(\hat{W}_{i,b})}{s_2-1}. \blacksquare$$

LEMMA 3 Suppose that the tree conditions from Lemma 2 hold i.e., we build regular, random-split, symmetric trees. Additionally, the tree is honest and $E\left[Y^d \middle| X=x\right]$ is Lipschitz continuous. Moreover, assume that $E\left[\left(Y^d\right)^2 \middle| X=x\right]$ is also Lipschitz continuous and $Var\left(Y^d \middle| X=x\right) > 0$. Further, the sampling rates satisfy $\beta_2 > \beta_1$. Then the tree variance has the following rates

$$Var\left(\sum_{i=1}^{s_2}\hat{W}_{i,b}Y_i\right) = O\left(N^{\beta_1-\beta_2}\right),$$
$$Var\left(\sum_{i=1}^{s_2}\hat{W}_{i,b}Y_i\right) = \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1-\beta_2}\right).$$

Proof: The variance of a tree estimator of $\mu_d(x)$ can be decomposed as

$$Var\left(\sum_{i=1}^{s_2}\hat{W}_{i,b}Y_i\right) = \sum_{i=1}^{s_2}Var\left(\hat{W}_{i,b}Y_i\right) + \sum_{i=1}^{s_2}\sum_{j\neq i}Cov\left(\hat{W}_{i,b}Y_i,\hat{W}_{j,b}Y_j\right).$$

The upper bound for $Var\left(\hat{W}_{i,b}Y_i\right)$ is

$$
\begin{aligned}
Var\left(\hat{W}_{i,b}Y_i\right) &= E\left[\hat{W}_{i,b}^2 Y_i^2\right] - E^2\left[\hat{W}_{i,b}Y_i\right] \\
&\le E\left[\hat{W}_{i,b}^2 Y_i^2\right] = E\left[E\left[\hat{W}_{i,b}^2 \middle| X_i\right] E\left[Y_i^2 \middle| X_i\right]\right] \\
&\le E\left[E\left[\hat{W}_{i,b}^2 \middle| X_i\right] E\left[\sup_{d,x} E\left[\left(Y_i\right)^2 \middle| X_i = x, D_i = d\right]\right]\right] \\
&= E\left[E\left[\hat{W}_{i,b}^2 \middle| X_i\right] E\left[\sup_{d,x} E\left[\left(Y_i^d\right)^2 \middle| X_i = x\right]\right]\right] \\
&\le E\left[\hat{W}_{i,b}^2\right] C_2 = O\left(N^{-2\beta_2+\beta_1}\right),
\end{aligned}
$$

using the identifying assumptions and Lipschitz continuity of $E\left[\left(Y_i^d\right)^2 \middle| X_i\right]$ on a bounded co-variate space. This yields that

$$
\sum_{i=1}^{s_2} Var\left(\hat{W}_{i,b}Y_i\right) \le s_2 E\left[\hat{W}_{i,b}^2\right] C_2 = O\left(N^{\beta_1-\beta_2}\right).
$$

Note that this upper bound converges to zero for $\beta_2 > \beta_1$. In order to derive the upper bound for the covariance part, we rewrite the covariance as

$$
\begin{aligned}
\left|Cov(\hat{W}_{i,b}Y_i, \hat{W}_{j,b}Y_j)\right| &= \left|Cov\left(\hat{W}_{i,b}Y_i, \left(1 - \sum_{k \ne j} \hat{W}_{k,b}\right) Y_j\right)\right| \\
&= \left|-\sum_{k \ne j} Cov\left(\hat{W}_{i,b}Y_i, \hat{W}_{k,b}Y_j\right)\right| \\
&= \left|\sum_{k \ne j} Cov\left(\hat{W}_{i,b}Y_i, \hat{W}_{k,b}Y_j\right)\right|.
\end{aligned}
$$

Next, we derive the upper bound of $\sum_{k \neq j} Cov\left(\hat{W}_{i,b} Y_i, \hat{W}_{k,b} Y_j\right)$

$$
\begin{aligned}
\sum_{k \neq j} Cov\left(\hat{W}_{i,b} Y_i, \hat{W}_{k,b} Y_j\right) &= \sum_{k \neq j} E\left[\hat{W}_{i,b} Y_i \hat{W}_{k,b} Y_j\right] - E\left[\hat{W}_{i,b} Y_i\right] E\left[\hat{W}_{k,b} Y_j\right] \\
&= \sum_{k \neq j} E\left[\hat{W}_{i,b} Y_i \hat{W}_{k,b}\right] E\left[Y_j\right] - E\left[\hat{W}_{i,b} Y_i\right] E\left[\hat{W}_{k,b}\right] E\left[Y_j\right] \\
&= \sum_{k \neq j} \left(E\left[\hat{W}_{i,b} Y_i \hat{W}_{k,b}\right] - E\left[\hat{W}_{i,b} Y_i\right] E\left[\hat{W}_{k,b}\right]\right) E\left[Y_j\right] \\
&= \sum_{k \neq j} \left(E\left[E\left[\hat{W}_{i,b} \hat{W}_{k,b} | X_i, X_k\right] E\left[Y_i | X_i\right]\right]\right. \\
&\quad \left. - E\left[E\left[\hat{W}_{i,b} | X_i\right] E\left[Y_i | X_i\right]\right] E\left[\hat{W}_{k,b}\right]\right) E\left[Y_j\right] \\
&= \sum_{k \neq j} \left(E\left[\left(E\left[\hat{W}_{i,b} \hat{W}_{k,b} | X_i, X_k\right] - E\left[\hat{W}_{i,b} | X_i\right] E\left[\hat{W}_{k,b}\right]\right) E\left[Y_i | X_i\right]\right]\right) E\left[Y_j\right] \\
&\leq \sum_{k \neq j} \left(E\left[\hat{W}_{i,b} \hat{W}_{k,b}\right] - E\left[\hat{W}_{i,b}\right] E\left[\hat{W}_{k,b}\right]\right) C_1^2 \\
&= \sum_{k \neq j} Cov(\hat{W}_{i,b}, \hat{W}_{k,b}) C_1^2 = \frac{Var(\hat{W}_{i,b})}{s_2 - 1} C_1^2,
\end{aligned}
$$

where the inequality uses the same supremum argument as the proof for the variance, the fact that the expected potential outcomes are Lipschitz continuous on a bounded covariate space, i.e., $E\left[Y_i^d | X_i\right] \in \left[-C_1, C_1\right]$ and $E\left[Y_i^d\right] \in \left[-C_1, C_1\right]$ for some positive constant C_1, and the fact that the final sum of covariances is positive and therefore the product certainly bounds the original sum from above. This yields an upper bound for

$$
\begin{aligned}
\sum_{i=1}^{s_2} \sum_{j \neq i} \left| Cov\left(\hat{W}_{i,b} Y_i, \hat{W}_{j,b} Y_j\right) \right| &\leq s_2 (s_2 - 1) \frac{Var(\hat{W}_{i,b})}{s_2 - 1} C_1^2 \\
&= O\left(N^{\beta_1 - \beta_2}\right)
\end{aligned}
.
$$

The overall upper bound for the tree variance is

$$
Var\left(\sum_{i=1}^{s_2} \hat{W}_{i,b} Y_i\right) = O\left(N^{\beta_1 - \beta_2}\right) .
$$

The constant depends on ε, the common support parameter, regularity parameter ν, number of treatments M and a constant related to Lipschitz continuity of the outcome variable.

Due to non-negativity of the variance, it is enough to find the lower bound for $E\left[\hat{W}_{i,b}^2 Y_i^2\right]$ to analyze the lower bound of $Var\left(\hat{W}_{i,b} Y_i\right)$.

$$
\begin{aligned}
E\left[\hat{W}_{i,b}^2 Y_i^2\right] &= E\left[E\left[\hat{W}_{i,b}^2 | X_i\right] E\left[Y_i^2 | X_i\right]\right] \\
&\geq E\left[E\left[\hat{W}_{i,b}^2 | X_i\right] E\left[\inf_{d,x} E\left[Y_i^2 | X_i = x, D_i = d\right]\right]\right] \\
&= E\left[E\left[\hat{W}_{i,b}^2 | X_i\right]\right] E\left[\inf_{d,x} E\left[\left(Y_i^d\right)^2 | X_i = x\right]\right] \\
&= E\left[\hat{W}_{i,b}^2\right] E\left[\inf_{d,x} E\left[\left(Y_i^d\right)^2 | X_i = x\right]\right] \\
&= \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 - 2\beta_2}\right).
\end{aligned}
$$

Therefore, $\sum_{i=1}^{s_2} Var\left(\hat{W}_{i,b} Y_i\right) = \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 - \beta_2}\right)$.

The lower bound for the covariance bound can be derived with help of the triangular inequality,

$$
\begin{aligned}
\left|Cov(\hat{W}_{i,b} Y_i, \hat{W}_{j,b} Y_j)\right| &= \left|E\left[\hat{W}_{i,b} Y_i \hat{W}_{j,b} Y_j\right] - E\left[\hat{W}_{i,b} Y_i\right] E\left[\hat{W}_{j,b} Y_j\right]\right| \\
&\geq \left\|E\left[\hat{W}_{i,b} Y_i \hat{W}_{j,b} Y_j\right]\right| - \left|E\left[\hat{W}_{i,b} Y_i\right] E\left[\hat{W}_{j,b} Y_j\right]\right\|
\end{aligned}
$$

and finding lower bounds for the individual elements:

$$
\begin{aligned}
\left|E\left[\hat{W}_{i,b} Y_i \hat{W}_{j,b} Y_j\right]\right| &= \left|E\left[E\left[\hat{W}_{i,b} Y_i \hat{w} \hat{W}_{j,b} Y_j | \lambda(S(x))\right]\right]\right| \\
&= \left|E\left[E\left[1/\left|S(x,d)\right|^2 Y_i Y_j | \lambda(S(x)), X_i, X_j \in S(x)\right] \Pr\left(X_i, X_j \in S(x) | \lambda(S(x))\right)\right]\right| \\
&\geq \frac{1}{\left(s_{2,d}\right)^2} E\left[\inf_{d,x} E\left[\left|Y_i Y_j\right| D_i, D_j = d, X_i, X_j = x\right] \left(\lambda(S(x))\right)^2\right] \\
&= \Omega\left(N^{-2\beta_2 - 2\beta_1}\right)
\end{aligned}
$$

$$\begin{aligned}\left|E\left[\hat{W}_{i,b}Y_i\right]\right| &= \left|E\left[E\left[\hat{W}_{i,b}Y_i \middle| \lambda(S(x))\right]\right]\right| \\ &= \left|E\left[E\left[\hat{W}_{i,b}Y_i \middle| X_i \in S(x), \lambda(S(x))\right] \Pr\left(X_i \in S(x) \middle| \lambda(S(x))\right)\right]\right| \\ &\geq \frac{1}{s_{2,d}} E\left[\inf_{d,x} E\left[|Y_i| \middle| D_i = d, X_i = x\right] \lambda(S(x))\right] \\ &= \Omega\left(N^{-\beta_2-\beta_1}\right)\end{aligned}$$

Since $\left|E\left[\hat{W}_{i,b}Y_i\hat{W}_{j,b}Y_j\right]\right| = \Omega\left(N^{-2\beta_2-2\beta_1}\right)$, $\left|E\left[\hat{W}_{i,b}Y_i\right]\right|\left|E\left[\hat{W}_{j,b}Y_j\right]\right| = \Omega\left(N^{-2\beta_2-2\beta_1}\right)$ and $\left|E\left[\hat{W}_{i,b}Y_i\right]\right|\left|E\left[\hat{W}_{j,b}Y_j\right]\right| = O\left(N^{-2\beta_2}\right)$, the final lower bound is

$$\left|Cov(\hat{W}_{i,b}Y_i, \hat{W}_{j,b}Y_j)\right| = \Omega\left(N^{-2\beta_2-2\beta_1}\right).$$

Putting these results together yields a lower bound for covariance:

$$\sum_{i=1}^{s_2}\sum_{j\neq i}\left|Cov\left(\hat{W}_{i,b}Y_i, \hat{W}_{j,b}Y_j\right)\right| = \Omega\left(N^{-2\beta_1}\right).$$

As the lower bound for the covariance exceeds the lower bound for the variance, the lower bound for the tree variance is determined by the one converging slower to zero i.e.

$$Var\left(\sum_{i=1}^{s_2}\hat{W}_{i,b}Y_i\right) = \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1-\beta_2}\right).$$

The constant depends on ε, the common support parameter, regularity parameter k and a constant related to Lipschitz continuity of the outcome variable. ■

LEMMA 4 Let the assumptions from Lemma 3 hold and $A_{i,N_2} := \hat{W}_{i,N}Y_i - E\left[\hat{W}_{i,N}Y_i\right]$, where $\hat{W}_{i,N}$ are the forest weights. Then $(A_{i,N_2})_{i=1,\dots,N_2}$ is a triangular array satisfying:

a) $E[A_{i,N_2}] = 0$,

b) $\sum_{i=1}^{N_2} E[A_{i,N_2}^2] < \infty$ for all N and i,

c) $\sigma_N^2 := Var(A_{1,N_2} + \dots + A_{N_2,N_2}) \xrightarrow[N\to\infty]{} 0$,

d) $\sum_{i=1}^{N_2} E\left[A_{i,N_2}^2 \underline{1}\left(\left|A_{i,N_2}\right| > \tilde{\varepsilon}\right)\right] \xrightarrow[N\to\infty]{} 0$ for all $\tilde{\varepsilon} > 0$,

e) There is a summable sequence $(\pi_r)_{r\in\mathbb{N}}$ such that for all $u \in \mathbb{N}$ and all indices $1 \le s_1 < s_2 < ... < s_u < s_u + r = t_1 \le t_2 \le N_2$, the following upper bounds for covariances hold true for all measurable functions $g:\mathbb{R}^u \to \mathbb{R}$ with $\|g\|_\infty = \sup_{x\in\mathbb{R}^u} |g(x)| \le 1$:

$$\left|\mathrm{cov}(g(A_{s_1,N_2},...,A_{s_u,N_2})A_{s_u,N_2}, A_{t_1,N_2})\right| \le \left(E[A_{s_u,N_2}^2] + E[A_{t_1,N_2}^2] + N_2^{-1}\right)\pi_r$$

and

$$\left|\mathrm{cov}(g(A_{s_1,N_2},...,A_{s_u,N_2}), A_{t_1,N_2}, A_{t_2,N_2})\right| \le \left(E[A_{t_1,N_2}^2] + E[A_{t_2,N_2}^2] + N_2^{-1}\right)\pi_r .$$

Proof:

a) $E\left[\hat{W}_{i,N} y_i - E\left[\hat{W}_{i,N} y_i\right]\right] = 0 \cdot$

b) $E\left[A_{i,N_2}^2\right] = Var\left(\hat{W}_{i,N} Y_i\right) = \left(E\left[\hat{W}_{i,N}^2 Y_i^2\right] - E^2\left[\hat{W}_{i,N} Y_i\right]\right)$

Deriving upper bounds for $E[\hat{W}_{i,N} Y_i]$ and $E[\hat{W}_{i,N}^2 Y_i^2]$, we use that $\hat{W}_{i,N}$ and Y_i are independent conditional on X_i and $E\left[Y^d \middle| X = x\right]$ and $E\left[\left(Y^d\right)^2 \middle| X = x\right]$ are Lipschitz continuous and therefore can be bounded from above by $C_1 < \infty$ and $C_2 < \infty$ respectively on a bounded covariate space. The first and second moment of the forest weights are

$$E\left[\hat{W}_{i,N}\right] = \frac{N_2 - s_2}{N_2} \cdot 0 + \frac{s_2}{N_2} E\left[\frac{1}{B}\sum_{b=1}^{B} \hat{W}_{i,N,b}\right] = \frac{s_2}{N_2}\frac{1}{B} B \frac{1}{s_2} = \frac{1}{N_2} \sim N^{-1}$$

and

$$
\begin{aligned}
E\left[\hat{W}_{i,N}^2\right] &= \frac{N_2 - s_2}{N_2}\cdot 0 + \frac{s_2}{N_2} E\left[\left(\frac{1}{B}\sum_{b=1}^{B}\hat{W}_{i,N,b}\right)^2\right] \\
&= \frac{s_2}{N_2}\frac{1}{B^2}\left[\sum_{b=1}^{B} E\left[\hat{W}_{i,N,b}^2\right] + \sum_{b=1}^{B}\sum_{b'\neq b} E\left[\hat{W}_{i,N,b}\hat{W}_{\hat{W}i,N,b'}\right]\right] \\
&= \frac{s_2}{N_2}\frac{1}{B^2}\left[BE\left[\hat{W}_{i,N,b}^2\right] + B(B-1)\left(Cov(\hat{W}_{i,N,b},\hat{W}_{i,N,b'}) + \left(E\left[\hat{W}_{i,N,b}\right]\right)^2\right)\right] \\
&\leq \frac{s_2}{N_2}\frac{1}{B^2}\left[BE\left[\hat{W}_{i,N,b}^2\right] + B(B-1)\left(Var(\hat{W}_{i,N,b}) + \left(E\left[\hat{W}_{i,N,b}\right]\right)^2\right)\right] \\
&\leq \frac{s_2}{N_2}\frac{1}{B^2}\left[B^2 E\left[\hat{W}_{i,N,b}^2\right]\right] = O\left(N^{-1-\beta_2+\beta_1}\right).
\end{aligned}
$$

These results can be used to further bound the following expectations:

$$
E\left[\hat{W}_{i,N}Y_i\right] = E\left[E\left[\hat{W}_{i,N}\,|X_i\right]E\left[Y_i\,|X_i\right]\right] = O\left(N^{-1}\right),
$$

$$
E[\hat{W}_{i,N}^2 Y_i^2] = E\left[E\left[\hat{W}_{i,N}^2\,|X_i\right]E\left[Y_i^2\,|X_i\right]\right] = O\left(N^{-1-\beta_2+\beta_1}\right).
$$

It follows that

$$
\sum_{i=1}^{N_2} E[A_{i,N_2}^2] = O\left(N^{-\beta_2+\beta_1}\right) < \infty .
$$

c)
$$
\begin{aligned}
Var(A_{1,N_2} + \ldots + A_{N_2,N_2}) &= Var\left(\sum_{i=1}^{N_2}\hat{W}_{i,N}Y_i - \sum_{i=1}^{N_2}E\left[\hat{W}_{i,N}Y_i\right]\right) = Var\left(\sum_{i=1}^{N_2}\hat{W}_{i,N}Y_i\right) \\
&= \sum_{i=1}^{N_2} Var\left(\hat{W}_{i,N}Y_i\right) + \sum_{i=1}^{N_2}\sum_{j\neq i} Cov\left(\hat{W}_{i,N}Y_i, \hat{W}_{j,N}Y_j\right) \\
&\leq \sum_{i=1}^{N_2} Var\left(\hat{W}_{i,N}Y_i\right) + \sum_{i=1}^{N_2}\sum_{j\neq i} \left|Cov\left(\hat{W}_{i,N}Y_i, \hat{W}_{j,N}Y_j\right)\right|
\end{aligned}
$$

Using the results from b), the first element is bounded at rate $O\left(N^{-\beta_2+\beta_1}\right)$ and by a similar logic as in the tree case the sum of all covariances has to decay to zero as fast as the sum of variances. These results yield that $Var(A_{1,N_2} + \ldots + A_{N_2,N_2}) \xrightarrow[N\to\infty]{} 0$.

d) Due to monotonicity and the result in b):

$$\sum_{i=1}^{N_2} E\left[A_{i,N_2}^2 \underline{1}\left(\left| A_{i,N_2} \right| > \tilde{\varepsilon} \right) \right] \le \sum_{i=1}^{N_2} E\left[A_{i,N_2}^2 \right] \to 0 \text{ for all } \tilde{\varepsilon} > 0 .$$

e) Due to exchangeability which implies strict stationarity and the result in c), it is possible to interpret $\left(A_{i,N_2} \right)_{i=1,\dots,N_2}$ as a ρ-mixing process. Since every ϕ-mixing process is also ρ-mixing, then along the Lemma 20.1 in Billingsley (1968) for ϕ-mixing processes, we can also bound the covariances of a ρ-mixing process as follows:[35]

$$\begin{aligned} &\left| C\text{ov}\left(g\left(A_{s_1,N_2},\dots,A_{s_u,N_2} \right) A_{s_u,N_2}, A_{t_1,N_2} \right) \right| \\ &\le 2\sqrt{\phi_{t_1-s_u}} \sqrt{E\left[g^2\left(A_{s_1,N_2},\dots,A_{s_u,N_2} \right) A_{s_u,N_2}^2 \right]} \sqrt{E\left[A_{t_1,N_2}^2 \right]} \end{aligned}$$

and

$$\left| C\text{ov}(g(A_{s_1,N_2},\dots,A_{s_u,N_2}), A_{t_1,N_2} A_{t_2,N_2} \right| \le 2\phi_{t_1-s_u} E\left[\left| A_{t_1,N_2} A_{t_2,N_2} \right| \right] \le 2\phi_{t_1-s_u} E\left[A_{t_1,N_2}^2 \right]$$

where the last inequality follows from the stationarity of the process. The two results can be further bounded

$$\begin{aligned} \left| C\text{ov}(g(A_{s_1,N_2},\dots,A_{s_u,N_2}) A_{s_u,N_2}, A_{t_1,N_2} \right| &\le 2\sqrt{\phi_{t_1-s_u}} \sqrt{E\left[A_{s_u,N_2}^2 \right]} \sqrt{E\left[A_{t_1,N_2}^2 \right]} \\ &\le \sqrt{\phi_{t_1-s_u}} \left(E\left[A_{s_u,N_2}^2 \right] + E\left[A_{t_1,N_2}^2 \right] \right) \\ &\le \sqrt{\phi_{t_1-s_u}} \left(E\left[A_{s_u,N_2}^2 \right] + E\left[A_{t_1,N_2}^2 \right] + N_2^{-1} \right) \end{aligned}$$

by property of the function $g()$ and inequality of arithmetic and geometric mean and

$$\left| C\text{ov}(g(A_{s_1,N_2},\dots,A_{s_u,N_2}), A_{t_1,N_2} A_{t_2,N_2} \right| \le \sqrt{\phi_{t_1-s_u}} \left(E\left[A_{t_1,N_2}^2 \right] + E\left[A_{t_2,N_2}^2 \right] + N_2^{-1} \right)$$

by stationarity and the fact that the mixing coefficients are smaller or equal to 1. Then the weak dependence conditions are fulfilled for $\pi_r = \sqrt{\phi_r}$. ■

[35] Bradley (1986) showed that the coefficients of the two processes fulfil the following inequality: $\rho_r = 2\sqrt{\phi_r}$.

LEMMA 5 Define A_{i,N_2} as in Lemma 4, then $\widehat{IATE}(m,l;x) - E\left[\widehat{IATE}(m,l;x)\right] = \sum_{i=1}^{N_2} A_{i,N_2}$ and

$$\frac{\widehat{IATE}(m,l;x) - E\left[\widehat{IATE}(m,l;x)\right]}{\sqrt{Var(\widehat{IATE}(m,l;x))}} \to N(0,1) \text{ .}$$

Proof: The normality proof for triangular arrays satisfying the conditions in Lemma 4 is proven in Neumann (2013) and can be applied on estimation of potential outcomes $\mu_m(x)$ and $\mu_l(x)$. As those are estimated on an honest data set, the difference of the two quantities will also follow a normal distribution. ■

THEOREM 2 Assume that there are i.i.d. data $(X_i, Y_i, D_i) \in [0,1]^p \times \mathbb{R} \times \{0,1,...,M-1\}$ and a given value of x. Moreover, features are independently and uniformly distributed $X_i \sim U\left([0,1]^p\right)$. Let T be an honest, regular and symmetric random split tree. Further assume that $E\left[Y^d \middle| X = x\right]$ and $E\left[\left(Y^d\right)^2 \middle| X = x\right]$ are Lipschitz continuous and $Var\left(Y^d \middle| X = x\right) > 0$. Then for $\beta_1 < \beta_2 < \frac{p+2}{p}\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1$,

$$\frac{\widehat{IATE}(m,l;x) - IATE(m,l;x)}{\sqrt{Var(\widehat{IATE}(m,l;x))}} \to N(0,1) \text{.}$$

Proof: Given the result in Lemma 5, it remains to show that

$$\frac{E\left[\widehat{IATE}(m,l;x)\right] - IATE(m,l;x)}{\sqrt{Var(\widehat{IATE}(m,l;x))}} \to 0 \text{.}$$

The final result will follow then from Slutsky's lemma. By Theorem 1, we have

$$\left| E\left[\widehat{IATE}(m,l;x)\right] - IATE(m,l;x) \right| = O\left(s_1^{-\log(1-\alpha)/p\log(\alpha)}\right).$$

From Corollary 1, we get

$$Var(\widehat{IATE}(m,l;x)) = \Omega\left(N^{\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 - \beta_2} \right).$$

It follows that

$$\frac{\left(E\left[\widehat{IATE}(m,l;x) \right] - IATE(m,l;x) \right)}{\sqrt{Var(\widehat{IATE}(m,l;x))}} = O\left(N^{-\left(\frac{1}{p}+\frac{1}{2}\right)\frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 + \frac{\beta_2}{2}} \right).$$

The ratio converges to zero when

$$\frac{p}{2+p}\beta_2 < \frac{\log(1-\alpha)}{\log(\alpha)}\beta_1 . \blacksquare$$

Appendix A.1.3: Proofs for Theorem 3

THEOREM 3 Let all assumptions from Theorem 2 hold and define $\widehat{ATE}(m,l)$ as an average of all corresponding $\widehat{IATE}(m,l;x)$. Then,

$$\frac{\widehat{ATE}(m,l) - ATE(m,l)}{\sqrt{Var(\widehat{ATE}(m,l))}} \to N(0,1).$$

Proof: Using the CLT for triangular arrays of weakly dependent random variables requires to check that all requirements in Lemma 4 hold for $A_{i,N_2} := \hat{W}_i^{ATE(m,l)} Y_i - E\left[\hat{W}_i^{ATE(m,l)} Y_i \right]$. The proof uses the observation that the rates for the ATE cannot be worse than for the $IATE$ and, therefore, the conditions in Lemma 4 will be satisfied for the ATE weights. Due to the pointwise convergence at different points x, the convergence rate is affected.

a) $E\left[A_{i,N_2} \right] = 0$ holds trivially.

b) $\sum_{i=1}^{N_2} E\left[A_{i,N_2}^2\right] = \sum_{i=1}^{N_2} Var\left(\hat{W}_i^{ATE(m,l)} Y_i\right)$. The upper bound on the individual variances is the upper bound of $E\left[\left(\hat{W}_i^{ATE(m,l)}\right)^2 Y_i^2\right]$:

$$E\left[\left(\hat{W}_i^{ATE(m,l)}\right)^2 Y_i^2\right] = E\left[\left(\frac{1}{N_2}\sum_{j=1}^{N_2} \hat{W}_i^{IATE(m,l,x_j)}\right)^2 Y_i^2\right] =$$
$$\leq \frac{1}{(N_2)^2} E\left[(N_2)^2 E\left[\left(\hat{W}_i^{IATE(m,l,x_j)}\right)^2 \middle| X_i\right] E\left[Y_i^2 | X_i\right]\right]$$
$$= O\left(N^{-1-\beta_2+\beta_1}\right).$$

The second requirement is also satisfied as $\sum_{i=1}^{N_2} Var\left(\hat{W}_i^{ATE(m,l)} Y_i\right) = O\left(N^{-\beta_2+\beta_1}\right) < \infty$.

c) $Var(A_{1,N_2} + ... + A_{N_2,N_2}) = Var\left(\sum_{i=1}^{N_2} \hat{W}_{i,N}^{ATE(m,l)} Y_i - \sum_{i=1}^{N_2} E\left[\hat{W}_{i,N}^{ATE(m,l)} Y_i\right]\right) = Var\left(\sum_{i=1}^{N_2} \hat{W}_{i,N}^{ATE(m,l)} Y_i\right)$

$$\leq \sum_{i=1}^{N} Var\left(\hat{W}_{i,N}^{ATE(m,l)} Y_i\right) + \sum_{i=1}^{N}\sum_{j\neq i} \left|Cov\left(\hat{W}_{i,N}^{ATE(m,l)} Y_i, \hat{W}_{j,N}^{ATE(m,l)} Y_j\right)\right|$$

Using the results from b), the first element is bounded at rate $O\left(N^{-\beta_2+\beta_1}\right)$ and by a similar logic as in the tree case the sum of all covariances must decay to zero as fast as the sum of the variances. These results yield that $Var(A_{1,N_2} + ... + A_{N_2,N_2}) \xrightarrow[N\to\infty]{} 0$.

d) Due to monotonicity and result in b):

$$\sum_{i=1}^{N_2} E\left[A_{i,N_2}^2 \underline{1}\left(\left|A_{i,N_2}\right| > \tilde{\varepsilon}\right)\right] \leq \sum_{i=1}^{N_2} E\left[A_{i,N_2}^2\right] \to 0 \text{ for all } \tilde{\varepsilon} > 0.$$

e) The proof follows the same logic as in Lemma 4.

With this, all requirements for the CLT for triangular arrays of weakly dependent random variables hold, so that we get

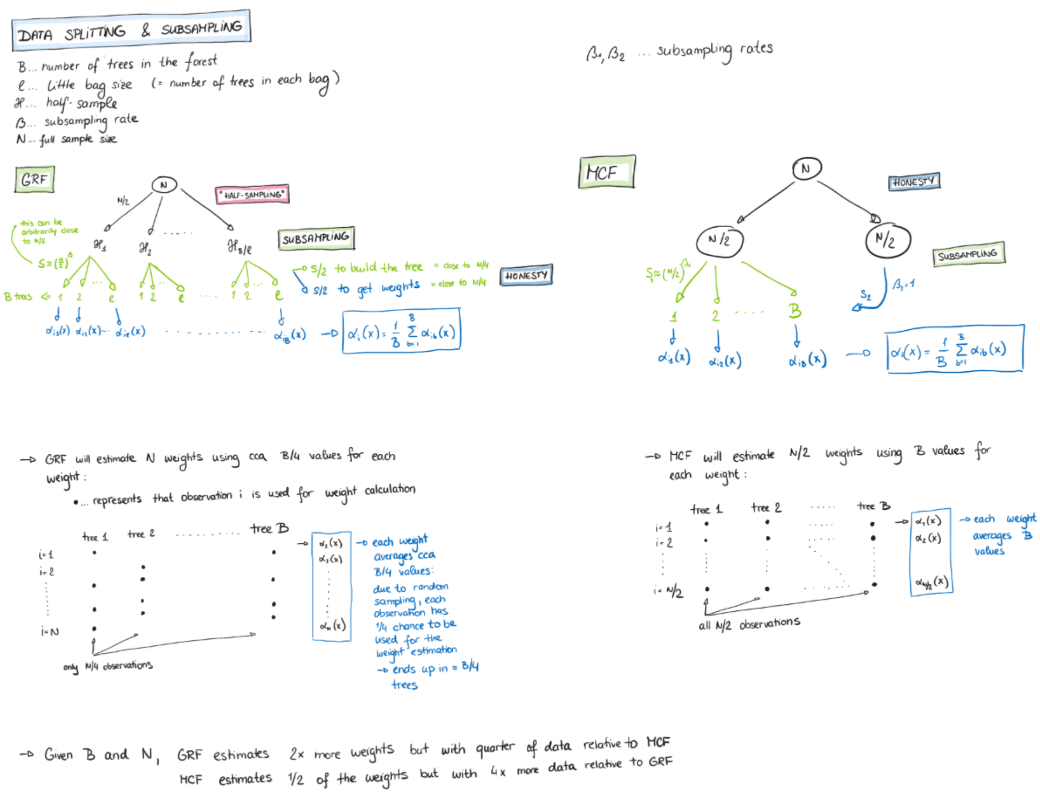
$$\frac{\widehat{ATE}(m,l) - ATE(m,l)}{\sqrt{Var(\widehat{ATE}(m,l))}} \to N(0,1).$$

Note that $E\left[\sum_{i=1}^{N_2} \hat{W}_i^{APO(d)} Y_i\right] = E\left[Y^d\right]$ due to exchangeability and the fact that the weights sum to 1. Therefore, we could have applied the CLT directly to the quantity of interest. The next corollary would follow a similar proof. ■

Appendix A.2: A comparison of *mcf* and *grf*

The Figure A.1 captures graphically the main differences between *mcf* and *grf* procedure regarding the one-sample and two-sample honesty and estimation of the weights.

Figure A.1: Diagram of grf and mcf



Appendix B: Additional details of the Monte Carlo study

Section B.1 explains the general simulation protocol used, while Section B.2 explains the data generating processes in detail. Section B.3 gives the details of the implementation of the various versions of the different estimators used.

B.1 Simulation protocol

Table B.1 shows the protocol employed in the Monte Carlo study.

Table B.1: Protocol of the Monte Carlo Study

1	Specify the data generating process with respect to (i) sample size, (ii) strength of selectivity into treatments, (iii) type of covariates, (iv) influence of covariates on non-treatment potential outcomes (including the degree of sparsity), (v) size of treatment effect and its heterogeneity, (vi) treatment share, and (vii) number of treatments
2	Draw training data of size *N*
3	Draw prediction data of (same) size *N*
4	Compute the true values of ATE and GATE and IATE on the prediction data
5	Train the different estimators on the training data
6	Predict ATE, GATEs, and IATEs on the prediction data
7	Repeat steps 2 to 6 *R* times (*R* = 1’000 * (2’500 / *sample size*))
8	Compute the performance measures
9	Repeat step 1 to 8 for different specifications

Note: The number of replications (*R*) declines such that the simulation noise remains approximately constant if the estimator is $\sqrt{N}$-convergent.

B.2 Data generating processes

The description of the data generating process covers the covariates, the selection (treatment) process, the outcome process, and the effects and their heterogeneity.

B.2.1 Covariates

The covariates are independent from each other and either normally or uniformly distributed,[36] both with mean zero and variance one.

[36] The reason for the uniform distribution is that the results for the *grf* and the *mcf* assume uniformly distributed covariates.

$$x_i^U \sim uniform(-\sqrt{12}/2, \sqrt{12}/2), \quad \dim(x_i^U) = p^U$$
$$x_i^N \sim N(0,1), \qquad \dim(x_i^N) = p^N$$
$$p = p^U + p^N, \quad \forall i = 1,...,N.$$

We also consider a scenario where the first 5 variables of X^N and X^U are generated as dummy variables, with $X^D = 2\underline{1}(X > 0) - 1$, also with mean zero.

In the simulations, we consider values of the following triples (p^N, p^U, p^D): (10, 10, 0), (5, 5, 0), (25, 25, 0), (20, 0, 0), (0, 20, 0), and (5, 5, 10). Thus, we capture cases of 10 to 50 covariates that may be of different types. (10, 10, 0) is the base specification.

B.2.2 Selection (treatment) process

We consider cases of two and four treatments where all treatments have equal shares. As an extension, in the binary treatment case there are also treatment shares of 25%. The treatment indicators are obtained from the quantiles of the following index function:

$$\breve{d}_i = \frac{\lambda \, x_i \beta}{\sqrt{\sum_{j=1}^{p} \beta_j^2 / 1.25}} + u_i, \quad \beta = (1, 1-1/k, ..., 1/k, 0, ..., 0), \quad \dim(\beta) = p, \quad \lambda \in [0, 0.42, 1.25],$$
$$u_i \sim N(0,1), \quad \forall i = 1,...,N.$$

The impact of the covariates declines with their order. The parameter λ determines the strength of the selection process, from a randomized experiment (λ=0) to the case of strong selection (λ=1.25 for uniform and mixed covariates; in case of normally distributed covariates grid for λ is 0, 0.45, and 1.5). In the intermediate setting, a machine learning estimation (using 1’000’000 observations) of the selection equation will obtain an out-of-sample R^2 of about 10%, while in the extreme case this value rises to about 42%. As it turns that the strength of selectivity is an important parameter when considering the performance of estimators, the base scenario considers all three selectivity levels. The baseline scenario in Appendix C.2 is based on the medium selectivity.

B.2.3 IATEs

The IATEs are deterministic functions of the covariates. All specifications of the IATEs are based on the following linear index, which is like the one used for the selection process:

$$\breve{\Delta}_i = \frac{\alpha\, x_i \gamma}{\sqrt{\sum_{j=1}^{p} \beta_j^2 / 1.25}}, \quad \gamma = (1, 1-1/k, ..., 1/k, 0, ..., 0), \quad \dim(\gamma) = p, \ \alpha \in [0, \alpha^{med}]$$

$$\forall i = 1, ..., N.$$

The value of α^{med} captures the strength of heterogeneity. Its exact value depends on the type of non-linearity and the R^2 of the potential non-treatment outcome process to ensure that the implications of the different effect strengths on the outcome remain stable.

The following variations of heterogeneity are considered for the IATE:

$$IATE_i = \breve{\Delta}_i + 1: \qquad linear$$
$$IATE_i = F^{\log istic}(\breve{\Delta}_i) + 0.5: \qquad nonlinear$$
$$IATE_i = \frac{\breve{\Delta}_i^2 - 1.25}{\sqrt{3}} + 1: \qquad quadratic$$
$$IATE_i = f^{WA}(\tilde{x}_{1,i}) f^{WA}(\tilde{x}_{2,i}) - 1.8: \quad step\ function\ (Wager - Athey)$$
$$F^{\text{logistic}}: c.d.f.\text{ of logistic distribution}, \quad f^{WA}(x) = 1 + \frac{1}{1 + \mathrm{e}^{-20(x-1/3)}}, \quad \forall i = 1, ..., N.$$

They represent a linear function and three non-linear functions. Note that the last function is very similar to the DGP used in Wager and Athey (2018). For the latter, the heterogeneity depends only on two variables (which are the ones that are most important for the selection and the outcome processes). $\tilde{x}_{1,i}$ and $\tilde{x}_{2,i}$ are transformed versions of $x_{1,i}$ and $x_{2,i}$, where the exact transformation used depends on the distribution of each variable. The baseline scenario in Appendix C.2 is based on the step function only.

The resulting ATE in these cases is always one, independent of the heterogeneity specification. In addition to these cases, we also consider a case of the IATEs all being equal to zero.

B.2.4 Outcome processes

The outcome process consists first of specifying a process for the non-treatment potential outcome. The potential outcome with treatment is obtained by adding the IATE plus noise to the non-treatment outcome.

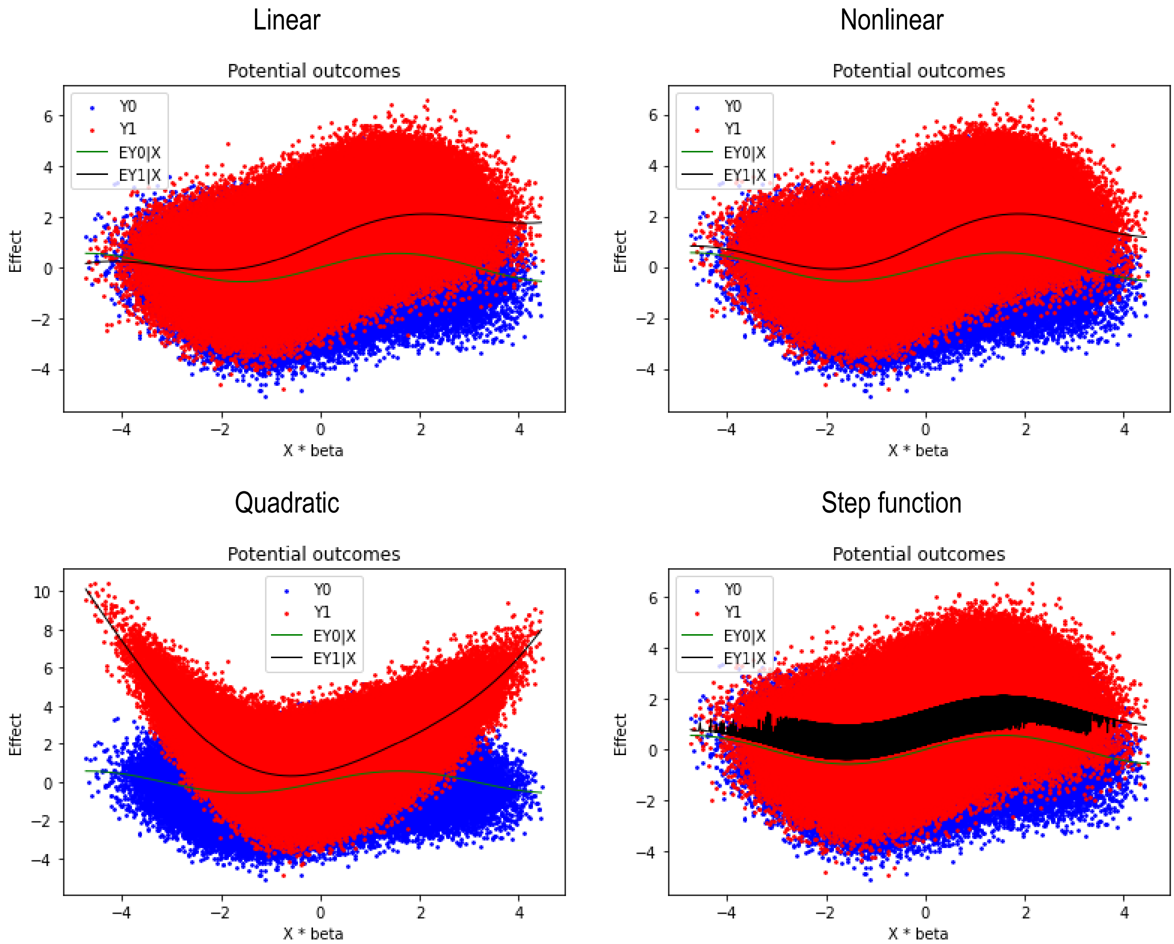
The following non-linear outcome process is specified for the potential non-treatment outcome:

$$\breve{y}_i = \frac{x_i\beta}{\sqrt{\sum_{j=1}^{p}\beta_j^2 / 1.25}}, \quad \beta = (1, 1-1/k, ..., 1/k, 0, ..., 0), \quad \dim(\beta) = p,$$
$$y_i^0 = \delta \sin(\breve{y}_i) + \varepsilon_i^0, \quad \delta \in [0, \delta^{med}, \delta^{strong}], \quad \varepsilon_i^0 \sim N(0,1),$$
$$y_i^1 = y_i^0 + IATE_i + \varepsilon_i^1, \qquad \varepsilon_i^1 \sim N(0,1), \quad \forall i = 1, ..., N.$$

δ^{med} and δ^{strong} are chosen such that the R^2 in the outcome process of the potential non-treatment outcome is about 10% (base specification) and 45%.

Figure B.1 shows the relation of the potential outcomes and their expectations with the index $x_i\beta$ for the case of p^N=p^U=10, p=20, k^N=k^U=5, k=10, and δ^{med} with respect to the different heterogeneities. Figure B.2 shows the same relationship for the stronger effect size δ^{strong}.

Figure B.1: Shape of potential outcomes for different shapes of IATEs (δ^{med})

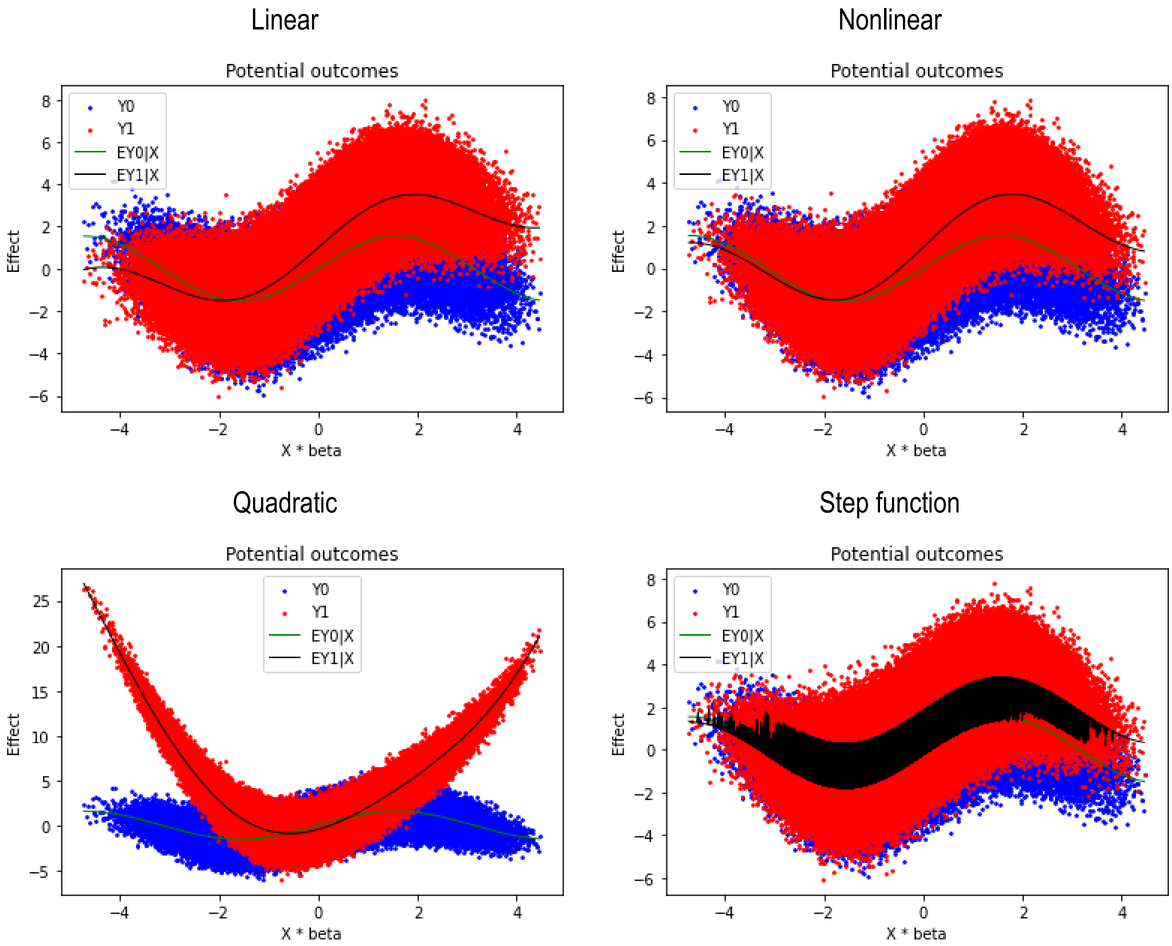


Note: Figures are based on 1'000'000 observations.

Figures B.1 show that the differences between the linear and the non-linear case are rather small, at least compared to the quadratic IATEs and those based on the step function. In particular the latter, for which the IATEs depend only on the first two most important covariates, show behaviour that is not being well approximated by any parsimonious parametric function of the linear index. Thus, we conjecture that this type of effect heterogeneity will be most difficult to estimate well.

Increasing the predictiveness of the covariates for the non-treatment potential outcome (Figure B.2), denoted by the blue dots and the green line, leads to substantially more pronounced non-linearity with respect to the linear index. This is true for all specifications of the IATEs, but particularly so for the quadratic one.

Figure B.2: Shape of potential outcomes for different shapes of IATEs (δ^{strong})



Note: Figures are based on 1’000’000 observations.

B.2.5 Overview of simulations and outcome tables

Table B.2 gives an overview where to find the simulations results for the different data generating processes in Appendix C.

Table B.2: Overview of specifications and their locations

Table	*Hetero-geneity*	Selec-tivity	*k*	*p*	R^2 (y^0)	*X*	*N*	No of treat-ments	Treat-ment share	GATE groups
C.1	none	none	10	20	10%	U, N	stand	2	50%	5
C.2	none	middle	10	20	10%	U, N	stand	2	50%	5
C.3	none	strong	10	20	10%	U, N	stand	2	50%	5
C.4	**linear**	None	10	20	10%	U, N	stand	2	50%	5
C.5	**linear**	middle	10	20	10%	U, N	stand	2	50%	5
C.6	**linear**	strong	10	20	10%	U, N	stand	2	50%	5
C.7	**nonlin**	None	10	20	10%	U, N	stand	2	50%	5
C.8	**nonlin**	middle	10	20	10%	U, N	stand	2	50%	5
C.9	**nonlin**	strong	10	20	10%	U, N	stand	2	50%	5
C.10	**quadrat**	none	10	20	10%	U, N	stand	2	50%	5
C.11	**quadrat**	middle	10	20	10%	U, N	stand	2	50%	5
C.12	**quadrat**	strong	10	20	10%	U, N	stand	2	50%	5
C.13	step	none	10	20	10%	U, N	stand	2	50%	5
C.14	step	middle	10	20	10%	U, N	stand	2	50%	5
C.15	step	strong	10	20	10%	U, N	stand	2	50%	5
C.16	step	middle	**5**	**10**	10%	U, N	stand	2	50%	5
C.17	step	middle	**25**	**50**	10%	U, N	stand	2	50%	5
C.18	step	middle	**4**	20	10%	U, N	stand	2	50%	5
C.19	step	middle	**16**	20	10%	U, N	stand	2	50%	5
C.20	step	middle	10	20	**0%**	U, N	stand	2	50%	5
C.21	step	middle	10	20	**45%**	U, N	stand	2	50%	5
C.22	step	middle	10	20	10%	**U**	stand	2	50%	5
C.23	step	middle	10	20	10%	**N**	stand	2	50%	5
C.24	step	middle	10	20	10%	**U, N, D**	stand	2	50%	5
C.25	step	middle	10	20	10%	U, N	stand	**4**	50%	5
C.26	step	middle	10	20	10%	U, N	**40’000**	2	50%	5
C.27	step	middle	10	20	10%	U, N	stand	2	**25%**	5
C.28	step	none	10	20	10%	U, N	stand	2	50%	**5-40**
C.29	step	middle	10	20	10%	U, N	stand	2	50%	**5-40**
C.30	step	strong	10	20	10%	U, N	stand	2	50%	**5-40**

Note: Deviations from the benchmark specifications are in bold font. *N stand* implies that *N* = 2’500 and *N* = 10’000 are in the same table. *X U,N, D* denotes uniformly and normally distributed covariates, as well as dummy variables.

B.3 Implementation of the estimators

To avoid additional computational costs in an already computationally very expensive simulation study, none of the estimators has been explicitly tuned. Either default values are used for the tuning parameters, or tuning parameters have been set to a fixed value beforehand. In an empirical application, the quality of estimation could be improved by hyperparameter tuning. For more details on hyperparameter tuning for *dml*, see Bach, Schacht, Chernozhukov, Klaassen, and Spindler (2024). Forest estimators often minimize out-of-bag error to find optimal hyperparameters. For details on the tuning procedures, please refer to the documentation of *grf* and *mcf*.

B.3.1 MCF

The results of the *mcf* are computed with the Python version of the package which is available on PyPI. Most of the simulations have been performed with version 0.3.3., but some of the latest simulations used also the newer and faster version 0.4.3 (if true, this is indicated in the respective table). Although some default values changed a bit between the versions, results were very similar. In the following, sample A denotes one half of the data set used to build the forest (training data) and sample B denotes the other half of the sample used to estimate the effects (honest data).

Contrary to the default values, the minimum leaf size in each tree equals 5. The number of trees contained in any forest equals 1000. Trees are formed on random subsamples drawn without replacement (subsampling) with a sample size of 50% of the size of sample A. Regarding estimation of MCE and finding close neighbours, closeness is based on a Random Forest based prognostic score (see Hansen, 2008), $[\hat{\mu}_0(x),....,\hat{\mu}_{M-1}(x)]$ weighted by the Mahalanobis distance, as forming the neighbours by simplified Mahalanobis matching suffers in large-dimensional problems.

Two variants of this estimator are used, with and without local centring. Local centring is implemented in the *mcf* package by running a regression random forest (using the Python package scikit-learn) to predict *E(Y|X)* (*X* does not include the treatment). These (out-of-training-sample) predictions are obtained in a 5-fold cross-fitting scheme. The predictions are subtracted from the observed *Y* to obtain a centred outcome, Y^{cent}. Y^{cent} instead of *Y* is subsequently used as outcome in the *mcf* algorithm. The simulations show that this centred version outperforms the uncentred version when there is non-random selection. Recentring is implemented in the following way:

1) Estimate the trees that define the Random Forest for $E(Y \mid X = x)$ in sample A.

2) Recentring of outcomes in sample A: Split sample A randomly into K equally sized parts, $A\text{-}1$ to $A\text{-}K$. Use the outcomes in the union of the $K\text{-}1$-folds $A\text{-}1$ to $A\text{-}(K\text{-}1)$ to obtain the Random Forest predictions given the forest estimated in step 1). Use these predictions to predict $E(Y \mid X = x)$ in fold $A\text{-}K$. Do this for all-possible combinations of folds (cross-fitting as used in k-fold cross-fitting). Subtract the predictions from the actual values of Y.
3) Redo step 2 in sample B using the estimated forests of sample A.

Concerning the specifics of the local centring algorithm, there are a couple of points worth mentioning. First, to avoid overfitting, the outcomes of observation 'i' are not used to predict itself. Therefore, the chosen implementation is based on cross-fitting.

Second, weights-based inference requires avoiding a dependence of the weights in sample B on outcomes of sample A. However, since recentring uses outcome variables independent of the treatment state, this could induce a correlation between the recentred outcomes in different treatment states. This finite sample correlation will be ignored here (as in Athey, Tibshirani, and Wager, 2019).

Third, the number of folds is a tuning parameter that influences the noise added to the recentred outcome by subtracting an estimated quantity. The simulation results indicate that the computationally most attractive choice of $K=2$ may be too small in medium sized samples and that a somewhat larger number of folds may be needed to avoid much additional noise to the estimators.

The nonparametric regressions that enter the estimation of the standard errors are based on k-NN estimation with number of neighbours equal to $2\ sqrt(N)$.

If inference is not of interest, a more efficient *mcf* estimator can be obtained by switching the roles of the samples used for building the forest and populating the leaves with outcome values, and subsequently averaging the two estimates.

B.3.2 GRF

The reported results are based on the R package *grf* (version 2.3.1). The default forest parameters are 2000 trees (twice more than in *mcf)*, minimum node size is 5 and number of variables tested for a split is a random draw from a Poisson distribution with parameter *min(p, round_up((p+20)^(1/2)))*.

The default implementation of the *grf* uses out-of-bag predictions of nuisance parameters to remove confounding effects in order to find the best splits on which the IATEs are estimated. As this implementation is not in line with the theory in Athey et al. (2019), a centred version of the algorithm, using *K*-fold cross-fitting to calculate the residualized outcome that is later fed as an input to the *grf*, is added to the simulation.

For later estimation of GATEs and ATEs, so-called DR scores need to be constructed. GATE and ATE can be estimated directly via functions *average_treatment_effect()* and *best_linear_predictor()*. Both functions need the estimated forest as an input to estimate the effects on the data used for building the forest and estimating the forest weights. Here, the simulation deviates from the protocol, as currently *grf* does not provide an option to estimate the DR scores on a new (prediction) data set. The package provides function *get_scores()* which computes the estimated component of a DR score on the data set that was used for the forest if needed. The current implementation of GATE estimation avoids the *best_linear_predictor()* and regresses the scores from *get_score()* on group dummies exploiting the homoscedastic error terms for estimation of standard errors.

B.3.3 DML

The key element of all *dml* estimators is the DR component of the *dml* score. For treatments *m* and *l*, this DR component is defined as:

$$\Gamma_{m,l}^{dml}(X,Y,D;\eta(X)) = \mu_m(X) - \mu_l(X) + \frac{\underline{1}(D=m)(Y-\mu_m(X))}{p_m(X)} - \frac{\underline{1}(D=l)(Y-\mu_l(X))}{p_l(X)}$$

$$\mu_d(x) = E\left(Y \middle| D = d, X = x\right), \quad p_d(x) = P(D = d \middle| X = x)$$

The nuisance parameters, $\eta(X) = (\mu_m(X), \mu_l(X), p_m(X), p_l(X))$ are estimated with regression forests. As usual, the necessary cross-fitting is implemented via 5-fold cross-fitting. To obtain effects, the estimated DR components of the score for the desired treatment contrasts are formed. The ATE is obtained by averaging these differences.

GATEs and IATEs are obtained by regression-type approaches in which the estimated DR component of the score serves as dependent variable. The GATEs are computed as OLS-coefficients of a (saturated) regression model with the indicators for the GATEs groups acting as independent variables. IATEs are estimated by using X as independent variables either in a regression random forest[37] or in an OLS estimation. When OLS regressions are used, inference is based on heteroscedasticity-robust covariance matrix of the corresponding coefficients. No inference is obtained for the random forest based IATEs.

As the weights may lead to small sample issues, in particular when selection probabilities get close to zero, two versions of the *dml* estimator are considered. The first one is taking the weights as they are, while the second version normalises to sum of the weights to one (N)[38] and truncates 'too large' weights. Too large here means that a single weight is larger than 5% of the sum of all weights (if so, it is truncated at 5%). Since the normalised versions appears to outperform the non-normalized one in many simulations, the latter is presented in the main body of the text.

37 For methods prone to overfitting, an additional data split might be beneficial that enables training of the model and estimation of IATE to be done on different parts of the data set as recommended by Knaus (2022). This step was not implemented for comparability of the estimates across methods.

38 Similarly, as in the normalized DR learner in Knaus (2022).

B.3.4 OLS

As a benchmark estimator, a two-sample (for binary treatments) OLS estimator is implemented in a standard way. However, since there are substantial non-linearities in the DGP, it is not surprising that there are many scenarios in which this estimator performs badly. Therefore, OLS results are not reported in the main part of the paper.

Appendix C: Detailed results of the Monte Carlo study

C.1 Base specifications

In this section, we vary the selectivity and the type of the IATE jointly and keep the other parameters fixed at their base values (i.e., k=10, p=20, $R^2(y^0)$ = 10%, X^N and X^U, binary treatment, treatment share 50%, N=2'500 and N =10'000).

Table C.1: No IATE, no selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurto-sis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	-0.003	0.048	0.060	0.060	0.01	-0.05	0.000	96	79
GATE	*5*		-0.003	0.052	0.065	0.065	0.01	0.05	0.002	96	81
IATE	*N*		-0.003	0.079	-	0.100	-	-	-	97	85
IATE eff	*N*		-0.001	0.056	-	0.070	-	-	-	-	-
ATE	*1*	*mcf*	0.001	0.048	0.061	0.061	0.03	0.07	-0.002	94	77
GATE	*5*	*cent*	0.001	0.051	0.065	0.065	0.03	0.13	-0.001	95	78
IATE	*N*		0.001	0.080	-	0.100	-	-	-	96	82
IATE eff	*N*		0.001	0.056	-	0.070	-	-	-	-	-
ATE	*1*	*grf*	-0.001	0.034	0.042	0.042	-0.14	-0.01	-0.000	95	82
GATE	*5*		-0.001	0.075	0.094	0.094	0.04	0.07	-0.001	95	80
IATE	*N*		-0.001	0.048	-	0.060	-	-	-	99	92
ATE	*1*	*grf*	0.003	0.032	0.041	0.041	0.12	0.01	0.001	95	82
GATE	*5*	*cent*	0.003	0.074	0.093	0.093	0.03	-0.00	-0.000	95	80
IATE	*N*		0.003	0.047	-	0.059	-	-	-	99	92
ATE	*1*	*dml*	-0.002	0.033	0.042	0.042	0.07	-0.07	0.005	98	84
GATE	*5*		-0.002	0.074	0.093	0.093	-0.02	0.04	0.001	95	81
IATE	*N*	*ols*	-0.002	0.167	-	0.210	-	-	-	32	21
IATE	*N*	*rf*	-0.002	0.221	-	0.281	-	-	-	-	-
ATE	*1*	*dml-*	-0.002	0.033	0.042	0.042	0.08	-0.07	0.004	97	84
GATE	*5*	*norm*	-0.002	0.073	0.093	0.093	-0.02	0.06	0.000	95	80
IATE	*N*	*ols*	-0.002	0.166	-	0.209	-	-	-	31	21
IATE	*N*	*rf*	-0.002	0.218	-	0.278	-	-	-	-	-
ATE	*1*	*ols*	-0.001	0.033	0.041	0.041	0.08	-0.05	-0.012	83	62
GATE	*5*		-0.001	0.072	0.091	0.091	-0.03	0.10	-0.026	84	64
IATE	*N*		-0.001	0.163	-	0.205	-	-	-	83	64

Note: Table to be continued.

Table C.1 - continued: No IATE, no selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurto-sis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.001	0.022	0.028	0.028	0.12	0.03	0.003	96	82
GATE	*5*		0.001	0.025	0.031	0.031	0.09	-0.30	0.002	97	84
IATE	*N*		0.001	0.049	-	0.062	-	-	-	98	87
IATE eff	*N*		0.000	0.035	-	0.043	-	-	-	-	-
ATE	*1*	*mcf*	0.003	0.022	0.027	0.028	0.14	-0.12	0.002	97	80
GATE	*5*	*cent*	0.003	0.020	0.030	0.030	0.11	-0.21	0.002	96	81
IATE	*N*		0.003	0.048	-	0.060	-	-	-	97	84
IATE eff	*N*		0.002	0.034	-	0.042	-	-	-	-	-
ATE	*1*	*grf*	-0.001	0.015	0.020	0.020	-0.11	0.57	0.001	96	81
GATE	*5*		-0.001	0.037	0.047	0.047	-0.13	0.15	-0.001	94	80
IATE	*N*		-0.001	0.030	-	0.038	-	-	-	99.8	98
ATE	*1*	*grf*	0.002	0.016	0.020	0.020	0.02	0.13	0.000	96	79
GATE	*5*	*cent*	0.002	0.036	0.045	0.045	-0.10	-0.12	0.000	95	79
IATE	*N*		0.002	0.030	-	0.038	-	-	-	99.9	98
ATE	*1*	*dml*	-0.001	0.015	0.019	0.019	-0.11	-0.09	0.004	98	87
GATE	*5*		-0.001	0.036	0.045	0.045	0.04	0.03	0.001	95	81
IATE	*N*	*ols*	-0.001	0.080	-	0.100	-	-	-	16	11
IATE	*N*	*rf*	-0.001	0.153	-	0.195	-	-	-	-	-
ATE	*1*	*dml-*	-0.001	0.015	0.019	0.019	-0.11	-0.08	0.004	98	87
GATE	*5*	*norm*	-0.001	0.036	0.045	0.045	0.04	0.02	0.001	95	81
IATE	*N*	*ols*	-0.001	0.079	-	0.100	-	-	-	16	11
IATE	*N*	*rf*	-0.001	0.152	-	0.194	-	-	-	-	-
ATE	*1*	*ols*	0.000	0.015	0.019	0.019	-0.17	-0.04	-0.005	86	66
GATE	*5*		0.000	0.036	0.045	0.045	0.01	-0.03	-0.013	85	63
IATE	*N*		0.035	0.079	-	0.100	-	-	-	84	64

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 500 replications used for 2'500 / 10'000 obs.

Table C.2: No IATE, medium selectivity

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.160	0.160	0.061	0.171	0.12	0.10	0.000	25	9
GATE	*5*		0.160	0.161	0.073	0.176	0.09	0.10	0.003	42	19
IATE	*N*		0.160	0.167	-	0.194	-	-	-	77	48
IATE eff	*N*		0.163	0.164	-	0.164	-	-	-	-	-
ATE	*1*	*mcf*	0.022	0.051	0.060	0.064	0.14	0.31	-0.001	94	76
GATE	*5*	*cent*	0.022	0.060	0.073	0.076	0.13	0.27	0.001	95	79
IATE	*N*		0.022	0.086	-	0.109	-	-	-	96	82
IATE eff	*N*		0.025	0.063	-	0.079	-	-	-	-	-
ATE	*1*	*grf*	0.091	0.092	0.043	0.101	-0.03	-0.23	-0.002	41	19
GATE	*5*		0.091	0.108	0.093	0.131	0.02	-0.00	-0.001	83	60
IATE	*N*		0.090	0.094	-	0.109	-	-	-	88	66
ATE	*1*	*grf*	0.027	0.040	0.041	0.049	0.02	-0.25	-0.000	90	71
GATE	*5*	*cent*	0.027	0.076	0.091	0.095	0.01	0.11	0.001	94	78
IATE	*N*		0.028	0.053	-	0.066	-	-	-	98	90
ATE	*1*	*dml*	0.016	0.038	0.044	0.048	0.00	-0.23	0.004	96	81
GATE	*5*		0.016	0.079	0.097	0.099	-0.05	-0.05	0.001	95	80
IATE	*N*	*ols*	0.017	0.176	-	0.221	-	-	-	33	22
IATE	*N*	*rf*	0.009	0.232	-	0.297	-	-	-	-	-
ATE	*1*	*dml-*	0.017	0.038	0.044	0.047	-0.01	-0.26	0.004	96	79
GATE	*5*	*norm*	0.017	0.079	0.097	0.099	-0.05	-0.05	0.001	95	79
IATE	*N*	*ols*	0.017	0.175	-	0.222	-	-	-	32	22
IATE	*N*	*rf*	0.010	0.231	-	0.299	-	-	-	-	-
ATE	*1*	*ols*	0.007	0.035	0.043	0.044	-0.01	-0.19	-0.012	84	62
GATE	*5*		0.007	0.077	0.093	0.096	-0.06	-0.01	-0.026	83	61
IATE	*N*		0.007	0.165	-	0.208	-	-	-	83	63

Note: Table to be continued.

Table C.2- continued: No IATE, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.138	0.138	0.028	0.141	0.15	-0.02	0.002	0	0
GATE	*5*		0.138	0.138	0.037	0.143	0.12	-0.14	0.001	11	4
IATE	*N*		0.138	0.140	-	0.156	-	-	-	62	33
IATE eff	*N*		0.138	0.138	-	0.148	-	-	-	-	-
ATE	*1*	*mcf*	0.001	0.021	0.027	0.027	0.27	0.03	0.003	97	84
GATE	*5*	*cent*	0.001	0.031	0.038	0.039	0.21	0.04	0.003	97	84
IATE	*N*		0.001	0.056	-	0.071	-	-	-	97	84
IATE eff	*N*		0.001	0.042	-	0.053	-	-	-	-	-
ATE	*1*	*grf*	0.059	0.059	0.020	0.062	-0.23	-0.20	0.000	16	7
GATE	*5*		0.059	0.064	0.046	0.075	-0.07	-0.14	-0.000	74	48
IATE	*N*		0.058	0.061	-	0.070	-	-	-	98	85
ATE	*1*	*grf*	0.011	0.019	0.021	0.024	-0.13	0.21	-0.001	90	74
GATE	*5*	*cent*	0.011	0.038	0.046	0.048	-0.05	-0.10	-0.001	94	76
IATE	*N*		0.013	0.033	-	0.041	-	-	-	99.8	97
ATE	*1*	*dml*	0.008	0.014	0.023	0.023	-0.02	0.31	0.003	96	82
GATE	*5*		0.009	0.040	0.049	0.050	0.12	0.05	0.000	94	80
IATE	*N*	*ols*	0.009	0.086	-	0.108	-	-	-	17	11
IATE	*N*	*rf*	0.004	0.161	-	0.207	-	-	-	-	-
ATE	*1*	*dml-*	0.009	0.018	0.021	0.023	-0.03	0.32	0.003	96	82
GATE	*5*	*norm*	0.009	0.040	0.049	0.050	0.12	0.04	0.000	94	80
IATE	*N*	*ols*	0.009	0.086	-	0.108	-	-	-	17	11
IATE	*N*	*rf*	0.004	0.160	-	0.207	-	-	-	-	-
ATE	*1*	*ols*	0.007	0.017	0.020	0.022	0.02	0.19	-0.005	84	64
GATE	*5*		0.007	0.038	0.046	0.053	0.09	0.13	-0.013	78	55
IATE	*N*		0.007	0.083	-	0.104	-	-	-	83	63

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.3: No IATE, strong selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.375	0.375	0.063	0.380	0.03	-0.03	0.000	0	0
GATE	*5*		0.375	0.375	0.082	0.384	0.02	0.17	0.010	3	1
IATE	*N*		0.375	0.375	-	0.394	-	-	-	21	6
IATE eff	*N*		0.378	0.378	-	0.388	-	-	-	-	-
ATE	*1*	*mcf*	0.011	0.045	0.056	0.057	0.03	-0.13	0.008	97	85
GATE	*5*	*cent*	0.011	0.060	0.076	0.077	0.04	0.11	0.013	98	86
IATE	*N*		0.011	0.091	-	0.115	-	-	-	97	86
IATE eff	*N*		0.013	0.066	-	0.084	-	-	-	-	-
ATE	*1*	*grf*	0.264	0.264	0.049	0.268	-0.07	-0.03	-0.010	0	0
GATE	*5*		0.264	0.264	0.093	0.280	-0.06	-0.00	-0.005	17	6
IATE	*N*		0.256	0.256	-	0.265	-	-	-	32	8
ATE	*1*	*grf*	0.068	0.070	0.045	0.081	0.03	-0.06	-0.006	58	35
GATE	*5*	*cent*	0.068	0.090	0.089	0.112	0.02	0.14	-0.001	87	67
IATE	*N*		0.076	0.084	-	0.101	-	-	-	94	78
ATE	*1*	*dml*	0.136	0.137	0.058	0.148	-0.32	1.36	-0.003	32	14
GATE	*5*		0.136	0.155	0.124	0.185	-0.67	5.35	-0.014	69	48
IATE	*N*	*ols*	0.136	0.235	-	0.297	-	-	-	30	20
IATE	*N*	*rf*	0.114	0.290	-	0.480	-	-	-	-	-
ATE	*1*	*dml-*	0.129	0.129	0.055	0.140	-0.03	0.01	0.002	38	16
GATE	*5*	*norm*	0.129	0.148	0.121	0.177	-0.13	0.02	-0.005	76	54
IATE	*N*	*ols*	0.129	0.177	-	0.292	-	-	-	33	22
IATE	*N*	*rf*	0.129	0.296	-	0.401	-	-	-	-	-
ATE	*1*	*ols*	0.109	0.110	0.053	0.121	0.00	0.04	-0.014	25	12
GATE	*5*		0.109	0.117	0.102	0.152	-0.03	0.01	-0.029	60	41
IATE	*N*		0.109	0.192	-	0.240	-	-	-	77	56

Note: Table to be continued.

Table C.3 - continued: No IATE, strong selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.334	0.334	0.030	0.336	-0.03	0.60	0.001	0	0
GATE	*5*		0.334	0.334	0.045	0.337	-0.07	0.08	0.006	0	0
IATE	*N*		0.334	0.334	-	0.345	-	-	-	5	1
IATE eff	*N*		0.335	0.335	-	0.340	-	-	-	-	-
ATE	*1*	*mcf*	-0.041	0.043	0.027	0.049	0.17	0.38	0.006	80	48
GATE	*5*	*cent*	-0.041	0.048	0.041	0.058	-0.07	0.06	0.008	90	70
IATE	*N*		-0.041	0.073	-	0.091	-	-	-	94	69
IATE eff	*N*		-0.041	0.060	-	0.074	-	-	-	-	-
ATE	*1*	*grf*	0.192	0.192	0.024	0.194	-0.37	0.18	-0.004	0	0
GATE	*5*		0.192	0.192	0.047	0.198	-0.00	-0.01	-0.002	2	0
IATE	*N*		0.187	0.187	-	0.193	-	-	-	53	18
ATE	*1*	*grf*	0.027	0.031	0.025	0.037	-0.08	-0.07	-0.005	68	45
GATE	*5*	*cent*	0.027	0.044	0.047	0.055	-0.09	0.06	-0.003	90	69
IATE	*N*		0.039	0.049	-	0.061	-	-	-	99	94
ATE	*1*	*dml*	0.079	0.079	0.034	0.086	0.034	0.54	3.01	27	11
GATE	*5*		0.079	0.091	0.076	0.111	0.076	0.20	6.61	66	46
IATE	*N*	*ols*	0.079	0.133	-	0.171	-	-	-	16	11
IATE	*N*	*rf*	0.061	0.208	-	0.411	-	-	-	-	-
ATE	*1*	*dml-*	0.070	0.070	0.031	0.077	0.000	0.15	0.33	37	18
GATE	*5*	*norm*	0.070	0.084	0.071	0.101	-0.007	-0.07	0.13	75	54
IATE	*N*	*ols*	0.070	0.129	-	0.162	-	-	-	18	12
IATE	*N*	*rf*	0.051	0.213	-	0.336	-	-	-	-	-
ATE	*1*	*ols*	0.111	0.111	0.026	0.111	0.10	-0.32	-0.007	0	0
GATE	*5*		0.111	0.113	0.051	0.126	0.04	-0.10	-0.015	26	14
IATE	*N*		0.111	0.129	-	0.156	-	-	-	61	41

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 observations.

Table C.4: Linear IATE, no selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	-0.004	0.049	0.062	0.062	-0.02	-0.06	-0.021	96	80
GATE	*5*		-0.004	0.091	0.078	0.114	0.00	0.19	0.001	83	61
IATE	*N*		-0.004	0.200	-	0.252	-	-	-	69	49
IATE eff	*N*		-0.004	0.193	-	0.242	-	-	-	-	-
ATE	*1*	*mcf*	0.002	0.049	0.062	0.062	0.01	0.08	-0.001	95	78
GATE	*5*	*cent*	0.001	0.095	0.080	0.118	0.01	0.22	-0.008	75	55
IATE	*N*		0.002	0.208	-	0.261	-	-	-	62	43
IATE eff	*N*		0.004	0.201	-	0.252	-	-	-	-	-
ATE	*1*	*grf*	-0.001	0.034	0.043	0.043	-0.15	0.01	0.000	95	82
GATE	*5*		-0.001	0.077	0.097	0.097	0.03	0.07	-0.001	95	80
IATE	*N*		-0.001	0.214	-	0.268	-	-	-	57	40
ATE	*1*	*grf*	0.004	0.033	0.042	0.042	0.09	0.02	0.001	94	81
GATE	*5*	*cent*	0.004	0.075	0.094	0.094	0.04	0.04	0.000	95	80
IATE	*N*		0.004	0.212	-	0.266	-	-	-	56	40
ATE	*1*	*dml*	-0.002	0.034	0.042	0.042	0.06	-0.07	0.005	98	84
GATE	*5*		-0.002	0.075	0.095	0.095	-0.02	0.04	0.001	95	81
IATE	*N*	*ols*	-0.002	0.168	-	0.212	-	-	-	34	22
IATE	*N*	*rf*	-0.002	0.245	-	0.310	-	-	-	-	-
ATE	*1*	*dml-*	-0.002	0.034	0.042	0.042	0.06	-0.07	0.005	97	83
GATE	*5*	*norm*	-0.002	0.075	0.095	0.095	-0.02	0.06	0.000	95	80
IATE	*N*	*ols*	-0.002	0.167	-	0.210	-	-	-	32	21
IATE	*N*	*rf*	-0.002	0.244	-	0.308	-	-	-	-	-
ATE	*1*	*ols*	-0.001	0.033	0.042	0.042	0.01	-0.08	-0.013	81	62
GATE	*5*		-0.002	0.073	0.092	0.092	-0.04	0.09	-0.027	84	64
IATE	*N*		-0.001	0.163	-	0.205	-	-	-	83	63

Note: Table to be continued.

Table C.4 - continued: Linear IATE, no selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.001	0.023	0.029	0.029	0.108	-0.20	0.004	98	82
GATE	*5*		0.000	0.066	0.042	0.080	-0.04	-0.27	0.000	67	46
IATE	*N*		0.001	0.162	-	0.205	-	-	-	61	43
IATE eff	*N*		0.000	0.158	-	0.201	-	-	-	-	-
ATE	*1*	*mcf*	0.003	0.022	0.028	0.028	0.13	-0.15	0.002	97	79
GATE	*5*	*cent*	0.003	0.062	0.047	0.077	0.01	-0.25	-0.008	65	46
IATE	*N*		0.003	0.171	-	0.216	-	-	-	52	36
IATE eff	*N*		0.002	0.168	-	0.212	-	-	-	-	-
ATE	*1*	*grf*	-0.000	0.015	0.020	0.020	-0.11	0.51	0.001	96	82
GATE	*5*		-0.001	0.037	0.047	0.047	-0.15	0.18	-0.000	94	81
IATE	*N*		-0.001	0.175	-	0.221	-	-	-	61	44
ATE	*1*	*grf*	0.003	0.016	0.020	0.021	0.02	0.15	-0.000	96	82
GATE	*5*	*cent*	0.003	0.036	0.045	0.045	-0.07	-0.10	0.001	95	80
IATE	*N*		0.003	0.174	-	0.219	-	-	-	61	44
ATE	*1*	*dml*	-0.001	0.016	0.020	0.020	-0.10	0.00	0.004	98	86
GATE	*5*		-0.001	0.037	0.046	0.046	0.02	0.01	0.001	95	82
IATE	*N*	*ols*	-0.001	0.080	-	0.101	-	-	-	16	11
IATE	*N*	*rf*	-0.001	0.187	-	0.236	-	-	-	-	-
ATE	*1*	*dml-*	-0.001	0.016	0.020	0.020	-0.11	-0.01	0.003	98	86
GATE	*5*	*norm*	-0.001	0.037	0.046	0.046	0.02	0.01	0.001	95	82
IATE	*N*	*ols*	-0.001	0.080	-	0.101	-	-	-	17	11
IATE	*N*	*rf*	-0.001	0.186	-	0.235	-	-	-	-	-
ATE	*1*	*ols*	-0.000	0.016	0.020	0.020	-0.16	0.01	-0.005	84	65
GATE	*5*		-0.001	0.037	0.045	0.046	-0.01	0.01	-0.013	84	64
IATE	*N*		0.000	0.079	-	0.100	-	-	-	84	64

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs.

Table C.5: Linear IATE, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.225	0.225	0.063	0.234	0.11	0.11	0.001	6	1
GATE	*5*		0.225	0.227	0.081	0.252	0.05	0.09	0.001	29	16
IATE	*N*		0.225	0.280	-	0.340	-	-	-	53	36
IATE eff	*N*		0.228	0.276	-	0.333	-	-	-	-	-
ATE	*1*	*mcf*	0.056	0.067	0.061	0.082	0.09	0.28	-0.001	84	63
GATE	*5*	*cent*	0.055	0.102	0.080	0.126	0.06	0.17	-0.003	76	54
IATE	*N*		0.056	0.215	-	0.269	-	-	-	62	43
IATE eff	*N*		0.059	0.208	-	0.260	-	-	-	-	-
ATE	*1*	*grf*	0.128	0.128	0.044	0.135	-0.04	-0.23	-0.002	16	5
GATE	*5*		0.128	0.137	0.096	0.161	0.01	-0.02	-0.002	72	47
IATE	*N*		0.122	0.248	-	0.309	-	-	-	49	33
ATE	*1*	*grf*	0.050	0.055	0.042	0.065	0.02	-0.18	-0.000	77	53
GATE	*5*	*cent*	0.050	0.086	0.093	0.108	0.01	0.08	0.000	91	72
IATE	*N*		0.049	0.230	-	0.287	-	-	-	51	36
ATE	*1*	*dml*	0.030	0.044	0.045	0.054	-0.02	-0.23	0.004	93	72
GATE	*5*		0.030	0.083	0.099	0.104	-0.06	-0.05	0.001	94	78
IATE	*N*	*ols*	0.031	0.179	-	0.225	-	-	-	33	22
IATE	*N*	*rf*	0.020	0.258	-	0.328	-	-	-	-	-
ATE	*1*	*dml-*	0.032	0.044	0.045	0.055	-0.02	-0.21	0.004	93	73
GATE	*5*	*norm*	0.032	0.083	0.099	0.104	-0.05	-0.05	0.001	94	78
IATE	*N*	*ols*	0.032	0.178	-	0.225	-	-	-	33	22
IATE	*N*	*rf*	0.021	0.257	-	0.327	-	-	-	-	-
ATE	*1*	*ols*	0.007	0.036	0.031	0.044	-0.04	-0.16	-0.013	83	61
GATE	*5*		0.006	0.078	0.094	0.097	-0.06	-0.01	-0.027	82	61
IATE	*N*		0.007	0.165	-	0.208	-	-	-	83	61

Note: Table to be continued.

Table C.5 - continued: Linear IATE, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.191	0.191	0.029	0.192	0.14	-0.10	0.003	0	0
GATE	*5*		0.191	0.191	0.044	0.206	0.12	-0.16	0.000	14	7
IATE	*N*		0.191	0.236	-	0.284	-	-	-	43	28
IATE eff	*N*		0.191	0.233	-	0.279	-	-	-	-	-
ATE	*1*	*mcf*	0.026	0.031	0.027	0.037	0.21	0.20	0.003	90	66
GATE	*5*	*cent*	0.025	0.064	0.044	0.078	0.19	0.06	-0.002	69	48
IATE	*N*		0.026	0.177	-	0.223	-	-	-	53	36
IATE eff	*N*		0.026	0.173	-	0.219	-	-	-	-	-
ATE	*1*	*grf*	0.084	0.084	0.020	0.087	-0.22	-0.32	0.001	2	0
GATE	*5*		0.084	0.086	0.047	0.098	-0.07	-0.11	-0.000	55	30
IATE	*N*		0.078	0.204	-	0.254	-	-	-	53	37
ATE	*1*	*grf*	0.024	0.027	0.022	0.033	-0.10	0.20	-0.001	79	50
GATE	*5*	*cent*	0.025	0.046	0.047	0.057	-0.02	-0.08	-0.001	88	68
IATE	*N*		0.024	0.195	-	0.244	-	-	-	55	38
ATE	*1*	*dml*	0.018	0.023	0.022	0.028	-0.07	0.51	0.003	90	74
GATE	*5*		0.018	0.042	0.049	0.053	0.12	0.00	0.000	93	78
IATE	*N*	*ols*	0.018	0.087	-	0.110	-	-	-	17	11
IATE	*N*	*rf*	0.011	0.195	-	0.248	-	-	-	-	-
ATE	*1*	*dml-*	0.018	0.023	0.022	0.028	-0.08	0.53	0.003	90	74
GATE	*5*	*norm*	0.018	0.042	0.049	0.053	0.12	-0.01	0.000	93	78
IATE	*N*	*ols*	0.019	0.087	-	0.110	-	-	-	17	11
IATE	*N*	*rf*	0.011	0.195	-	0.248	-	-	-	-	-
ATE	*1*	*ols*	0.007	0.017	0.021	0.022	0.05	0.17	-0.015	81	66
GATE	*5*		0.007	0.043	0.042	0.053	0.08	0.12	-0.014	77	56
IATE	*N*		0.007	0.083	-	0.104	-	-	-	83	66

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.6: Linear IATE, strong selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.528	0.528	0.065	0.532	0.03	0.06	-0.001	0	0
GATE	*5*		0.528	0.528	0.086	0.532	0.03	0.09	0.009	1	0
IATE	*N*		0.528	0.535	-	0.597	-	-	-	20	11
IATE eff	*N*		0.531	0.536	-	0.593	-	-	-	-	-
ATE	*1*	*mcf*	0.085	0.088	0.057	0.102	0.03	-0.06	0.007	76	49
GATE	*5*	*cent*	0.084	0.120	0.078	0.157	0.01	0.11	0.012	73	52
IATE	*N*		0.085	0.241	-	0.301	-	-	-	61	42
IATE eff	*N*		0.087	0.234	-	0.292	-	-	-	-	-
ATE	*1*	*grf*	0.348	0.348	0.050	0.352	-0.05	-0.07	-0.010	0	0
GATE	*5*		0.348	0.348	0.096	0.372	-0.06	-0.03	-0.005	10	3
IATE	*N*		0.331	0.382	-	0.461	-	-	-	32	21
ATE	*1*	*grf*	0.117	0.118	0.046	0.126	0.04	-0.07	-0.006	19	7
GATE	*5*	*cent*	0.118	0.142	0.090	0.175	0.02	0.15	-0.002	65	47
IATE	*N*		0.120	0.274	-	0.342	-	-	-	46	31
ATE	*1*	*dml*	0.197	0.198	0.060	0.207	-0.57	3.72	-0.003	9	2
GATE	*5*		0.197	0.208	0.129	0.208	-0.95	10.36	-0.016	54	33
IATE	*N*	*ols*	0.198	0.278	-	0.347	-	-	-	26	17
IATE	*N*	*rf*	0.170	0.344	-	0.568	-	-	-	-	-
ATE	*1*	*dml-*	0.188	0.188	0.056	0.196	-0.02	-0.02	0.002	9	2
GATE	*5*	*norm*	0.188	0.197	0.123	0.229	-0.13	0.02	-0.016	61	39
IATE	*N*	*ols*	0.189	0.271	-	0.336	-	-	-	29	20
IATE	*N*	*rf*	0.159	0.344	-	0.450	-	-	-	-	-
ATE	*1*	*ols*	0.109	0.110	0.053	0.121	-0.03	-0.09	-0.015	25	12
GATE	*5*		0.109	0.127	0.103	0.153	-0.04	0.01	-0.030	60	41
IATE	*N*		0.109	0.193	-	0.240	-	-	-	77	57

Note: Table to be continued.

Table C.6 - continued: Linear IATE, strong selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.469	0.469	0.031	0.470	-0.08	0.68	0.001	0	0
GATE	*5*		0.468	0.468	0.048	0.477	-0.06	0.08	0.005	0	0
IATE	*N*		0.469	0.475	-	0.526	-	-	-	13	7
IATE eff	*N*		0.469	0.475	-	0.523	-	-	-	-	-
ATE	*1*	*mcf*	0.012	0.024	0.028	0.030	0.26	0.68	0.005	98	85
GATE	*5*	*cent*	0.012	0.093	0.042	0.113	-0.03	0.15	0.007	57	38
IATE	*N*		0.012	0.208	-	0.263	-	-	-	51	35
IATE eff	*N*		0.012	0.204	-	0.258	-	-	-	-	-
ATE	*1*	*grf*	0.256	0.256	0.024	0.257	-0.31	0.07	-0.004	0	0
GATE	*5*		0.255	0.255	0.049	0.274	0.00	-0.05	-0.002	4	2
IATE	*N*		0.240	0.317	-	0.388	-	-	-	38	25
ATE	*1*	*grf*	0.058	0.058	0.026	0.064	-0.08	-0.07	-0.006	24	10
GATE	*5*	*cent*	0.058	0.096	0.048	0.119	-0.10	0.04	-0.003	54	38
IATE	*N*		0.065	0.251	-	0.314	-	-	-	47	32
ATE	*1*	*dml*	0.127	0.127	0.035	0.132	0.47	3.25	-0.003	6	2
GATE	*5*		0.127	0.133	0.078	0.153	-0.02	7.79	-0.024	44	26
IATE	*N*	*ols*	0.128	0.169	-	0.211	-	-	-	13	9
IATE	*N*	*rf*	0.104	0.259	-	0.462	-	-	-	-	-
ATE	*1*	*dml-*	0.117	0.117	0.032	0.121	0.47	3.25	-0.003	6	2
GATE	*5*	*norm*	0.117	0.122	0.072	0.141	-0.02	7.79	-0.024	44	26
IATE	*N*	*ols*	0.117	0.161	-	0.199	-	-	-	13	9
IATE	*N*	*rf*	0.092	0.259	-	0.374	-	-	-	-	-
ATE	*1*	*ols*	0.111	0.111	0.027	0.114	-0.08	-0.28	-0.008	0	0
GATE	*5*		0.111	0.112	0.051	0.126	0.02	-0.10	-0.015	25	14
IATE	*N*		0.111	0.129	-	0.156	-	-	-	60	40

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.7: Nonlinear IATE, no selectivity

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.000	0.048	0.060	0.060	0.03	0.00	0.006	97	85
GATE	*5*		-0.001	0.080	0.077	0.110	0.06	0.25	0.004	86	65
IATE	*N*		-0.001	0.208	-	0.252	-	-	-	70	48
IATE eff	*N*		-0.001	0.201	-	0.240	-	-	-	-	-
ATE	*1*	*mcf*	0.002	0.049	0.062	0.062	0.00	0.06	-0.001	94	78
GATE	*5*	*cent*	0.002	0.095	0.081	0.118	0.02	0.20	-0.008	76	55
IATE	*N*		0.002	0.220	-	0.264	-	-	-	56	37
IATE eff	*N*		0.004	0.214	-	0.255	-	-	-	-	-
ATE	*1*	*grf*	-0.001	0.034	0.043	0.043	-0.15	0.03	0.000	95	82
GATE	*5*		-0.001	0.077	0.097	0.097	0.03	0.07	-0.001	95	80
IATE	*N*		-0.001	0.229	-	0.271	-	-	-	50	33
ATE	*1*	*grf*	0.004	0.033	0.042	0.042	0.10	0.04	0.000	94	81
GATE	*5*	*cent*	0.004	0.075	0.094	0.095	0.04	0.06	0.000	95	80
IATE	*N*		0.004	0.228	-	0.269	-	-	-	49	33
ATE	*1*	*dml*	-0.002	0.034	0.042	0.043	0.05	-0.07	0.005	98	84
GATE	*5*		-0.002	0.076	0.095	0.095	-0.01	0.04	0.001	95	81
IATE	*N*	*ols*	-0.002	0.172	-	0.218	-	-	-	32	21
IATE	*N*	*rf*	-0.002	0.248	-	0.313	-	-	-	-	-
ATE	*1*	*dml-*	-0.002	0.034	0.043	0.043	0.06	-0.06	0.005	97	83
GATE	*5*	*norm*	-0.002	0.075	0.095	0.095	-0.02	0.06	0.000	95	80
IATE	*N*	*ols*	-0.002	0.171	-	0.218	-	-	-	31	21
IATE	*N*	*rf*	-0.002	0.246	-	0.313	-	-	-	-	-
ATE	*1*	*ols*	-0.001	0.034	0.042	0.042	0.05	-0.09	-0.013	82	74
GATE	*5*		-0.002	0.073	0.092	0.092	-0.03	0.09	-0.027	84	64
IATE	*N*		-0.001	0.168	-	0.212	-	-	-	84	64

Note: Table to be continued.

Table C.7 - continued: Nonlinear IATE, no selectivity

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.000	0.022	0.030	0.030	0.25	0.03	0.003	98	86
GATE	*5*		-0.001	0.059	0.042	0.073	0.08	-0.04	0.001	75	53
IATE	*N*		0.000	0.169	-	0.203	-	-	-	61	41
IATE eff	*N*		0.000	0.165	-	0.197	-	-	-	-	-
ATE	*1*	*mcf*	0.003	0.022	0.028	0.028	0.11	-0.15	0.002	97	78
GATE	*5*	*cent*	0.003	0.062	0.048	0.077	0.01	-0.23	-0.009	66	47
IATE	*N*		0.003	0.183	-	0.218	-	-	-	45	30
IATE eff	*N*		0.002	0.180	-	0.214	-	-	-	-	-
ATE	*1*	*grf*	-0.000	0.016	0.020	0.020	-0.10	0.50	0.001	96	83
GATE	*5*		-0.001	0.037	0.048	0.048	-0.16	0.19	-0.000	94	81
IATE	*N*		-0.001	0.189	-	0.223	-	-	-	54	36
ATE	*1*	*grf*	0.003	0.016	0.020	0.021	0.02	0.15	0.000	96	79
GATE	*5*	*cent*	0.003	0.036	0.045	0.045	-0.07	-0.09	0.001	95	80
IATE	*N*		0.003	0.187	-	0.221	-	-	-	54	36
ATE	*1*	*dml*	-0.001	0.016	0.020	0.020	-0.09	-0.01	0.004	98	86
GATE	*5*		-0.001	0.037	0.046	0.046	0.02	0.01	0.001	95	82
IATE	*N*	*ols*	-0.001	0.088	-	0.111	-	-	-	15	10
IATE	*N*	*rf*	-0.001	0.189	-	0.239	-	-	-	-	-
ATE	*1*	*dml-*	-0.001	0.016	0.020	0.020	-0.09	-0.02	0.003	98	86
GATE	*5*	*norm*	-0.001	0.037	0.046	0.046	0.02	0.00	0.001	95	82
IATE	*N*	*ols*	-0.001	0.087	-	0.111	-	-	-	15	10
IATE	*N*	*rf*	-0.001	0.189	-	0.239	-	-	-	-	-
ATE	*1*	*ols*	0.000	0.016	0.020	0.020	0.01	-0.01	-0.005	84	66
GATE	*5*		-0.001	0.037	0.045	0.046	0.01	0.01	-0.013	84	64
IATE	*N*		0.000	0.087	-	0.111	-	-	-	81	64

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs.

Table C.8: Nonlinear IATE, medium selectivity

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.219	0.219	0.060	0.227	-0.03	-0.14	0.004	6	1
GATE	*5*		0.219	0.221	0.080	0.244	0.07	0.07	0.005	31	18
IATE	*N*		0.219	0.276	-	0.336	-	-	-	58	41
IATE eff	*N*		0.219	0.268	-	0.336	-	-	-	-	-
ATE	*1*	*mcf*	0.056	0.067	0.061	0.083	0.09	0.29	-0.001	84	64
GATE	*5*	*cent*	0.056	0.102	0.081	0.126	0.05	0.15	-0.003	77	54
IATE	*N*		0.056	0.225	-	0.271	-	-	-	57	38
IATE eff	*N*		0.059		-	0.261	-	-	-	-	-
ATE	*1*	*grf*	0.128	0.128	0.044	0.135	-0.04	-0.24	-0.002	16	5
GATE	*5*		0.128	0.137	0.097	0.162	0.01	-0.04	-0.002	71	47
IATE	*N*		0.123	0.259	-	0.310	-	-	-	45	30
ATE	*1*	*grf*	0.050	0.055	0.042	0.065	0.03	-0.16	-0.000	77	52
GATE	*5*	*cent*	0.050	0.087	0.093	0.108	0.01	0.08	-0.000	91	72
IATE	*N*		0.050	0.244	-	0.289	-	-	-	45	30
ATE	*1*	*dml*	0.030	0.043	0.045	0.054	-0.02	-0.21	0.004	93	73
GATE	*5*		0.029	0.083	0.099	0.104	-0.05	-0.05	0.001	94	79
IATE	*N*	*ols*	0.030	0.183	-	0.230	-	-	-	32	22
IATE	*N*	*rf*	0.019	0.260	-	0.331	-	-	-	-	-
ATE	*1*	*dml-*	0.031	0.044	0.045	0.055	-0.03	-0.24	0.004	93	72
GATE	*5*	*norm*	0.031	0.083	0.099	0.104	-0.05	-0.05	0.001	94	78
IATE	*N*	*ols*	0.031	0.182	-	0.230	-	-	-	32	21
IATE	*N*	*rf*	0.020	0.259	-	0.330	-	-	-	-	-
ATE	*1*	*ols*	0.008	0.036	0.044	0.045	-0.04	-0.16	-0.013	83	60
GATE	*5*		0.008	0.079	0.094	0.098	-0.06	-0.06	-0.027	82	61
IATE	*N*		0.008	0.170	-	0.215	-	-	-	82	62

Note: Table to be continued.

Table C.8 - continued: Nonlinear IATE, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.181	0.181	0.030	0.184	0.22	-0.47	0.002	0	0
GATE	*5*		0.181	0.181	0.044	0.194	0.14	0.20	0.002	13	7
IATE	*N*		0.181	0.228	-	0.275	-	-	-	50	35
IATE eff	*N*		0.181	0.223	-	0.269	-	-	-	-	-
ATE	*1*	*mcf*	0.026	0.031	0.027	0.038	0.21	0.26	0.003	90	68
GATE	*5*	*cent*	0.025	0.063	0.045	0.077	0.18	0.05	-0.002	70	50
IATE	*N*		0.026	0.187	-	0.223	-	-	-	47	31
IATE eff	*N*		0.026	0.184	-	0.219	-	-	-	-	-
ATE	*1*	*grf*	0.084	0.084	0.020	0.086	-0.22	-0.35	0.001	2	0
GATE	*5*		0.084	0.086	0.047	0.097	-0.08	-0.14	-0.000	55	30
IATE	*N*		0.078	0.214	-	0.254	-	-	-	48	32
ATE	*1*	*grf*	0.024	0.027	0.022	0.033	-0.08	0.18	-0.001	78	51
GATE	*5*	*cent*	0.024	0.046	0.047	0.056	-0.03	-0.10	-0.001	89	69
IATE	*N*		0.024	0.207	-	0.244	-	-	-	48	32
ATE	*1*	*dml*	0.017	0.022	0.022	0.028	-0.05	0.49	0.003	91	75
GATE	*5*		0.017	0.042	0.049	0.053	0.12	-0.01	0.000	94	78
IATE	*N*	*ols*	0.018	0.094	-	0.120	-	-	-	16	10
IATE	*N*	*rf*	0.010	0.198	-	0.251	-	-	-	-	-
ATE	*1*	*dml-*	0.017	0.022	0.022	0.028	-0.07	0.50	0.003	90	74
GATE	*5*	*norm*	0.017	0.042	0.049	0.053	0.12	-0.01	0.000	93	78
IATE	*N*	*ols*	0.018	0.094	-	0.120	-	-	-	16	10
IATE	*N*	*rf*	0.011	0.198	-	0.251	-	-	-	-	-
ATE	*1*	*ols*	0.009	0.018	0.021	0.023	0.07	0.16	-0.005	81	65
GATE	*5*		0.008	0.039	0.047	0.055	0.08	0.12	-0.013	75	53
IATE	*N*		0.009	0.091	-	0.116	-	-	-	79	59

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.9: Nonlinear IATE, strong selectivity

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.517	0.517	0.062	0.521	-0.07	-0.05	0.002	0	0
GATE	*5*		0.517	0.517	0.087	0.532	0.01	0.05	0.012	1	0
IATE	*N*		0.518	0.520	-	0.589	-	-	-	28	16
IATE eff	*N*		0.518	0.520	-	0.581	-	-	-	-	-
ATE	*1*	*mcf*	0.091	0.094	0.057	0.108	0.03	-0.04	0.007	73	44
GATE	*5*	*cent*	0.091	0.129	0.078	0.160	0.04	0.12	0.012	72	51
IATE	*N*		0.091	0.252	-	0.303	-	-	-	56	38
IATE eff	*N*		0.093	0.246	-	0.294	-	-	-	-	-
ATE	*1*	*grf*	0.356	0.356	0.051	0.359	-0.04	-0.09	-0.010	0	0
GATE	*5*		0.356	0.356	0.096	0.379	-0.06	-0.05	-0.006	9	3
IATE	*N*		0.341	0.386	-	0.469	-	-	-	35	24
ATE	*1*	*grf*	0.124	0.125	0.046	0.133	0.04	-0.08	-0.006	15	5
GATE	*5*	*cent*	0.125	0.146	0.091	0.178	0.02	0.15	-0.002	64	45
IATE	*N*		0.128	0.289	-	0.346	-	-	-	41	27
ATE	*1*	*dml*	0.198	0.199	0.061	0.207	-0.53	3.20	-0.003	9	2
GATE	*5*		0.198	0.208	0.130	0.240	-0.89	9.06	-0.016	53	33
IATE	*N*	*ols*	0.199	0.280	-	0.350	-	-	-	26	17
IATE	*N*	*rf*	0.170	0.345	-	0.563	-	-	-	-	-
ATE	*1*	*dml-*	0.189	0.189	0.057	0.197	-0.02	-0.01	0.002	11	3
GATE	*5*	*norm*	0.188	0.198	0.124	0.228	-0.13	0.03	-0.005	62	39
IATE	*N*	*ols*	0.189	0.273	-	0.339	-	-	-	30	20
IATE	*N*	*rf*	0.159	0.346	-	0.452	-	-	-	-	-
ATE	*1*	*ols*	0.124	0.125	0.053	0.135	-0.03	0.10	-0.015	17	8
GATE	*5*		0.124	0.139	0.103	0.165	-0.04	0.01	-0.030	55	36
IATE	*N*		0.124	0.202	-	0.253	-	-	-	75	54

Note: Table to be continued.

Table C.9 - continued: Nonlinear IATE, strong selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.455	0.455	0.032	0.456	-0.02	-0.16	0.001	0	0
GATE	*5*		0.455	0.455	0.048	0.463	-0.02	-0.14	0.007	0	0
IATE	*N*		0.455	0.458	-	0.513	-	-	-	21	12
IATE eff	*N*		0.453	0.454	-	0.506	-	-	-	-	-
ATE	*1*	*mcf*	0.017	0.026	0.028	0.033	0.26	0.69	0.005	96	83
GATE	*5*	*cent*	0.017	0.091	0.043	0.110	-0.03	0.12	0.007	59	40
IATE	*N*		0.017	0.218	-	0.260	-	-	-	45	30
IATE eff	*N*		0.018	0.215	-	0.255	-	-	-	-	-
ATE	*1*	*grf*	0.261	0.261	0.024	0.262	-0.27	-0.01	-0.004	0	0
GATE	*5*		0.261	0.261	0.049	0.278	0.03	-0.02	-0.002	3	1
IATE	*N*		0.248	0.322	-	0.392	-	-	-	38	25
ATE	*1*	*grf*	0.063	0.063	0.026	0.068	-0.06	-0.11	-0.006	20	8
GATE	*5*	*cent*	0.064	0.095	0.049	0.118	-0.10	0.01	-0.003	54	37
IATE	*N*		0.072	0.265	-	0.314	-	-	-	41	27
ATE	*1*	*dml*	0.126	0.126	0.035	0.130	0.51	3.24	-0.003	6	2
GATE	*5*		0.125	0.131	0.078	0.150	0.02	7.37	-0.014	45	26
IATE	*N*	*ols*	0.126	0.170	-	0.212	-	-	-	13	9
IATE	*N*	*rf*	0.102	0.259	-	0.459	-	-	-	-	-
ATE	*1*	*dml-*	0.114	0.114	0.032	0.119	0.16	0.35	0.000	6	2
GATE	*5*	*norm*	0.114	0.120	0.072	0.138	-0.10	0.17	-0.006	54	35
IATE	*N*	*ols*	0.115	0.162	-	0.200	-	-	-	15	10
IATE	*N*	*rf*	0.089	0.260	-	0.375	-	-	-	-	-
ATE	*1*	*ols*	0.126	0.126	0.027	0.129	0.08	-0.12	-0.008	0	0
GATE	*5*		0.126	0.127	0.052	0.140	0.02	-0.11	-0.015	20	11
IATE	*N*		0.126	0.145	-	0.175	-	-	-	55	36

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.10: Quadratic IATE, no selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	-0.003	0.053	0.065	0.065	0.01	-0.22	0.001	96	80
GATE	*5*		-0.003	0.090	0.074	0.110	0.06	0.02	0.003	81	59
IATE	*N*		-0.003	0.450	-	0.648	-	-	-	30	18
IATE eff	*N*		-0.003	0.448	-	0.644	-	-	-	-	-
ATE	*1*	*mcf*	0.002	0.053	0.065	0.065	0.00	-0.07	-0.001	95	80
GATE	*5*	*cent*	0.002	0.091	0.074	0.110	0.07	0.20	-0.002	79	55
IATE	*N*		0.002	0.455	-	0.653	-	-	-	27	55
IATE eff	*N*		0.004	0.453	-	0.649	-	-	-	-	-
ATE	*1*	*grf*	-0.002	0.037	0.046	0.046	-0.13	0.06	-0.001	95	80
GATE	*5*		-0.002	0.082	0.103	0.103	0.01	-0.03	-0.001	95	80
IATE	*N*		-0.001	0.465	-	0.668	-	-	-	24	15
ATE	*1*	*grf*	0.003	0.036	0.045	0.045	0.04	-0.03	0.000	95	81
GATE	*5*	*cent*	0.003	0.081	0.101	0.101	-0.02	-0.02	-0.000	95	79
IATE	*N*		0.004	0.465	-	0.665	-	-	-	23	15
ATE	*1*	*dml*	-0.001	0.036	0.045	0.046	-0.01	0.01	0.005	97	84
GATE	*5*		-0.002	0.080	0.101	0.101	0.00	0.06	0.002	95	81
IATE	*N*	*ols*	-0.002	0.505	-	0.720	-	-	-	13	8
IATE	*N*	*rf*	-0.004	0.446	-	0.626	-	-	-	-	-
ATE	*1*	*dml-*	-0.002	0.036	0.046	0.046	-0.01	0.01	0.004	97	83
GATE	*5*	*norm*	-0.002	0.080	0.101	0.101	0.00	0.07	0.000	95	80
IATE	*N*	*ols*	-0.002	0.505	-	0.720	-	-	-	12	8
IATE	*N*	*rf*	-0.004	0.444	-	0.625	-	-	-	-	-
ATE	*1*	*ols*	-0.002	0.037	0.046	0.046	0.04	0.03	-0.014	85	61
GATE	*5*		-0.002	0.080	0.101	0.101	0.01	0.09	-0.029	83	65
IATE	*N*		-0.002	0.505	-	0.720	-	-	-	39	25

Note: Table to be continued.

Table C.10 - continued: Quadratic IATE, no selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.001	0.024	0.030	0.030	0.22	-0.02	0.004	97	84
GATE	*5*		0.001	0.062	0.040	0.073	0.15	-0.15	0.001	68	45
IATE	*N*		0.001	0.405	-	0.584	-	-	-	21	13
IATE eff	*N*		0.000	0.404	-	0.583	-	-	-	-	-
ATE	*1*	*mcf*	0.002	0.024	0.029	0.029	0.21	0.01	0.002	96	82
GATE	*5*	*cent*	0.002	0.059	0.042	0.070	0.16	-0.09	-0.002	69	45
IATE	*N*		0.002	0.409	-	0.588	-	-	-	19	12
IATE eff	*N*		0.002	0.408	-	0.586	-	-	-	-	-
ATE	*1*	*grf*	-0.001	0.017	0.021	0.021	-0.02	0.21	0.001	95	83
GATE	*5*		-0.001	0.041	0.051	0.051	-0.06	-0.115	-0.001	94	79
IATE	*N*		-0.001	0.423	-	0.607	-	-	-	25	16
ATE	*1*	*grf*	0.002	0.018	0.022	0.022	-0.01	-0.13	0.000	96	79
GATE	*5*	*cent*	0.002	0.039	0.048	0.048	-0.08	-0.08	0.001	96	81
IATE	*N*		0.002	0.422	-	0.605	-	-	-	25	16
ATE	*1*	*dml*	0.000	0.016	0.020	0.020	0.00	-0.12	0.004	99	86
GATE	*5*		0.000	0.039	0.049	0.049	0.06	-0.07	0.001	95	81
IATE	*N*	*ols*	0.000	0.483	-	0.693	-	-	-	3	2
IATE	*N*	*rf*	-0.004	0.339	-	0.479	-	-	-	-	-
ATE	*1*	*dml-*	0.000	0.016	0.020	0.020	0.00	-0.13	0.004	99	86
GATE	*5*	*norm*	0.000	0.039	0.049	0.049	0.06	-0.06	0.001	96	81
IATE	*N*	*ols*	0.000	0.483	-	0.693	-	-	-	3	2
IATE	*N*	*rf*	-0.004	0.339	-	0.479	-	-	-	-	-
ATE	*1*	*ols*	0.001	0.016	0.021	0.021	0.02	0.09	-0.005	84	68
GATE	*5*		0.000	0.040	0.050	0.050	0.00	-0.04	-0.014	83	64
IATE	*N*		0.000	0.483	-	0.693	-	-	-	17	11

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.11: Quadratic IATE, medium selectivity

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.123	0.124	0.064	0.139	0.10	0.08	0.001	52	27
GATE	*5*		0.123	0.155	0.082	0.179	0.11	0.02	0.003	54	35
IATE	*N*		0.123	0.504	-	0.667	-	-	-	25	16
IATE eff	*N*		0.123	0.503	-	0.663	-	-	-	-	-
ATE	*1*	*mcf*	-0.036	0.059	0.064	0.074	0.05	0.25	-0.002	90	73
GATE	*5*	*cent*	-0.036	0.099	0.080	0.127	0.25	0.18	-0.001	79	60
IATE	*N*		-0.036	0.446	-	0.660	-	-	-	31	19
IATE eff	*N*		-0.034	0.444	-	0.656	-	-	-	-	-
ATE	*1*	*grf*	0.030	0.044	0.045	0.054	-0.05	-0.14	-0.001	90	68
GATE	*5*		0.030	0.100	0.099	0.125	0.05	0.01	-0.001	88	68
IATE	*N*		0.014	0.480	-	0.682	-	-	-	22	14
ATE	*1*	*grf*	-0.046	0.053	0.044	0.064	0.05	-0.30	-0.001	79	57
GATE	*5*	*cent*	-0.046	0.100	0.096	0.128	-0.02	0.10	0.001	86	69
IATE	*N*		-0.057	0.456	-	0.682	-	-	-	25	16
ATE	*1*	*dml*	-0.014	0.040	0.048	0.050	-0.04	-0.22	0.004	96	82
GATE	*5*		-0.014	0.090	0.104	0.113	-0.04	0.02	0.002	93	77
IATE	*N*	*ols*	-0.014	0.509	-	0.730	-	-	-	14	9
IATE	*N*	*rf*	-0.019	0.467	-	0.667	-	-	-	-	-
ATE	*1*	*dml-*	-0.013	0.040	0.048	0.050	-0.04	-0.25	0.004	96	82
GATE	*5*	*norm*	-0.013	0.090	0.104	0.113	-0.02	0.02	0.002	93	77
IATE	*N*	*ols*	-0.014	0.508	-	0.730	-	-	-	14	9
IATE	*N*	*rf*	-0.019	0.467	-	0.666	-	-	-	-	-
ATE	*1*	*ols*	-0.104	0.105	0.049	0.115	0.01	-0.10	-0.016	20	10
GATE	*5*		-0.104	0.156	0.101	0.156	-0.05	0.02	-0.031	56	41
IATE	*N*		-0.104	0.526	-	0.798	-	-	-	56	41

Note: Table to be continued.

Table C.11 - continued: Quadratic IATE, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.092	0.092	0.029	0.097	0.25	0.05	0.003	18	3
GATE	*5*		0.091	0.124	0.045	0.137	0.26	0.14	0.002	33	17
IATE	*N*		0.091	0.452	-	0.612	-	-	-	19	12
IATE eff	*N*		0.091	0.451	-	0.609	-	-	-	-	-
ATE	*1*	*mcf*	-0.059	0.060	0.029	0.066	0.29	0.42	0.002	53	24
GATE	*5*	*cent*	-0.060	0.079	0.044	0.106	0.28	0.21	0.000	68	51
IATE	*N*		-0.060	0.401	-	0.611	-	-	-	23	15
IATE eff	*N*		-0.059	0.399	-	0.608	-	-	-	-	-
ATE	*1*	*grf*	-0.002	0.017	0.021	0.021	-0.22	0.03	0.001	95	82
GATE	*5*		-0.002	0.060	0.049	0.076	-0.04	-0.13	-0.001	82	60
IATE	*N*		-0.030	0.446	-	0.655	-	-	-	22	14
ATE	*1*	*grf*	-0.061	0.061	0.023	0.065	-0.06	0.41	-0.001	20	8
GATE	*5*	*cent*	-0.061	0.073	0.050	0.095	-0.02	-0.21	-0.001	71	53
IATE	*N*		-0.083	0.432	-	0.662	-	-	-	25	16
ATE	*1*	*dml*	-0.012	0.021	0.023	0.026	0.02	0.03	0.003	94	80
GATE	*5*		-0.013	0.048	0.053	0.062	0.07	0.02	0.000	91	74
IATE	*N*	*ols*	-0.013	0.482	-	0.698	-	-	-	3	2
IATE	*N*	*rf*	-0.016	0.370	-	0.537	-	-	-	-	-
ATE	*1*	*dml-*	-0.012	0.021	0.023	0.026	0.02	0.03	0.003	94	80
GATE	*5*	*norm*	-0.013	0.048	0.053	0.062	0.07	0.01	0.000	91	74
IATE	*N*	*ols*	-0.013	0.482	-	0.698	-	-	-	3	2
IATE	*N*	*rf*	-0.016	0.370	-	0.537	-	-	-	-	-
ATE	*1*	*ols*	-0.104	0.104	0.023	0.107	-0.01	-0.11	-0.007	0	0
GATE	*5*		-0.104	0.135	0.052	0.186	0.03	0.15	-0.017	44	31
IATE	*N*		-0.105	0.505	-	0.776	-	-	-	18	11

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs.

Table C.12: Quadratic IATE, strong selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.261	0.261	0.066	0.269	-0.03	0.06	-0.001	2	0
GATE	*5*		0.261	0.280	0.091	0.316	0.03	0.09	0.002	27	19
IATE	*N*		0.261	0.601	-	0.743	-	-	-	20	12
IATE eff	*N*		0.264	0.600	-	0.739	-	-	-	-	-
ATE	*1*	*mcf*	-0.182	0.182	0.057	0.191	-0.02	-0.06	0.007	16	4
GATE	*5*	*cent*	-0.182	0.187	0.078	0.231	0.03	0.03	0.008	52	33
IATE	*N*		-0.182	0.442	-	0.717	-	-	-	46	29
IATE eff	*N*		-0.180	0.437	-	0.712	-	-	-	-	-
ATE	*1*	*grf*	0.028	0.046	0.049	0.056	-0.10	-0.08	-0.008	85	62
GATE	*5*		0.029	0.120	0.094	0.146	-0.06	0.01	-0.003	76	54
IATE	*N*		-0.020	0.479	-	0.697	-	-	-	25	16
ATE	*1*	*grf*	-0.208	0.208	0.046	0.213	0.02	-0.04	-0.006	0	0
GATE	*5*	*cent*	-0.208	0.212	0.090	0.253	0.02	0.14	-0.001	46	28
IATE	*N*		-0.235	0.435	-	0.732	-	-	-	44	27
ATE	*1*	*dml*	-0.037	0.056	0.059	0.069	-0.02	-0.09	-0.003	87	69
GATE	*5*		-0.037	0.154	0.123	0.194	-0.10	0.48	-0.010	73	55
IATE	*N*	*ols*	-0.037	0.550	-	0.798	-	-	-	14	9
IATE	*N*	*rf*	-0.058	0.513	-	0.759	-	-	-	-	-
ATE	*1*	*dml-*	-0.048	0.061	0.057	0.074	0.01	-0.07	0.000	86	66
GATE	*5*	*norm*	-0.048	0.156	0.121	0.197	-0.08	0.04	-0.003	75	57
IATE	*N*	*ols*	-0.047	0.550	-	0.803	-	-	-	16	10
IATE	*N*	*rf*	-0.070	0.515	-	0.750	-	-	-	-	-
ATE	*1*	*ols*	-0.369	0.369	0.057	0.373	-0.01	0.03	-0.018	0	0
GATE	*5*		-0.369	0.392	0.108	0.521	-0.07	-0.02	-0.034	33	24
IATE	*N*		-0.369	0.675	-	1.091	-	-	-	38	24

Note: Table to be continued.

Table C.12 - continued: Quadratic IATE, strong selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.198	0.198	0.032	0.201	-0.04	0.57	0.001	0	0
GATE	*5*		0.198	0.231	0.052	0.256	0.09	0.09	-0.001	15	8
IATE	*N*		0.198	0.552	-	0.702	-	-	-	15	9
IATE eff	*N*		0.198	0.551	-	0.700	-	-	-	-	-
ATE	*1*	*mcf*	-0.254	0.254	0.029	0.255	0.34	0.51	0.004	0	0
GATE	*5*	*cent*	-0.254	0.254	0.044	0.280	0.11	0.13	0.003	3	1
IATE	*N*		-0.254	0.420	-	0.721	-	-	-	40	25
IATE eff	*N*		-0.254	0.418	-	0.719	-	-	-	-	-
ATE	*1*	*grf*	-0.065	0.065	0.025	0.069	-0.21	-0.02	-0.004	16	6
GATE	*5*		-0.065	0.094	0.049	0.129	-0.03	-0.02	-0.002	66	49
IATE	*N*		-0.124	0.448	-	0.703	-	-	-	30	18
ATE	*1*	*grf*	-0.264	0.264	0.026	0.265	-0.04	-0.21	-0.006	0	0
GATE	*5*	*cent*	-0.264	0.264	0.049	0.287	-0.05	0.06	-0.003	2	0
IATE	*N*		-0.299	0.433	-	0.754	-	-	-	49	31
ATE	*1*	*dml*	-0.069	0.071	0.036	0.078	0.58	1.44	-0.004	38	19
GATE	*5*		-0.069	0.123	0.079	0.167	0.73	5.73	-0.014	65	49
IATE	*N*	*ols*	-0.069	0.507	-	0.761	-	-	-	5	3
IATE	*N*	*rf*	-0.083	0.446	-	0.772	-	-	-	-	-
ATE	*1*	*dml-*	-0.078	0.079	0.033	0.085	0.20	0.18	-0.001	31	14
GATE	*5*	*norm*	-0.078	0.124	0.074	0.169	0.07	0.28	-0.007	67	50
IATE	*N*	*ols*	-0.078	0.505	-	0.762	-	-	-	5	3
IATE	*N*	*rf*	-0.094	0.448	-	0.702	-	-	-	-	-
ATE	*1*	*ols*	-0.366	0.366	0.030	0.368	0.14	0.02	-0.011	0	0
GATE	*5*		-0.367	0.375	0.055	0.510	0.06	-0.08	-0.019	25	17
IATE	*N*		-0.367	0.656	-	1.078	-	-	-	17	11

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 observations.

Table C.13: Step-function IATE, no selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	-0.003	0.048	0.061	0.061	0.005	0.00	-0.10	97	83
GATE	*5*		-0.004	0.092	0.087	0.115	0.000	0.01	0.02	86	67
IATE	*N*		-0.004	0.158	-	0.193	-	-	-	83	62
IATE eff	*N*		-0.002	0.147	-	0.178	-	-	-	-	-
ATE	*1*	*mcf*	0.003	0.048	0.061	0.061	0.03	0.06	-0.001	94	78
GATE	*5*	*cent*	0.002	0.090	0.091	0.114	0.02	0.09	-0.010	83	64
IATE	*N*		0.002	0.162	-	0.199	-	-	-	77	56
IATE eff	*N*		0.004	0.151	-	0.184	-	-	-	-	-
ATE	*1*	*grf*	-0.001	0.035	0.043	0.043	-0.13	-0.04	-0.000	95	80
GATE	*5*		-0.001	0.076	0.096	0.096	0.06	0.09	-0.001	95	80
IATE	*N*		-0.001	0.158	-	0.190	-	-	-	76	55
ATE	*1*	*grf*	0.003	0.033	0.042	0.042	0.07	-0.02	0.000	95	80
GATE	*5*	*cent*	0.004	0.075	0.094	0.094	0.01	0.01	0.000	95	80
IATE	*N*		0.004	0.155	-	0.187	-	-	-	77	57
ATE	*1*	*dml*	-0.002	0.034	0.043	0.043	0.04	-0.12	0.005	97	85
GATE	*5*		-0.002	0.075	0.094	0.095	-0.02	0.04	0.002	95	81
IATE	*N*	*ols*	-0.002	0.230	-	0.287	-	-	-	24	16
IATE	*N*	*rf*	-0.002	0.273	-	0.345	-	-	-	-	-
ATE	*1*	*dml-*	-0.002	0.034	0.043	0.043	0.04	-0.11	0.005	97	84
GATE	*5*	*norm*	-0.002	0.075	0.094	0.094	-0.02	0.05	0.002	95	80
IATE	*N*	*ols*	-0.002	0.229	-	0.286	-	-	-	23	15
IATE	*N*	*rf*	-0.002	0.270	-	0.342	-	-	-	-	-
ATE	*1*	*ols*	-0.001	0.034	0.042	0.042	0.04	-0.11	-0.013	82	61
GATE	*5*		-0.002	0.073	0.092	0.093	-0.04	0.09	-0.027	84	64
IATE	*N*		-0.001	0.228	-	0.284	-	-	-	68	48

Note: Table to be continued.

Table C.13 - continued: Step-function IATE, no selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.001	0.022	0.028	0.028	0.13	-0.15	0.005	98	84
GATE	*5*		0.001	0.049	0.045	0.063	0.07	-0.21	0.002	87	66
IATE	*N*		0.001	0.097	-	0.121	-	-	-	86	68
IATE eff	*N*		0.000	0.089	-	0.111	-	-	-	-	-
ATE	*1*	*mcf*	0.004	0.022	0.028	0.028	0.028	0.15	0.002	97	81
GATE	*5*	*cent*	0.004	0.045	0.045	0.058	0.045	0.17	-0.002	87	68
IATE	*N*		0.004	0.096	-	0.121	-	-	-	83	63
IATE eff	*N*		0.003	0.088	-	0.110	-	-	-	-	-
ATE	*1*	*grf*	-0.000	0.016	0.020	0.020	-0.17	0.39	0.001	96	83
GATE	*5*		-0.001	0.037	0.047	0.047	-0.09	0.13	-0.001	94	80
IATE	*N*		-0.001	0.083	-	0.105	-	-	-	91	76
ATE	*1*	*grf*	0.002	0.017	0.021	0.021	0.01	0.02	0.000	96	79
GATE	*5*	*cent*	0.002	0.036	0.046	0.046	-0.07	-0.05	0.001	95	80
IATE	*N*		0.002	0.082	-	0.103	-	-	-	91	76
ATE	*1*	*dml*	0.000	0.016	0.020	0.020	-0.13	0.25	0.003	97	86
GATE	*5*		-0.001	0.037	0.046	0.046	0.06	-0.02	0.001	95	82
IATE	*N*	*ols*	-0.001	0.179	-	0.219	-	-	-	7	5
IATE	*N*	*rf*	0.000	0.202	-	0.257	-	-	-	-	-
ATE	*1*	*dml-*	0.000	0.016	0.020	0.020	-0.14	0.25	0.003	98	85
GATE	*5*	*norm*	-0.001	0.037	0.046	0.046	0.05	-0.03	0.001	95	82
IATE	*N*	*ols*	-0.001	0.179	-	0.219	-	-	-	7	4
IATE	*N*	*rf*	0.000	0.201	-	0.256	-	-	-	-	-
ATE	*1*	*ols*	0.000	0.016	0.020	0.020	-0.09	-0.02	-0.005	84	66
GATE	*5*		-0.001	0.037	0.046	0.046	0.01	-0.03	-0.013	85	62
IATE	*N*		-0.001	0.179	-	0.219	-	-	-	45	30

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.14: Step-function IATE, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.175	0.175	0.062	0.186	0.11	0.13	0.002	21	7
GATE	*5*		0.175	0.178	0.090	0.211	0.06	-0.04	0.008	53	32
IATE	*N*		0.175	0.214	-	0.268	-	-	-	67	50
IATE eff	*N*		0.178	0.205	-	0.258	-	-	-	-	-
ATE	*1*	*mcf*	0.037	0.057	0.061	0.072	0.09	0.26	0.000	91	72
GATE	*5*	*cent*	0.037	0.094	0.090	0.121	0.05	0.07	-0.006	84	65
IATE	*N*		0.037	0.174	-	0.212	-	-	-	75	53
IATE eff	*N*		0.040	0.163	-	0.197	-	-	-	-	-
ATE	*1*	*grf*	0.097	0.097	0.044	0.107	-0.03	-0.24	-0.002	38	17
GATE	*5*		0.097	0.113	0.096	0.137	0.03	-0.00	-0.003	81	57
IATE	*N*		0.092	0.188	-	0.231	-	-	-	70	50
ATE	*1*	*grf*	0.037	0.046	0.042	0.056	-0.00	-0.20	-0.000	85	64
GATE	*5*	*cent*	0.037	0.081	0.093	0.102	-0.01	0.07	-0.000	93	76
IATE	*N*		0.034	0.175	-	0.210	-	-	-	72	51
ATE	*1*	*dml*	0.019	0.040	0.045	0.045	-0.03	-0.24	0.004	96	79
GATE	*5*		0.019	0.081	0.099	0.101	-0.04	-0.05	0.001	95	80
IATE	*N*	*ols*	0.019	0.237	-	0.296	-	-	-	25	16
IATE	*N*	*rf*	0.012	0.285	-	0.364	-	-	-	-	-
ATE	*1*	*dml-*	0.020	0.040	0.045	0.050	-0.04	-0.27	0.004	95	78
GATE	*5*	*norm*	0.020	0.081	0.081	0.101	-0.04	-0.05	0.001	95	79
IATE	*N*	*ols*	0.020	0.236	-	0.295	-	-	-	24	16
IATE	*N*	*rf*	0.013	0.284	-	0.362	-	-	-	-	-
ATE	*1*	*ols*	0.007	0.037	0.044	0.045	-0.07	-0.20	-0.013	83	63
GATE	*5*		0.007	0.080	0.094	0.100	-0.06	0.01	-0.027	81	61
IATE	*N*		0.007	0.231	-	0.287	-	-	-	67	47

Note: Table to be continued.

Table C.14 - continued: Step-function IATE, medium selectivity

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						Estimation of std. errors		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.144	0.144	0.029	0.147	0.16	-0.02	0.003	0	0
GATE	*5*		0.144	0.144	0.048	0.158	0.16	-0.03	-0.001	21	7
IATE	*N*		0.144	0.158	-	0.193	-	-	-	45	63
IATE eff	*N*		0.144	0.153	-	0.186	-	-	-	-	-
ATE	*1*	*mcf*	0.014	0.025	0.027	0.030	0.21	0.24	0.003	96	82
GATE	*5*	*cent*	0.013	0.049	0.045	0.063	0.22	0.17	0.004	86	69
IATE	*N*		0.013	0.106	-	0.131	-	-	-	81	59
IATE eff	*N*		0.013	0.100	-	0.122	-	-	-	-	-
ATE	*1*	*grf*	0.064	0.064	0.020	0.067	-0.18	-0.41	0.001	13	4
GATE	*5*		0.064	0.068	0.047	0.080	-0.03	-0.15	-0.001	71	46
IATE	*N*		0.063	0.096	-	0.126	-	-	-	86	71
ATE	*1*	*grf*	0.015	0.022	0.022	0.027	-0.08	0.05	-0.001	87	69
GATE	*5*	*cent*	0.016	0.041	0.048	0.051	-0.04	-0.03	-0.001	92	75
IATE	*N*		0.016	0.089	-	0.111	-	-	-	89	73
ATE	*1*	*dml*	0.010	0.019	0.022	0.024	-0.16	0.94	0.003	94	82
GATE	*5*		0.010	0.041	0.050	0.051	0.14	0.04	0.000	94	80
IATE	*N*	*ols*	0.010	0.181	-	0.222	-	-	-	8	5
IATE	*N*	*rf*	0.005	0.214	-	0.275	-	-	-	-	-
ATE	*1*	*dml-*	0.010	0.019	0.022	0.024	-0.17	0.97	0.003	94	82
GATE	*5*	*norm*	0.010	0.041	0.050	0.051	0.13	0.04	0.000	94	80
IATE	*N*	*ols*	0.010	0.181	-	0.222	-	-	-	8	5
IATE	*N*	*rf*	0.005	0.214	-	0.274	-	-	-	-	-
ATE	*1*	*ols*	0.008	0.017	0.021	0.022	0.08	0.21	-0.005	81	65
GATE	*5*		0.007	0.046	0.047	0.058	0.09	0.22	-0.014	74	53
IATE	*N*		0.007	0.182	-	0.222	-	-	-	44	30

Note: For GATE and IATE the results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95%/80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs.

Table C.15: Step-function IATE, strong selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.437	0.437	-0.001	0.441	0.03	-0.02	-0.001	0	0
GATE	*5*		0.436	0.436	0.001	0.455	0.00	0.06	0.001	1	0
IATE	*N*		0.436	0.438	-	0.494	-	-	-	32	16
IATE eff	*N*		0.439	0.439	-	0.489	-	-	-	-	-
ATE	*1*	*mcf*	0.046	0.060	0.058	0.074	0.01	0.05	0.008	92	72
GATE	*5*	*cent*	0.046	0.112	0.085	0.146	-0.03	0.11	0.009	81	62
IATE	*N*		0.046	0.212	-	0.252	-	-	-	68	45
IATE eff	*N*		0.048	0.205	-	0.240	-	-	-	-	-
ATE	*1*	*grf*	0.284	0.284	0.051	0.288	-0.06	-0.06	-0.010	0	0
GATE	*5*		0.284	0.284	0.101	0.309	-0.06	0.06	-0.011	19	8
IATE	*N*		0.272	0.315	-	0.394	-	-	-	49	37
ATE	*1*	*grf*	0.090	0.091	0.046	0.101	0.04	-0.04	-0.006	41	20
GATE	*5*	*cent*	0.090	0.119	0.096	0.148	0.00	0.09	-0.007	75	56
IATE	*N*		0.092	0.254	-	0.297	-	-	-	55	33
ATE	*1*	*dml*	0.147	0.148	0.059	0.159	-0.30	1.42	-0.003	27	11
GATE	*5*		0.147	0.165	0.127	0.195	-0.61	4.66	-0.014	66	46
IATE	*N*	*ols*	0.148	0.287	-	0.362	-	-	-	25	17
IATE	*N*	*rf*	0.125	0.344	-	0.533	-	-	-	-	-
ATE	*1*	*dml-*	0.140	0.140	0.056	0.150	0.00	-0.07	0.002	23	13
GATE	*5*	*norm*	0.140	0.157	0.123	0.187	-0.13	0.01	-0.005	74	52
IATE	*N*	*ols*	0.140	0.285	-	0.357	-	-	-	28	18
IATE	*N*	*rf*	0.117	0.349	-	0.461	-	-	-	-	-
ATE	*1*	*ols*	0.111	0.112	0.054	0.123	-0.02	0.01	-0.015	25	13
GATE	*5*		0.111	0.133	0.103	0.160	-0.04	0.03	-0.029	58	39
IATE	*N*		0.111	0.990	-	0.314	-	-	-	65	47

Note: Table to be continued.

Table C.15 - continued: Step-function IATE, strong selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.375	0.375	0.031	0.377	-0.06	0.70	0.001	0	0
GATE	*5*		0.375	0.375	0.053	0.384	-0.10	-0.10	-0.002	0	0
IATE	*N*		0.375	0.375	-	0.409	-	-	-	18	6
IATE eff	*N*		0.375	0.375	-	0.407	-	-	-	-	-
ATE	*1*	*mcf*	-0.013	0.025	0.028	0.031	0.18	0.26	0.007	96	86
GATE	*5*	*cent*	-0.014	0.085	0.045	0.100	0.01	0.03	0.007	65	41
IATE	*N*		-0.014	0.156	-	0.185	-	-	-	66	43
IATE eff	*N*		-0.014	0.151	-	0.151	-	-	-	-	-
ATE	*1*	*grf*	0.211	0.211	0.025	0.213	-0.31	-0.09	-0.004	0	0
GATE	*5*		0.211	0.211	0.052	0.220	0.00	-0.04	-0.005	3	0
IATE	*N*		0.205	0.208	-	0.253	-	-	-	59	39
ATE	*1*	*grf*	0.042	0.044	0.027	0.050	-0.00	-0.18	-0.006	46	27
GATE	*5*	*cent*	0.043	0.063	0.051	0.080	-0.09	0.03	-0.006	75	56
IATE	*N*		0.053	0.123	-	0.162	-	-	-	82	66
ATE	*1*	*dml*	0.088	0.088	0.035	0.095	0.45	2.84	-0.004	22	8
GATE	*5*		0.088	0.099	0.091	0.119	0.10	6.78	-0.015	62	42
IATE	*N*	*ols*	0.088	0.208	-	0.262	-	-	-	11	7
IATE	*N*	*rf*	0.069	0.268	-	0.479	-	-	-	-	-
ATE	*1*	*dml-*	0.079	0.079	0.032	0.085	0.12	0.21	0.000	28	12
GATE	*5*	*norm*	0.078	0.090	0.073	0.108	-0.08	0.04	-0.007	71	49
IATE	*N*	*ols*	0.079	0.205	-	0.256	-	-	-	12	8
IATE	*N*	*rf*	0.058	0.275	-	0.418	-	-	-	-	-
ATE	*1*	*ols*	0.113	0.113	0.027	0.116	0.07	-0.30	-0.008	0	0
GATE	*5*		0.113	0.118	0.052	0.134	0.05	-0.07	-0.015	27	17
IATE	*N*		0.113	0.200	-	0.253	-	-	-	46	31

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

C.2 Extensions to the base specifications

So far, we investigate the performance of the estimations when we vary the IATE and the degree of selectivity, as these dimensions are key parameters in any unconfoundedness setting. However, there are many other dimensions for which it is interesting to see if the conclusions concerning estimator behaviour change. Since it is infeasible to vary all of them simultaneously, we chose the setting with medium selectivity, the step function specification of IATE heterogeneity, as well as the other parameters from the previous specification (k=10, p=20, $R^2(y^0)$ = 10%, X^U and X^N, 2 treatment groups, a treatment share of 50%, 5 GATEs, N=2'500 and N=10'000). Then, we vary these parameters only individually, instead of varying them jointly.

Therefore, the following tables will only show the dimension that deviates from this baseline specification.

Table C.16: Fewer covariates (p = 10, k= p / 2)

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.134	0.135	0.060	0.147	-0.09	-0.04	0.005	44	21
GATE	*5*		0.133	0.144	0.091	0.177	-0.01	0.09	0.005	70	50
IATE	*N*		0.134	0.196	-	0.246	-	-	-	80	62
IATE eff	*N*		0.136	0.181	-	0.228	-	-	-	-	-
ATE	*1*	*mcf*	0.006	0.049	0.060	0.061	-0.09	0.01	0.002	96	83
GATE	*5*	*cent*	0.005	0.088	0.093	0.112	-0.02	-0.01	0.002	90	73
IATE	*N*		0.005	0.171	-	0.211	-	-	-	86	65
IATE eff	*N*		0.005	0.154	-	0.188	-	-	-	-	-
ATE	*1*	*grf*	0.043	0.050	0.044	0.061	0.05	-0.02	-0.001	82	60
GATE	*5*		0.043	0.085	0.097	0.107	-0.05	-0.01	-0.003	92	73
IATE	*N*		0.040	0.158	-	0.194	-	-	-	78	58
ATE	*1*	*grf*	0.011	0.036	0.045	0.046	0.16	0.04	-0.002	93	76
GATE	*5*	*cent*	0.012	0.079	0.098	0.099	0.03	0.12	-0.003	94	78
IATE	*N*		0.013	0.157	-	0.190	-	-	-	78	58
ATE	*1*	*dml*	0.000	0.039	0.049	0.049	-0.08	0.14	0.001	97	83
GATE	*5*		0.000	0.084	0.105	0.105	0.00	-0.02	0.003	95	81
IATE	*N*	*ols*	0.000	0.214	-	0.266	-	-	-	19	12
IATE	*N*	*rf*	-0.002	0.336	-	0.433	-	-	-	-	-
ATE	*1*	*dml-*	0.002	0.039	0.049	0.049	-0.08	0.13	0.002	96	81
GATE	*5*	*norm*	0.002	0.083	0.103	0.104	-0.01	-0.04	0.000	95	80
IATE	*N*	*ols*	0.002	0.212	-	0.264	-	-	-	17	11
IATE	*N*	*rf*	0.000	0.329	-	0.423	-	-	-	-	-
ATE	*1*	*ols*	0.009	0.037	0.046	0.047	-0.02	0.06	-0.015	81	63
GATE	*5*		0.008	0.086	0.096	0.107	0.01	0.14	-0.029	79	57
IATE	*N*		0.008	0.207	-	0.257	-	-	-	60	41

Note: Table to be continued.

Table C.16 - continued: Fewer covariates (p = 10, k= p / 2)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.092	0.092	0.030	0.097	0.09	-0.35	0.002	18	5
GATE	*5*		0.091	0.094	0.051	0.113	0.12	0.07	0.003	62	40
IATE	*N*		0.091	0.132	-	0.166	-	-	-	83	66
IATE eff	*N*		0.093	0.120	-	0.152	-	-	-	-	-
ATE	*1*	*mcf*	-0.009	0.026	0.030	0.032	0.10	-0.29	0.001	95	80
GATE	*5*	*cent*	-0.010	0.049	0.049	0.061	0.13	0.10	0.003	92	72
IATE	*N*		-0.009	0.113	-	0.140	-	-	-	90	71
IATE eff	*N*		-0.008	0.098	-	0.121	-	-	-	-	-
ATE	*1*	*grf*	0.020	0.023	0.021	0.029	-0.13	-0.04	0.000	82	62
GATE	*5*		0.019	0.042	0.049	0.053	0.08	0.03	-0.002	92	76
IATE	*N*		0.020	0.086	-	0.111	-	-	-	90	75
ATE	*1*	*grf*	0.003	0.017	0.022	0.022	0.13	0.58	-0.001	93	78
GATE	*5*	*cent*	0.003	0.040	0.049	0.050	0.02	0.12	-0.002	95	80
IATE	*N*		0.007	0.086	-	0.109	-	-	-	90	75
ATE	*1*	*dml*	-0.003	0.019	0.023	0.023	0.00	-0.45	0.003	98	82
GATE	*5*		-0.003	0.041	0.051	0.051	0.10	0.04	0.001	96	81
IATE	*N*	*ols*	-0.003	0.176	-	0.214	-	-	-	5	3
IATE	*N*	*rf*	-0.003	0.214	-	0.340	-	-	-	-	-
ATE	*1*	*dml-*	-0.003	0.019	0.023	0.023	0.00	-0.45	0.001	98	82
GATE	*5*	*norm*	-0.003	0.041	0.051	0.051	0.01	0.04	0.000	95	80
IATE	*N*	*ols*	-0.003	0.176	-	0.214	-	-	-	5	3
IATE	*N*	*rf*	-0.003	0.263	-	0.334	-	-	-	-	-
ATE	*1*	*ols*	0.004	0.017	0.021	0.021	-0.01	-0.23	-0.006	82	65
GATE	*5*		0.003	0.053	0.048	0.065	0.09	0.13	-0.014	67	50
IATE	*N*		0.003	0.178	-	0.215	-	-	-	35	23

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.17: More covariates (p = 50, k= p / 2)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.223	0.223	0.061	0.231	-0.02	0.05	0.002	5	1
GATE	*5*		0.223	0.224	0.085	0.253	-0.04	0.22	-0.001	33	14
IATE	*N*		0.223	0.244	-	0.304	-	-	-	53	40
IATE eff	*N*		0.223	0.238	-	0.297	-	-	-	-	-
ATE	*1*	*mcf*	0.124	0.125	0.060	0.138	0.02	-0.03	0.001	48	22
GATE	*5*	*cent*	0.124	0.136	0.085	0.174	-0.01	0.22	-0.006	66	46
IATE	*N*		0.124	0.196	-	0.247	-	-	-	60	43
IATE eff	*N*		0.125	0.189	-	0.239	-	-	-	-	-
ATE	*1*	*grf*	0.178	0.178	0.043	0.183	-0.01	0.06	0.000	2	0
GATE	*5*		0.178	0.180	0.095	0.202	-0.08	0.10	0.000	53	27
IATE	*N*		0.172	0.252	-	0.308	-	-	-	52	35
ATE	*1*	*grf*	0.107	0.107	0.042	0.115	0.01	-0.03	0.001	29	11
GATE	*5*	*cent*	0.107	0.119	0.093	0.142	0.03	-0.06	0.000	79	55
IATE	*N*		0.102	0.231	-	0.271	-	-	-	53	32
ATE	*1*	*dml*	0.092	0.093	0.045	0.103	-0.04	-0.07	0.003	52	25
GATE	*5*		0.093	0.111	0.098	0.135	-0.03	0.09	-0.003	83	61
IATE	*N*	*ols*	0.093	0.308	-	0.386	-	-	-	38	26
IATE	*N*	*rf*	0.084	0.250	-	0.314	-	-	-	-	-
ATE	*1*	*dml-*	0.092	0.092	0.046	0.102	-0.06	-0.06	0.003	53	25
GATE	*5*	*norm*	0.092	0.111	0.099	0.135	-0.02	0.09	-0.002	84	62
IATE	*N*	*ols*	0.092	0.310	-	0.389	-	-	-	39	26
IATE	*N*	*rf*	0.083	0.251	-	0.316	-	-	-	-	-
ATE	*1*	*ols*	0.012	0.039	0.048	0.049	-0.09	-0.02	-0.016	77	58
GATE	*5*		0.012	0.080	0.099	0.101	0.00	0.09	-0.032	81	62
IATE	*N*		0.012	0.295	-	0.369	-	-	-	75	55

Note: Table to be continued.

Table C.17 - continued: More covariates (p = 50, k= p / 2)

			Estimation of effects							**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.201	0.201	0.031	0.204	0.01	-0.07	0.001	0	0
GATE	*5*		0.202	0.202	0.047	0.213	0.05	-0.08	-0.002	1	0
IATE	*N*		0.202	0.205	-	0.240	-	-	-	38	23
IATE eff	*N*		0.202	0.204	-	0.235	-	-	-	-	-
ATE	*1*	*mcf*	0.093	0.093	0.029	0.097	0.14	0.12	0.001	11	3
GATE	*5*	*cent*	0.093	0.094	0.045	0.116	0.11	-0.03	-0.003	50	31
IATE	*N*		0.093	0.130	-	0.164	-	-	-	59	43
IATE eff	*N*		0.093	0.164	-	0.159	-	-	-	-	-
ATE	*1*	*grf*	0.144	0.144	0.020	0.145	-0.17	0.29	0.001	0	0
GATE	*5*		0.144	0.144	0.046	0.152	0.05	-0.09	0.001	13	3
IATE	*N*		0.141	0.148	-	0.188	-	-	-	66	50
ATE	*1*	*grf*	0.073	0.073	0.020	0.076	-0.10	0.06	0.000	6	2
GATE	*5*	*cent*	0.072	0.074	0.045	0.085	0.07	-0.24	0.001	66	39
IATE	*N*		0.068	0.111	-	0.142	-	-	-	80	63
ATE	*1*	*dml*	0.070	0.070	0.021	0.073	0.10	0.03	0.003	14	4
GATE	*5*		0.070	0.073	0.049	0.085	-0.03	0.00	-0.001	66	42
IATE	*N*	*ols*	0.070	0.208	-	0.259	-	-	-	14	9
IATE	*N*	*rf*	0.062	0.183	-	0.232	-	-	-	-	-
ATE	*1*	*dml-*	0.068	0.068	0.021	0.071	0.11	0.02	0.003	16	4
GATE	*5*	*norm*	0.069	0.072	0.049	0.084	-0.03	0.01	-0.001	69	44
IATE	*N*	*ols*	0.068	0.208	-	0.260	-	-	-	15	10
IATE	*N*	*rf*	0.055	0.185	-	0.234	-	-	-	-	-
ATE	*1*	*ols*	0.011	0.020	0.022	0.024	0.03	-0.19	-0.006	79	56
GATE	*5*		0.011	0.041	0.041	0.051	-0.02	0.09	-0.014	79	60
IATE	*N*		0.011	0.200	-	0.248	-	-	-	59	41

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.18: Sparser model (p = 20, k= 4)

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.132	0.133	0.061	0.146	0.04	0.01	0.003	47	22
GATE	*5*		0.132	0.142	0.091	0.175	0.01	-0.07	0.001	69	49
IATE	*N*		0.132	0.189	-	0.237	-	-	-	74	57
IATE eff	*N*		0.135	0.189	-	0.237	-	-	-	-	-
ATE	*1*	*mcf*	0.011	0.050	0.063	0.063	0.07	0.08	0.000	95	79
GATE	*5*	*cent*	0.010	0.091	0.093	0.115	0.05	-0.02	-0.004	87	68
IATE	*N*		0.010	0.172	-	0.209	-	-	-	77	55
IATE eff	*N*		0.013	0.160	-	0.192	-	-	-	-	-
ATE	*1*	*grf*	0.043	0.051	0.044	0.062	-0.09	-0.03	-0.001	80	58
GATE	*5*		0.043	0.085	0.096	0.107	0.07	0.06	-0.003	91	74
IATE	*N*		0.040	0.184	-	0.220	-	-	-	69	48
ATE	*1*	*grf*	0.022	0.040	0.044	0.049	0.03	0.04	-0.002	91	73
GATE	*5*	*cent*	0.022	0.079	0.095	0.099	0.00	-0.01	-0.001	94	77
IATE	*N*		0.021	0.180	-	0.214	-	-	-	70	49
ATE	*1*	*dml*	0.000	0.036	0.045	0.045	-0.01	-0.14	0.005	97	85
GATE	*5*		0.000	0.082	0.102	0.102	-0.10	-0.04	0.001	95	80
IATE	*N*	*ols*	0.000	0.240	-	0.299	-	-	-	28	17
IATE	*N*	*rf*	-0.004	0.292	-	0.375	-	-	-	-	-
ATE	*1*	*dml-*	0.002	0.036	0.045	0.045	-0.01	-0.14	0.004	97	84
GATE	*5*	*norm*	0.002	0.081	0.101	0.102	-0.10	-0.02	0.000	95	80
IATE	*N*	*ols*	0.002	0.238	-	0.297	-	-	-	25	16
IATE	*N*	*rf*	-0.003	0.288	-	0.370	-	-	-	-	-
ATE	*1*	*ols*	0.002	0.030	0.044	0.044	-0.06	-0.14	-0.012	84	63
GATE	*5*		0.002	0.086	0.096	0.107	-0.03	-0.01	-0.028	78	58
IATE	*N*		0.002	0.232	-	0.289	-	-	-	67	47

Note: Table to be continued.

Table C.18 - continued: Sparser model (p = 20, k= 4)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.097	0.097	0.029	0.102	0.16	-0.31	0.003	14	2
GATE	*5*		0.097	0.099	0.050	0.119	0.08	-0.27	0.001	56	36
IATE	*N*		0.097	0.127	-	0.161	-	-	-	75	57
IATE eff	*N*		0.097	0.119	-	0.151	-	-	-	-	-
ATE	*1*	*mcf*	-0.002	0.023	0.029	0.029	0.07	-0.19	0.003	98	84
GATE	*5*	*cent*	-0.002	0.049	0.049	0.062	0.12	-0.13	0.002	89	71
IATE	*N*		-0.003	0.106	-	0.131	-	-	-	83	61
IATE eff	*N*		-0.002	0.098	-	0.119	-	-	-	-	-
ATE	*1*	*grf*	0.022	0.026	0.021	0.031	-0.15	-0.25	-0.000	81	57
GATE	*5*		0.022	0.043	0.049	0.054	0.05	-0.07	-0.002	92	73
IATE	*N*		0.023	0.091	-	0.118	-	-	-	87	72
ATE	*1*	*grf*	0.007	0.019	0.023	0.024	0.19	0.07	-0.002	91	74
GATE	*5*	*cent*	0.008	0.040	0.048	0.050	0.01	-0.05	-0.001	93	77
IATE	*N*		0.010	0.091	-	0.114	-	-	-	88	71
ATE	*1*	*dml*	0.001	0.017	0.022	0.022	-0.14	0.63	0.003	96	86
GATE	*5*		0.000	0.040	0.050	0.051	0.09	0.06	0.001	95	80
IATE	*N*	*ols*	0.000	0.182	-	0.223	-	-	-	8	5
IATE	*N*	*rf*	-0.002	0.218	-	0.282	-	-	-	-	-
ATE	*1*	*dml-*	0.001	0.017	0.022	0.022	-0.15	0.62	0.003	96	87
GATE	*5*	*norm*	0.001	0.040	0.050	0.050	0.09	0.06	0.000	95	79
IATE	*N*	*ols*	0.001	0.182	-	0.223	-	-	-	8	5
IATE	*N*	*rf*	-0.002	0.216	-	0.279	-	-	-	-	-
ATE	*1*	*ols*	0.004	0.017	0.021	0.022	-0.09	0.30	-0.006	84	66
GATE	*5*		0.004	0.056	0.052	0.068	-0.01	0.08	-0.014	64	44
IATE	*N*		0.004	0.184	-	0.225	-	-	-	44	29

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.19: Less sparse model (p = 20, k= 16)

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.192	0.192	0.061	0.201	0.08	0.23	0.002	14	4
GATE	*5*		0.191	0.193	0.089	0.224	0.07	0.17	-0.003	47	25
IATE	*N*		0.191	0.225	-	0.279	-	-	-	64	48
IATE eff	*N*		0.194	0.216	-	0.270	-	-	-	-	-
ATE	*1*	*mcf*	0.053	0.065	0.060	0.080	0.05	0.43	0.001	87	65
GATE	*5*	*cent*	0.052	0.096	0.088	0.126	0.05	0.29	-0.005	82	65
IATE	*N*		0.052	0.174	-	0.214	-	-	-	75	53
IATE eff	*N*		0.055	0.164	-	0.200	-	-	-	-	-
ATE	*1*	*grf*	0.125	0.125	0.044	0.133	-0.01	-0.09	-0.002	17	5
GATE	*5*		0.125	0.134	0.097	0.159	0.00	-0.03	-0.003	72	48
IATE	*N*		0.119	0.193	-	0.240	-	-	-	70	51
ATE	*1*	*grf*	0.048	0.054	0.042	0.064	0.04	-0.18	-0.000	79	55
GATE	*5*	*cent*	0.047	0.085	0.093	0.106	-0.01	0.09	-0.001	92	74
IATE	*N*		0.044	0.174	-	0.209	-	-	-	73	52
ATE	*1*	*dml*	0.032	0.044	0.044	0.055	-0.02	-0.15	0.005	93	73
GATE	*5*		0.032	0.083	0.098	0.103	-0.02	-0.05	0.000	93	79
IATE	*N*	*ols*	0.032	0.236	-	0.295	-	-	-	24	16
IATE	*N*	*rf*	0.024	0.282	-	0.359	-	-	-	-	-
ATE	*1*	*dml-*	0.032	0.044	0.044	0.055	-0.02	-0.17	0.004	92	73
GATE	*5*	*norm*	0.032	0.083	0.099	0.104	-0.01	-0.06	0.000	93	78
IATE	*N*	*ols*	0.032	0.236	-	0.295	-	-	-	24	16
IATE	*N*	*rf*	0.024	0.283	-	0.359	-	-	-	-	-
ATE	*1*	*ols*	0.008	0.036	0.044	0.045	-0.07	-0.19	-0.013	84	61
GATE	*5*		0.008	0.078	0.094	0.097	-0.05	0.03	-0.027	82	62
IATE	*N*		0.008	0.230	-	0.286	-	-	-	67	48

Note: Table to be continued.

Table C.19 - continued: Less sparse model (p = 20, k= 16)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.168	0.168	0.029	0.170	0.06	0.07	0.003	0	0
GATE	*5*		0.167	0.167	0.047	0.179	0.15	0.04	-0.001	7	2
IATE	*N*		0.167	0.176	-	0.210	-	-	-	57	39
IATE eff	*N*		0.167	0.172	-	0.203	-	-	-	-	-
ATE	*1*	*mcf*	0.029	0.032	0.026	0.039	-0.05	0.10	0.004	90	65
GATE	*5*	*cent*	0.028	0.051	0.044	0.069	0.15	0.12	0.001	84	67
IATE	*N*		0.028	0.107	-	0.133	-	-	-	80	59
IATE eff	*N*		0.028	0.100	-	0.124	-	-	-	-	-
ATE	*1*	*grf*	0.093	0.093	0.020	0.095	-0.15	-0.44	0.001	0	0
GATE	*5*		0.092	0.093	0.047	0.104	0.04	-0.11	-0.001	49	25
IATE	*N*		0.089	0.109	-	0.141	-	-	-	82	66
ATE	*1*	*grf*	0.024	0.027	0.021	0.032	-0.05	0.06	0.000	80	51
GATE	*5*	*cent*	0.025	0.043	0.046	0.054	-0.04	-0.07	-0.000	90	72
IATE	*N*		0.024	0.089	-	0.112	-	-	-	89	73
ATE	*1*	*dml*	0.020	0.025	0.22	0.030	-0.31	1.16	0.002	89	68
GATE	*5*		0.020	0.042	0.49	0.053	0.07	-0.03	0.000	93	77
IATE	*N*	*ols*	0.020	0.181	-	0.222	-	-	-	7	5
IATE	*N*	*rf*	0.014	0.212	-	0.271	-	-	-	-	-
ATE	*1*	*dml-*	0.020	0.025	0.22	0.030	-0.33	1.20	0.002	89	70
GATE	*5*	*norm*	0.020	0.042	0.49	0.053	0.07	-0.03	0.000	93	77
IATE	*N*	*ols*	0.020	0.181	-	0.222	-	-	-	7	5
IATE	*N*	*rf*	0.014	0.212	-	0.272	-	-	-	-	-
ATE	*1*	*ols*	0.009	0.018	0.021	0.023	-0.11	0.33	-0.006	79	64
GATE	*5*		0.009	0.043	0.047	0.053	0.05	0.07	-0.014	77	57
IATE	*N*		0.009	0.181	-	0.221	-	-	-	45	30

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs.

Table C.20: Covariates irrelevant for Y^0 ($R^2(y^0)$ = 0%)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.024	0.050	0.059	0.064	0.10	0.07	0.00	93	76
GATE	*5*		0.023	0.089	0.087	0.115	0.05	0.04	-0.003	86	68
IATE	*N*		0.023	0.168	-	0.206	-	-	-	79	58
IATE eff	*N*		0.026	0.157	-	0.192	-	-	-	-	-
ATE	*1*	*mcf*	-0.034	0.056	0.060	0.069	0.059	0.184	0.000	92	73
GATE	*5*	*cent*	-0.034	0.098	0.089	0.120	0.044	0.124	-0.005	81	61
IATE	*N*		-0.034	0.176	-	0.213	-	-	-	75	53
IATE eff	*N*		-0.031	0.165	-	0.197	-	-	-	-	-
ATE	*1*	*grf*	0.013	0.036	0.043	0.045	-0.02	-0.27	-0.002	93	73
GATE	*5*		0.013	0.077	0.094	0.096	0.06	0.03	-0.003	93	78
IATE	*N*		0.009	0.161	-	0.197	-	-	-	75	56
ATE	*1*	*grf*	-0.011	0.035	0.042	0.044	0.03	-0.22	-0.001	94	78
GATE	*5*	*cent*	-0.010	0.075	0.093	0.094	-0.01	0.06	-0.001	94	78
IATE	*N*		-0.012	0.164	-	0.198	-	-	-	74	55
ATE	*1*	*dml*	-0.002	0.037	0.045	0.045	-0.04	-0.23	0.004	97	83
GATE	*5*		-0.002	0.078	0.098	0.098	-0.04	-0.06	0.000	95	80
IATE	*N*	*ols*	-0.001	0.247	-	0.307	-	-	-	22	15
IATE	*N*	*rf*	-0.002	0.283	-	0.362	-	-	-	-	-
ATE	*1*	*dml-*	-0.001	0.037	0.045	0.045	-0.04	-0.26	0.003	97	83
GATE	*5*	*norm*	-0.001	0.078	0.098	0.098	-0.04	-0.06	0.000	95	80
IATE	*N*	*ols*	-0.001	0.246	-	0.307	-	-	-	22	15
IATE	*N*	*rf*	-0.001	0.282	-	0.360	-	-	-	-	-
ATE	*1*	*ols*	-0.002	0.036	0.044	0.044	-0.06	-0.22	-0.013	83	62
GATE	*5*		-0.002	0.076	0.093	0.094	-0.07	0.01	-0.027	84	63
IATE	*N*		-0.002	0.241	-	0.299	-	-	-	64	45

Note: Table to be continued.

Table C.20 - continued: Covariates irrelevant for Y^0 ($R^2(y^0)$ = 0%)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.016	0.026	0.027	0.032	0.27	0.04	0.003	94	79
GATE	*5*		0.015	0.046	0.044	0.061	0.25	0.20	0.002	89	72
IATE	*N*		0.015	0.101	-	0.128	-	-	-	84	72
IATE eff	*N*		0.015	0.093	-	0.117	-	-	-	-	-
ATE	*1*	*mcf*	-0.035	0.038	0.03	0.044	0.24	0.24	0.004	81	57
GATE	*5*	*cent*	-0.036	0.062	0.04	0.073	0.21	0.19	0.002	76	50
IATE	*N*		-0.036	0.111	-	0.137	-	-	-	79	57
IATE eff	*N*		-0.036	0.104	-	0.127	-	-	-	-	-
ATE	*1*	*grf*	0.006	0.017	0.020	0.021	-0.13	-0.35	0.001	95	77
GATE	*5*		0.006	0.038	0.047	0.047	-0.01	-0.04	-0.001	95	78
IATE	*N*		0.005	0.089	-	0.114	-	-	-	89	73
ATE	*1*	*grf*	-0.013	0.020	0.022	0.026	-0.16	0.31	-0.001	90	67
GATE	*5*	*cent*	-0.013	0.040	0.047	0.050	-0.06	0.01	-0.001	92	77
IATE	*N*		-0.012	0.095	-	0.117	-	-	-	87	69
ATE	*1*	*dml*	-0.001	0.016	0.022	0.022	-0.25	1.08	0.003	96	84
GATE	*5*		-0.001	0.039	0.049	0.050	0.09	0.01	-0.001	94	80
IATE	*N*	*ols*	-0.001	0.197	-	0.241	-	-	-	7	4
IATE	*N*	*rf*	-0.001	0.213	-	0.273	-	-	-	-	-
ATE	*1*	*dml-*	-0.001	0.017	0.022	0.022	-0.28	1.13	0.002	96	85
GATE	*5*	*norm*	-0.001	0.039	0.049	0.050	0.09	0.01	-0.001	95	80
IATE	*N*	*ols*	-0.001	0.197	-	0.240	-	-	-	7	4
IATE	*N*	*rf*	-0.001	0.212	-	0.273	-	-	-	-	-
ATE	*1*	*ols*	-0.001	0.016	0.021	0.021	0.05	0.13	-0.006	84	66
GATE	*5*		-0.001	0.039	0.047	0.049	0.07	0.20	-0.014	82	62
IATE	*N*		-0.001	0.049	-	0.196	-	-	-	40	27

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.21: Covariates very important for Y^0 ($R^2(y^0)$ = 45%)

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.421	0.421	0.074	0.427	0.09	0.17	0.016	0	0
GATE	*5*		0.420	0.420	0.111	0.447	0.06	-0.07	0.007	9	2
IATE	*N*		0.420	0.427	-	0.497	-	-	-	45	31
IATE eff	*N*		0.422	0.427	-	0.488	-	-	-	-	-
ATE	*1*	*mcf*	0.165	0.165	0.068	0.178	0.13	0.31	0.002	33	12
GATE	*5*	*cent*	0.164	0.173	0.104	0.215	0.09	0.06	-0.007	63	42
IATE	*N*		0.164	0.248	-	0.308	-	-	-	65	47
IATE eff	*N*		0.164	0.237	-	0.295	-	-	-	-	-
ATE	*1*	*grf*	0.228	0.228	0.050	0.234	-0.02	-0.20	-0.000	1	0
GATE	*5*		0.228	0.230	0.110	0.254	-0.01	-0.03	-0.000	45	22
IATE	*N*		0.218	0.253	-	0.323	-	-	-	69	50
ATE	*1*	*grf*	0.117	0.117	0.047	0.126	0.01	-0.06	-0.000	29	12
GATE	*5*	*cent*	0.117	0.131	0.102	0.158	0.02	0.02	-0.000	78	54
IATE	*N*		0.109	0.193	-	0.246	-	-	-	77	58
ATE	*1*	*dml*	0.061	0.067	0.050	0.079	-0.04	-0.20	0.006	84	57
GATE	*5*		0.061	0.101	0.110	0.126	-0.02	0.00	0.004	92	74
IATE	*N*	*ols*	0.062	0.318	-	0.395	-	-	-	23	15
IATE	*N*	*rf*	0.041	0.340	-	0.434	-	-	-	-	-
ATE	*1*	*dml-*	0.064	0.068	0.050	0.081	-0.05	-0.20	0.005	82	54
GATE	*5*	*norm*	0.063	0.102	0.110	0.127	-0.03	0.00	0.003	92	74
IATE	*N*	*ols*	0.064	0.317	-	0.395	-	-	-	23	15
IATE	*N*	*rf*	0.044	0.339	-	0.432	-	-	-	-	-
ATE	*1*	*ols*	0.022	0.044	0.049	0.054	-0.09	-0.14	-0.015	78	56
GATE	*5*		0.022	0.105	0.103	0.129	-0.04	-0.02	-0.029	73	51
IATE	*N*		0.022	0.314	-	0.388	-	-	-	56	38

Note: Table to be continued.

Table C.21 - continued: Covariates very important for Y^0 ($R^2(y^0) = 45\%$)

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.346	0.346	0.034	0.347	-0.09	-0.12	0.011	0	0
GATE	*5*		0.345	0.345	0.059	0.356	0.04	-0.12	0.004	0	0
IATE	*N*		0.344	0.347	-	0.387	-	-	-	35	20
IATE eff	*N*		0.345	0.346	-	0.381	-	-	-	-	-
ATE	*1*	*mcf*	0.104	0.104	0.029	0.108	0.08	-0.05	0.005	10	2
GATE	*5*	*cent*	0.103	0.105	0.050	0.126	0.17	0.01	0.001	53	32
IATE	*N*		0.103	0.158	-	0.194	-	-	-	67	48
IATE eff	*N*		0.103	0.153	-	0.187	-	-	-	-	-
ATE	*1*	*grf*	0.152	0.152	0.022	0.153	-0.20	-0.51	0.002	0	0
GATE	*5*		0.151	0.151	0.052	0.161	-0.05	-0.18	0.001	20	6
IATE	*N*		0.147	0.162	-	0.200	-	-	-	72	53
ATE	*1*	*grf*	0.060	0.060	0.024	0.064	-0.00	-0.20	-0.001	29	9
GATE	*5*	*cent*	0.061	0.068	0.051	0.081	-0.04	-0.02	-0.001	74	50
IATE	*N*		0.061	0.118	-	0.146	-	-	-	83	64
ATE	*1*	*dml*	0.036	0.037	0.023	0.043	-0.14	0.80	0.004	77	46
GATE	*5*		0.036	0.052	0.053	0.065	0.20	0.10	0.001	90	73
IATE	*N*	*ols*	0.036	0.267	-	0.325	-	-	-	6	4
IATE	*N*	*rf*	0.021	0.325	-	0.320	-	-	-	-	-
ATE	*1*	*dml-*	0.037	0.038	0.023	0.043	-0.14	0.81	0.004	77	44
GATE	*5*	*norm*	0.036	0.052	0.053	0.065	0.19	0.09	0.001	89	72
IATE	*N*	*ols*	0.036	0.267	-	0.325	-	-	-	6	4
IATE	*N*	*rf*	0.022	0.319	-	0.319	-	-	-	-	-
ATE	*1*	*ols*	0.023	0.027	0.022	0.032	0.21	0.34	-0.005	69	49
GATE	*5*		0.022	0.080	0.051	0.095	0.11	0.28	0.109	48	30
IATE	*N*		0.022	0.273	-	0.332	-	-	-	32	21

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.22: Uniformly distributed covariates (X^U)

			Estimation of effects						**Estimation of std. errors**		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.179	0.179	0.058	0.188	0.115	-0.24	0.006	19	5
GATE	*5*		0.178	0.180	0.088	0.212	0.030	-0.12	0.000	53	31
IATE	*N*		0.179	0.214	-	0.267	-	-	-	67	50
IATE eff	*N*		0.180	0.205	-	0.256	-	-	-	-	-
ATE	*1*	*mcf*	0.038	0.046	0.056	0.068	0.12	-0.24	0.005	92	75
GATE	*5*	*cent*	0.038	0.092	0.087	0.118	-0.01	-0.10	-0.003	85	67
IATE	*N*		0.039	0.171	-	0.209	-	-	-	76	55
IATE eff	*N*		0.039	0.161	-	0.195	-	-	-	-	-
ATE	*1*	*grf*	0.095	0.095	0.042	0.103	0.03	-0.30	0.000	39	17
GATE	*5*		0.094	0.110	0.093	0.133	-0.02	0.09	0.000	83	59
IATE	*N*		0.089	0.188	-	0.230	-	-	-	71	50
ATE	*1*	*grf*	0.034	0.044	0.043	0.055	0.02	0.22	-0.001	86	67
GATE	*5*	*cent*	0.035	0.081	0.094	0.102	-0.03	-0.01	-0.002	93	76
IATE	*N*		0.032	0.174	-	0.209	-	-	-	72	51
ATE	*1*	*dml*	0.019	0.039	0.044	0.048	-0.07	-0.09	0.005	96	81
GATE	*5*		0.019	0.082	0.101	0.103	-0.01	-0.13	-0.001	94	78
IATE	*N*	*ols*	0.019	0.227	-	0.283	-	-	-	26	17
IATE	*N*	*rf*	0.012	0.285	-	0.363	-	-	-	-	-
ATE	*1*	*dml-*	0.020	0.040	0.044	0.049	-0.06	-0.12	0.005	96	80
GATE	*5*	*norm*	0.020	0.082	0.101	0.103	-0.01	-0.13	-0.001	94	78
IATE	*N*	*ols*	0.020	0.227	-	0.283	-	-	-	26	17
IATE	*N*	*rf*	0.013	0.284	-	0.362	-	-	-	-	-
ATE	*1*	*ols*	0.007	0.035	0.043	0.044	-0.03	-0.10	-0.012	83	64
GATE	*5*		0.007	0.082	0.096	0.102	-0.01	-0.13	-0.011	80	60
IATE	*N*		0.008	0.220	-	0.273	-	-	-	70	50

Note: Table to be continued.

Table C.22 - continued: Uniformly distributed covariates (X^U)

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						Estimation of std. errors		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.145	0.145	0.028	0.148	0.01	-0.58	0.004	0	0
GATE	*5*		0.145	0.145	0.046	0.158	0.01	-0.31	0.001	22	5
IATE	*N*		0.145	0.158	-	0.192	-	-	-	63	45
IATE eff	*N*		0.144	0.152	-	0.184	-	-	-	-	-
ATE	*1*	*mcf*	0.014	0.025	0.028	0.031	0.02	-0.40	0.003	95	78
GATE	*5*	*cent*	0.013	0.048	0.044	0.063	0.00	-0.23	0.002	87	68
IATE	*N*		0.013	0.107	-	0.131	-	-	-	81	59
IATE eff	*N*		0.011	0.099	-	0.120	-	-	-	-	-
ATE	*1*	*grf*	0.062	0.062	0.022	0.066	0.02	-0.32	-0.001	18	6
GATE	*5*		0.062	0.065	0.046	0.077	0.00	-0.17	0.001	73	49
IATE	*N*		0.060	0.095	-	0.124	-	-	-	87	72
ATE	*1*	*grf*	0.014	0.022	0.023	0.027	0.03	0.12	-0.002	88	66
GATE	*5*	*cent*	0.015	0.041	0.048	0.051	0.01	-0.06	-0.002	92	76
IATE	*N*		0.015	0.090	-	0.113	-	-	-	88	72
ATE	*1*	*dml*	0.009	0.020	0.023	0.025	-0.02	-0.21	0.002	96	79
GATE	*5*		0.009	0.040	0.049	0.050	-0.01	-0.06	0.000	95	79
IATE	*N*	*ols*	0.009	0.168	-	0.205	-	-	-	8	5
IATE	*N*	*rf*	0.004	0.215	-	0.276	-	-	-	-	-
ATE	*1*	*dml-*	0.009	0.020	0.023	0.025	-0.03	-0.17	0.002	96	79
GATE	*5*	*norm*	0.009	0.040	0.049	0.050	-0.01	-0.05	0.000	94	79
IATE	*N*	*ols*	0.010	0.168	-	0.205	-	-	-	8	5
IATE	*N*	*rf*	0.004	0.214	-	0.275	-	-	-	-	-
ATE	*1*	*ols*	0.008	0.019	0.022	0.023	-0.06	0.08	-0.007	82	62
GATE	*5*		0.007	0.047	0.048	0.058	-0.03	-0.17	-0.015	73	52
IATE	*N*		0.007	0.168	-	0.204	-	-	-	47	31

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.23: Normally distributed covariates (X^N)

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.191	0.191	0.062	0.201	-0.06	0.14	0.002	15	4
GATE	*5*		0.191	0.194	0.089	0.226	-0.12	0.17	-0.001	49	26
IATE	*N*		0.190	0.226	-	0.281	-	-	-	64	48
IATE eff	*N*		0.191	0.226	-	0.270	-	-	-	-	-
ATE	*1*	*mcf*	0.044	0.061	0.060	0.075	-0.10	-0.01	0.001	89	69
GATE	*5*	*cent*	0.044	0.097	0.088	0.125	-0.12	0.08	-0.004	82	64
IATE	*N*		0.044	0.176	-	0.214	-	-	-	75	53
IATE eff	*N*		0.045	0.165	-	0.200	-	-	-	-	-
ATE	*1*	*grf*	0.107	0.108	0.044	0.116	-0.12	0.01	-0.001	27	11
GATE	*5*		0.107	0.121	0.096	0.145	-0.05	-0.08	-0.002	78	54
IATE	*N*		0.102	0.193	-	0.238	-	-	-	70	50
ATE	*1*	*grf*	0.041	0.048	0.041	0.058	0.08	-0.03	0.000	84	62
GATE	*5*	*cent*	0.041	0.084	0.095	0.105	-0.03	-0.12	-0.002	92	74
IATE	*N*		0.039	0.179	-	0.214	-	-	-	71	50
ATE	*1*	*dml*	0.026	0.042	0.046	0.053	-0.06	0.01	0.003	93	77
GATE	*5*		0.026	0.083	0.101	0.104	-0.03	-0.03	-0.001	94	78
IATE	*N*	*ols*	0.026	0.240	-	0.300	-	-	-	24	16
IATE	*N*	*rf*	0.017	0.286	-	0.365	-	-	-	-	-
ATE	*1*	*dml-*	0.027	0.043	0.046	0.053	-0.06	0.02	0.003	93	77
GATE	*5*	*norm*	0.027	0.083	0.101	0.104	-0.03	-0.03	-0.001	94	78
IATE	*N*	*ols*	0.027	0.239	-	0.299	-	-	-	24	16
IATE	*N*	*rf*	0.017	0.285	-	0.363	-	-	-	-	-
ATE	*1*	*ols*	0.012	0.038	0.046	0.047	-0.02	0.25	-0.014	81	61
GATE	*5*		0.011	0.083	0.097	0.103	-0.04	-0.04	-0.029	80	58
IATE	*N*		0.011	0.233	-	0.291	-	-	-	66	47

Note: Table to be continued.

Table C.23 - continued: Normally distributed covariates (X^N)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.156	0.156	0.030	0.159	-0.15	-0.03	0.002	0	0
GATE	*5*		0.156	0.156	0.049	0.170	-0.08	-0.16	-0.001	16	5
IATE	*N*		0.155	0.168	-	0.205	-	-	-	60	43
IATE eff	*N*		0.155	0.163	-	0.198	-	-	-	-	-
ATE	*1*	*mcf*	0.018	0.028	0.029	0.034	-0.08	-0.29	0.001	93	73
GATE	*5*	*cent*	0.018	0.051	0.046	0.067	0.03	-0.09	0.000	84	67
IATE	*N*		0.018	0.110	-	0.136	-	-	-	80	67
IATE eff	*N*		0.017	0.102	-	0.125	-	-	-	-	-
ATE	*1*	*grf*	0.073	0.073	0.021	0.076	-0.04	0.24	-0.000	8	1
GATE	*5*		0.073	0.076	0.047	0.088	0.04	0.07	-0.001	66	39
IATE	*N*		0.071	0.103	-	0.133	-	-	-	84	68
ATE	*1*	*grf*	0.019	0.026	0.024	0.031	-0.30	0.01	-0.003	82	56
GATE	*5*	*cent*	0.019	0.042	0.048	0.053	0.08	-0.11	-0.002	92	73
IATE	*N*		0.021	0.089	-	0.113	-	-	-	89	73
ATE	*1*	*dml*	0.016	0.022	0.023	0.028	0.07	-0.07	0.002	92	71
GATE	*5*		0.015	0.042	0.050	0.053	0.05	-0.05	-0.001	93	78
IATE	*N*	*ols*	0.015	0.185	-	0.227	-	-	-	7	5
IATE	*N*	*rf*	0.009	0.216	-	0.276	-	-	-	-	-
ATE	*1*	*dml-*	0.016	0.022	0.023	0.028	0.07	-0.05	0.002	93	70
GATE	*5*	*norm*	0.016	0.042	0.050	0.053	0.06	-0.06	-0.001	93	77
IATE	*N*	*ols*	0.015	0.185	-	0.227	-	-	-	7	5
IATE	*N*	*rf*	0.009	0.216	-	0.276	-	-	-	-	-
ATE	*1*	*ols*	0.013	0.021	0.022	0.025	-0.05	-0.23	-0.006	75	55
GATE	*5*		0.012	0.050	0.048	0.062	0.02	-0.16	-0.014	70	51
IATE	*N*		0.012	0.186	-	0.227	-	-	-	43	29

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs.

Table C.24: Normally, uniformly distributed, and indicator covariates (X^D, X^N, X^U)

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.160	0.161	0.062	0.172	0.21	0.30	0.003	30	10
GATE	*5*		0.151	0.157	0.086	0.192	0.08	-0.01	-0.001	60	39
IATE	*N*		0.160	0.200	-	0.247	-	-	-	77	58
IATE eff	*N*		0.163	0.189	-	0.231	-	-	-	-	-
ATE	*1*	*mcf*	0.054	0.066	0.062	0.082	0.18	0.46	0.001	87	65
GATE	*5*	*cent*	0.045	0.094	0.158	0.122	0.04	-0.01	-0.005	83	65
IATE	*N*		0.054	0.158	-	0.197	-	-	-	86	65
IATE eff	*N*		0.057	0.143	-	0.175	-	-	-	-	-
ATE	*1*	*grf*	0.092	0.093	0.044	0.102	-0.03	-0.20	-0.001	41	21
GATE	*5*		0.093	0.112	0.099	0.137	0.03	-0.09	-0.003	82	60
IATE	*N*		0.093	0.180	-	0.219	-	-	-	74	55
ATE	*1*	*grf*	0.052	0.057	0.044	0.068	0.11	-0.11	-0.001	75	51
GATE	*5*	*cent*	0.054	0.089	0.097	0.113	0.05	-0.01	-0.002	90	73
IATE	*N*		0.049	0.162	-	0.198	-	-	-	78	59
ATE	*1*	*dml*	0.047	0.054	0.045	0.065	0.08	-0.21	0.004	86	64
GATE	*5*		0.049	0.092	0.103	0.115	-0.02	-0.01	0.001	93	75
IATE	*N*	*ols*	0.047	0.203	-	0.255	-	-	-	30	20
IATE	*N*	*rf*	0.047	0.320	-	0.408	-	-	-	-	-
ATE	*1*	*dml-*	0.048	0.055	0.045	0.066	0.08	-0.22	0.004	84	63
GATE	*5*	*norm*	0.050	0.092	0.103	0.115	-0.02	-0.02	0.000	92	75
IATE	*N*	*ols*	0.049	0.202	-	0.254	-	-	-	29	20
IATE	*N*	*rf*	0.040	0.318	-	0.406	-	-	-	-	-
ATE	*1*	*ols*	0.001	0.036	0.045	0.045	0.06	-0.21	0.045	82	62
GATE	*5*		0.009	0.083	0.096	0.104	0.01	-0.05	0.096	80	60
IATE	*N*		0.001	0.188	-	0.236	-	-	-	77	57

Note: Table to be continued.

Table C.24 - continued: Normally, uniformly distributed, and indicator covariates (X^D, X^N, X^U)

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.120	0.120	0.028	0.124	0.32	0.41	0.004	2	0
GATE	*5*		0.115	0.116	0.047	0.136	0.16	-0.07	0.000	41	24
IATE	*N*		0.120	0.145	-	0.178	-	-	-	78	59
IATE eff	*N*		0.120	0.135	-	0.165	-	-	-	-	-
ATE	*1*	*mcf*	0.022	0.028	0.027	0.035	0.13	0.29	0.004	93	75
GATE	*5*	*cent*	0.017	0.052	0.047	0.066	0.17	-0.08	-0.001	85	67
IATE	*N*		0.022	0.100	-	0.125	-	-	-	91	73
IATE eff	*N*		0.022	0.086	-	0.107	-	-	-	-	-
ATE	*1*	*grf*	0.053	0.054	0.021	0.057	0.15	0.16	0.000	30	10
GATE	*5*		0.054	0.061	0.050	0.074	0.01	-0.28	-0.002	78	53
IATE	*N*		0.056	0.083	-	0.107	-	-	-	90	76
ATE	*1*	*grf*	0.021	0.025	0.023	0.031	-0.14	0.17	-0.002	83	58
GATE	*5*	*cent*	0.022	0.043	0.048	0.054	-0.06	0.03	-0.000	92	74
IATE	*N*		0.022	0.073	-	0.094	-	-	-	93	81
ATE	*1*	*dml*	0.023	0.026	0.022	0.032	0.03	0.35	0.003	90	67
GATE	*5*		0.024	0.045	0.052	0.058	0.12	0.11	0.001	92	76
IATE	*N*	*ols*	0.023	0.126	-	0.156	-	-	-	12	8
IATE	*N*	*rf*	0.016	0.248	-	0.318	-	-	-	-	-
ATE	*1*	*dml-*	0.024	0.027	0.022	0.032	0.03	0.36	0.003	88	64
GATE	*5*	*norm*	0.025	0.046	0.052	0.058	0.12	0.11	0.000	92	75
IATE	*N*	*ols*	0.024	0.126	-	0.156	-	-	-	12	8
IATE	*N*	*rf*	0.017	0.245	-	0.315	-	-	-	-	-
ATE	*1*	*ols*	0.003	0.016	0.021	0.021	0.10	0.03	-0.005	85	67
GATE	*5*		0.010	0.050	0.049	0.061	0.06	0.11	-0.015	71	50
IATE	*N*		0.002	0.122	-	0.152	-	-	-	64	50

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs.

Table C.25: Four treatments

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
ATE	*1*	*mcf*	0.086	0.099	0.087	0.122	0.205	-0.06	0.008	87	67
GATE	*5*		0.086	0.125	0.112	0.160	0.09	-0.08	0.009	85	68
IATE	*N*		0.086	0.183	-	0.228	-	-	-	83	65
IATE eff	*N*		0.084	0.167	-	0.206	-	-	-	-	-
ATE	*1*	*mcf*	-0.019	0.069	0.084	0.086	0.21	-0.04	0.004	95	80
GATE	*5*	*cent*	-0.019	0.138	0.112	0.167	0.06	0.04	-0.005	77	55
IATE	*N*		-0.019	0.222	-	0.264	-	-	-	69	46
IATE eff	*N*		-0.022	0.214	-	0.250	-	-	-	-	-
ATE	*1*	*grf*	0.027	0.052	0.060	0.066	-0.01	0.41	-0.001	92	77
GATE	*5*		0.027	0.107	0.132	0.136	0.05	0.05	0.001	94	80
IATE	*N*		-0.025	0.200	-	0.236	-	-	-	69	48
ATE	*1*	*grf*	0.004	0.049	0.062	0.062	-0.06	0.23	-0.003	95	78
GATE	*5*	*cent*	0.004	0.106	0.133	0.134	0.03	-0.03	-0.001	95	79
IATE	*N*		-0.046	0.194	-	0.232	-	-	-	70	50
ATE	*1*	*dml*	0.012	0.057	0.070	0.071	-0.12	-0.14	0.003	96	80
GATE	*5*		0.012	0.122	0.153	0.154	0.03	0.11	0.000	95	81
IATE	*N*	*ols*	0.012	0.310	-	0.390	-	-	-	43	29
IATE	*N*	*rf*	0.004	0.405	-	0.545	-	-	-	-	-
ATE	*1*	*dml-*	0.014	0.056	0.068	0.069	-0.09	-0.12	0.002	96	78
GATE	*5*	*norm*	0.014	0.119	0.149	0.150	0.04	0.03	-0.002	94	79
IATE	*N*	*ols*	0.014	0.303	-	0.381	-	-	-	41	27
IATE	*N*	*rf*	0.006	0.402	-	0.523	-	-	-	-	-
ATE	*1*	*ols*	0.019	0.054	0.064	0.067	-0.13	-0.13	-0.033	64	44
GATE	*5*		0.018	0.113	0.137	0.142	-0.01	-0.02	-0.070	65	46
IATE	*N*		0.019	0.287	-	0.361	-	-	-	62	44

Note: Table to be continued.

Table C.25 - continued: Four treatments

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
ATE	*1*	*mcf*	0.075	0.077	0.044	0.087	0.00	0.65	0.005	69	38
GATE	*5*		0.075	0.087	0.043	0.110	0.00	0.41	0.004	74	55
IATE	*N*		0.075	0.124	-	0.155	-	-	-	66	47
IATE eff	*N*		0.077	0.114	-	0.293	-	-	-	-	-
ATE	*1*	*mcf*	-0.014	0.035	0.043	0.045	-0.01	0.97	0.002	94	81
GATE	*5*	*cent*	-0.014	0.069	0.061	0.084	-0.01	0.07	-0.002	84	67
IATE	*N*		-0.015	0.129	-	0.157	-	-	-	77	54
IATE eff	*N*		-0.012	0.121	-	0.144	-	-	-	-	-
ATE	*1*	*grf*	0.035	0.038	0.030	0.045	0.03	-0.21	0.000	81	56
GATE	*5*		0.035	0.060	0.067	0.076	0.08	-0.01	-0.001	91	74
IATE	*N*		0.005	0.089	-	0.111	-	-	-	90	74
ATE	*1*	*grf*	0.005	0.024	0.029	0.030	-0.12	0.20	0.000	96	82
GATE	*5*	*cent*	0.005	0.053	0.066	0.066	-0.04	-0.13	-0.000	94	78
IATE	*N*		-0.022	0.092	-	0.113	-	-	-	88	72
ATE	*1*	*dml*	0.008	0.026	0.031	0.032	0.08	-0.32	0.004	96	83
GATE	*5*		0.008	0.056	0.070	0.072	0.12	0.10	0.004	95	82
IATE	*N*	*ols*	0.008	0.204	-	0.254	-	-	-	16	10
IATE	*N*	*rf*	0.003	0.308	-	0.419	-	-	-	-	-
ATE	*1*	*dml-*	0.009	0.026	0.031	0.032	0.08	-0.39	0.004	97	84
GATE	*5*	*norm*	0.008	0.056	0.070	0.071	0.14	0.09	0.003	95	81
IATE	*N*	*ols*	0.008	0.204	-	0.253	-	-	-	15	10
IATE	*N*	*rf*	0.004	0.305	-	0.412	-	-	-	-	-
ATE	*1*	*ols*	0.022	0.028	0.029	0.036	0.06	-0.04	-0.013	61	44
GATE	*5*		0.021	0.057	0.064	0.072	0.06	0.12	-0.030	65	46
IATE	*N*		0.021	0.202	-	0.253	-	-	-	42	28

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1'000 / 250 replications used for 2'500 / 10'000 obs. Results are shown for the comparison of treatments 1 to 0.

For the *grf*, the results differ substantially for the different treatment comparisons (which all have the same effect size), with an RMSE of the ATE/GATE/IATE in the range of 0.066-0.154 / 0.136-0.196 / 0.089-0.116 for *N*=2'500 and 0.045-0.110 / 0.076-0.126 / 0.111-0.147 for *N*=10'000.

For the centred grf, the RMSE of the ATE/GATE/IATE is in the range of 0.062-0.083 / 0.134-0.145 / 0.228-0.239 for *N*=2'500 and 0.030-0.042 / 0.066-0.073 / 0.113-0.122 for *N*=10'000.

Table C.26: Larger sample

	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 40'000					
ATE	*1*	*mcf*	0.111	0.111	0.012	0.112	0.12	-0.92	0.004	0	0
GATE	*5*		0.111	0.111	0.022	0.115	0.30	0.33	0.004	0	0
IATE	*N*		0.111	0.120	-	0.142	-	-	-	45	67
IATE eff	*N*		0.112	0.117	-	0.137	-	-	-	-	-
ATE	*1*	*mcf*	0.004	0.010	0.013	0.014	-0.01	2.40	0.002	97	87
GATE	*5*	*cent*	0.003	0.028	0.021	0.035	-0.07	1.01	0.003	84	62
IATE	*N*		0.003	0.073	-	0.090	-	-	-	83	62
IATE eff	*N*		0.001	0.067	-	0.082	-	-	-	-	-
ATE	*1*	*grf*	0.044	0.044	0.010	0.045	-0.23	0.73	0.000	3	0
GATE	*5*		0.043	0.045	0.024	0.050	-0.23	0.14	-0.001	51	26
IATE	*N*		0.042	0.066	-	0.084	-	-	-	93	80
ATE	*1*	*grf*	0.005	0.010	0.011	0.012	0.12	-0.44	-0.001	87	71
GATE	*5*	*cent*	0.006	0.020	0.023	0.025	-0.08	-0.08	-0.000	92	76
IATE	*N*		0.007	0.059	-	0.073	-	-	-	96	84
ATE	*1*	*dml*	0.008	0.011	0.009	0.012	-0.28	0.14	0.003	95	79
GATE	*5*		0.007	0.020	0.023	0.025	0.23	0.40	0.001	93	82
IATE	*N*	*ols*	0.008	0.167	-	0.201	-	-	-	2	1
IATE	*N*	*rf*	0.008	0.164	-	0.212	-	-	-	-	-
ATE	*1*	*dml-*	0.008	0.011	0.009	0.012	-0.28	0.12	0.003	95	79
GATE	*5*	*norm*	0.008	0.020	0.023	0.025	0.23	0.40	0.001	93	82
IATE	*N*	*ols*	0.008	0.167	-	0.201	-	-	-	2	1
IATE	*N*	*rf*	0.008	0.164	-	0.212	-	-	-	-	-
ATE	*1*	*ols*	0.010	0.012	0.009	0.014	-0.11	-0.37	-0.002	66	42
GATE	*5*		0.010	0.034	0.022	0.039	0.09	0.26	-0.005	49	33
IATE	*N*		0.010	0.169	-	0.203	-	-	-	24	16

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 62 replications.

Table C.27: Only 25% treated

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
ATE	*1*	*mcf*	0.247	0.247	0.073	0.257	-0.04	-0.12	0.001	8	2
GATE	*5*		0.246	0.247	0.094	0.300	-0.05	-0.09	0.001	45	26
IATE	*N*		0.247	0.269	-	0.369	-	-	-	53	42
IATE eff	*N*		0.247	0.287	-	0.360	-	-	-	-	-
ATE	*1*	*mcf*	0.087	0.094	0.071	0.112	-0.04	0.07	0.000	76	50
GATE	*5*	*cent*	0.086	0.133	0.099	0.178	-0.02	-0.02	-0.007	75	57
IATE	*N*		0.086	0.234	-	0.280	-	-	-	61	40
IATE eff	*N*		0.089	0.227	-	0.269	-	-	-	-	-
ATE	*1*	*grf*	0.144	0.144	0.051	0.152	-0.02	0.11	-0.002	17	6
GATE	*5*		0.143	0.156	0.111	0.187	-0.00	0.00	-0.003	71	49
IATE	*N*		0.172	0.293	-	0.351	-	-	-	46	28
ATE	*1*	*grf*	0.075	0.078	0.050	0.090	0.05	-0.13	-0.001	65	39
GATE	*5*	*cent*	0.075	0.110	0.108	0.138	-0.09	0.07	-0.001	88	67
IATE	*N*		0.108	0.274	-	0.317	-	-	-	47	27
ATE	*1*	*dml*	0.025	0.051	0.059	0.064	-0.15	0.10	0.005	94	79
GATE	*5*		0.025	0.108	0.134	0.138	0.01	0.28	-0.003	93	78
IATE	*N*	*ols*	0.025	0.279	-	0.351	-	-	-	34	23
IATE	*N*	*rf*	0.010	0.359	-	0.525	-	-	-	-	-
ATE	*1*	*dml-*	0.029	0.051	0.056	0.063	-0.09	0.07	0.004	93	78
GATE	*5*	*norm*	0.029	0.105	0.129	0.133	0.04	0.04	-0.003	93	77
IATE	*N*	*ols*	0.030	0.272	-	0.341	-	-	-	32	22
IATE	*N*	*rf*	0.016	0.351	-	0.472	-	-	-	-	-
ATE	*1*	*ols*	0.022	0.046	0.053	0.058	-0.06	0.07	-0.021	71	54
GATE	*5*		0.022	0.097	0.114	0.122	0.03	-0.04	-0.046	72	52
IATE	*N*		0.022	0.257	-	0.321	-	-	-	62	43

Note: Table to be continued.

Table C.27 - continued: Only 25% treated

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10’000					
ATE	*1*	*mcf*	0.189	0.189	0.035	0.192	0.06	-0.10	0.002	0	0
GATE	*5*		0.188	0.188	0.053	0.214	0.03	-0.03	0.001	20	7
IATE	*N*		0.188	0.208	-	0.257	-	-	-	54	42
IATE eff	*N*		0.186	0.201	-	0.251	-	-	-	-	-
ATE	*1*	*mcf*	0.043	0.046	0.034	0.055	0.16	0.02	0.002	79	54
GATE	*5*	*cent*	0.042	0.066	0.052	0.091	0.02	0.01	0.001	80	63
IATE	*N*		0.042	0.134	-	0.166	-	-	-	75	54
IATE eff	*N*		0.042	0.127	-	0.157	-	-	-	-	-
ATE	*1*	*grf*	0.083	0.083	0.026	0.087	-0.14	0.45	-0.001	7	4
GATE	*5*		0.082	0.087	0.056	0.103	0.05	-0.19	-0.003	66	42
IATE	*N*		0.098	0.130	-	0.177	-	-	-	76	61
ATE	*1*	*grf*	0.030	0.033	0.026	0.039	0.05	-0.33	-0.002	72	51
GATE	*5*	*cent*	0.030	0.052	0.054	0.066	-0.03	0.07	-0.001	88	71
IATE	*N*		0.049	0.116	-	0.152	-	-	-	80	64
ATE	*1*	*dml*	0.012	0.024	0.028	0.030	-0.05	-0.23	0.003	94	81
GATE	*5*		0.011	0.053	0.065	0.067	0.03	-0.09	0.030	95	78
IATE	*N*	*ols*	0.011	0.193	-	0.238	-	-	-	12	7
IATE	*N*	*rf*	0.000	0.266	-	0.371	-	-	-	-	-
ATE	*1*	*dml-*	0.012	0.024	0.028	0.030	-0.03	-0.21	0.003	95	80
GATE	*5*	*norm*	0.012	0.053	0.065	0.067	0.04	-0.13	-0.001	95	78
IATE	*N*	*ols*	0.012	0.193	-	0.238	-	-	-	11	8
IATE	*N*	*rf*	0.001	0.264	-	0.367	-	-	-	-	-
ATE	*1*	*ols*	0.020	0.023	0.026	0.033	-0.13	0.15	-0.010	62	45
GATE	*5*		0.020	0.049	0.055	0.068	-0.06	0.04	0.055	66	46
IATE	*N*		0.020	0.193	-	0.236	-	-	0.236	43	29

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 1’000 / 250 replications used for 2’500 / 10’000 obs. Mcf results are based on version 0.4.1.

Table C.28: More GATE groups, no selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
GATE	*5*	*mcf*	-0.004	0.092	0.087	0.115	0.00	0.01	0.002	86	67
GATE	*10*		-0.004	0.095	0.090	0.118	-0.01	-0.04	-0.001	86	86
GATE	*20*		-0.004	0.097	0.060	0.120	-0.01	-0.02	-0.001	85	65
GATE	*40*		-0.004	0.099	0.093	0.123	-0.01	-0.01	-0.003	84	64
GATE	*5*	*mcf*	0.002	0.090	0.091	0.114	0.02	0.09	-0.010	83	64
GATE	*10*	*cent*	0.002	0.094	0.094	0.117	-0.02	0.02	-0.012	82	63
GATE	*20*		0.002	0.094	0.094	0.118	0.00	0.04	-0.011	83	63
GATE	*40*		0.002	0.096	0.096	0.120	0.00	0.03	-0.013	82	62
GATE	*5*	*grf*	-0.001	0.076	0.096	0.096	0.06	0.09	-0.001	95	80
GATE	*10*		-0.001	0.108	0.135	0.135	0.03	0.00	-0.001	95	79
GATE	*20*		-0.001	0.152	0.190	0.191	-0.01	0.02	-0.001	95	80
GATE	*40*		-0.001	0.215	0.270	0.270	0.01	0.05	-0.001	95	80
GATE	*5*	*grf*	0.004	0.075	0.094	0.094	0.01	0.01	0.000	95	80
GATE	*10*	*cent*	0.004	0.106	0.133	0.133	-0.00	0.02	0.000	95	80
GATE	*20*		0.004	0.150	0.189	0.189	0.02	0.07	-0.000	95	80
GATE	*40*		0.004	0.213	0.269	0.269	0.02	0.11	-0.001	95	80
GATE	*5*	*dml*	-0.002	0.075	0.094	0.095	-0.02	0.04	0.002	95	81
GATE	*10*		-0.002	0.108	0.134	0.134	-0.02	-0.06	0.002	95	81
GATE	*20*		-0.002	0.152	0.191	0.191	-0.01	0.07	0.001	95	80
GATE	*40*		-0.002	0.217	0.272	0.273	0.01	0.06	-0.003	95	79
GATE	*5*	*dml-*	-0.002	0.075	0.094	0.094	-0.02	0.05	0.002	95	80
GATE	*10*	*norm*	-0.002	0.107	0.133	0.133	-0.02	-0.05	0.000	95	80
GATE	*20*		-0.002	0.150	0.189	0.189	-0.01	0.07	-0.001	95	80
GATE	*40*		-0.002	0.214	0.268	0.269	0.00	0.06	-0.003	94	79
GATE	*5*	*ols*	-0.002	0.073	0.092	0.093	-0.04	0.09	-0.027	84	64
GATE	*10*		-0.002	0.105	0.131	0.131	-0.01	-0.03	-0.039	83	63
GATE	*20*		-0.002	0.149	0.187	0.187	-0.01	0.07	-0.056	83	63
GATE	*40*		-0.002	0.213	0.268	0.268	0.01	0.07	-0.082	83	63

Note: Table to be continued.

Table C.28 - continued: More GATE groups, no selectivity

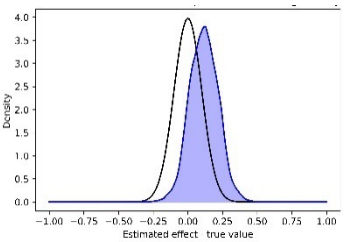
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
GATE	*5*	*mcf*	0.001	0.049	0.045	0.063	0.07	-0.21	0.002	87	66
GATE	*10*		0.001	0.053	0.048	0.066	0.01	-0.23	0.000	84	64
GATE	*20*		0.001	0.053	0.049	0.067	0.04	-0.25	0.000	85	64
GATE	*40*		0.001	0.054	0.050	0.068	0.04	-0.18	-0.001	84	64
GATE	*5*	*mcf*	0.004	0.045	0.045	0.058	0.05	0.17	-0.002	87	68
GATE	*10*	*cent*	0.004	0.050	0.049	0.063	0.07	-0.22	-0.005	85	64
GATE	*20*		0.004	0.049	0.049	0.062	0.09	-0.16	-0.004	85	65
GATE	*40*		0.004	0.050	0.050	0.063	0.09	-0.11	-0.005	85	64
GATE	*5*	*grf*	-0.001	0.037	0.047	0.047	-0.09	0.13	-0.001	94	80
GATE	*10*		-0.001	0.053	0.067	0.067	-0.03	0.03	-0.001	94	80
GATE	*20*		-0.001	0.076	0.094	0.095	-0.04	-0.05	-0.001	95	79
GATE	*40*		-0.001	0.107	0.134	0.134	-0.01	-0.00	-0.002	94	79
GATE	*5*	*grf*	0.002	0.036	0.046	0.046	-0.07	-0.05	0.001	95	80
GATE	*10*	*cent*	0.002	0.052	0.065	0.065	-0.04	-0.13	0.001	95	80
GATE	*20*		0.002	0.072	0.091	0.091	0.02	-0.02	0.002	96	81
GATE	*40*		0.002	0.104	0.130	0.130	0.03	-0.05	0.002	95	80
GATE	*5*	*dml*	-0.001	0.037	0.046	0.046	0.06	-0.02	0.001	95	82
GATE	*10*		-0.001	0.052	0.064	0.064	-0.04	-0.08	0.002	96	81
GATE	*20*		-0.001	0.075	0.094	0.094	-0.03	0.00	0.000	95	80
GATE	*40*		-0.001	0.106	0.131	0.132	-0.01	-0.02	0.000	95	80
GATE	*5*	*dml-*	-0.001	0.037	0.046	0.046	0.05	-0.03	0.001	95	82
GATE	*10*	*norm*	-0.001	0.052	0.064	0.064	-0.04	-0.07	0.002	96	81
GATE	*20*		-0.001	0.075	0.093	0.093	-0.03	0.01	0.000	95	80
GATE	*40*		-0.001	0.106	0.132	0.132	-0.01	-0.02	0.000	95	80
GATE	*5*	*ols*	0.000	0.037	0.046	0.046	0.01	-0.03	-0.013	85	62
GATE	*10*		-0.001	0.052	0.064	0.064	-0.04	-0.13	-0.018	84	63
GATE	*20*		-0.001	0.074	0.093	0.093	-0.03	-0.02	-0.028	83	63
GATE	*40*		-0.001	0.105	0.131	0.132	0.01	0.00	-0.039	83	63

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 250 replications used for 10'000 obs.

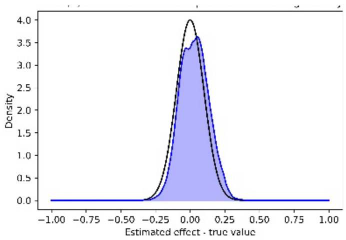
Table C.28 – continued: More GATE groups, no selectivity - Distribution of GATE minus true value of the centred mcf

N = 2’500

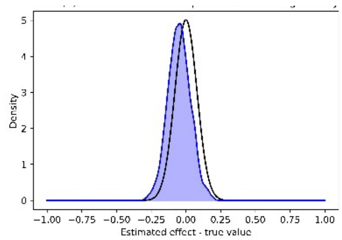
True GATE: 0.50



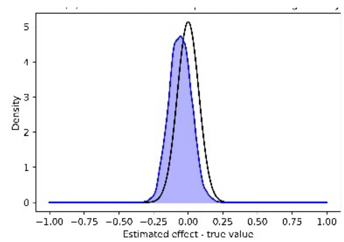
True GATE: 0.77



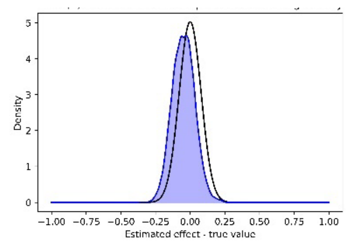
True GATE: 1.20



True GATE: 1.24

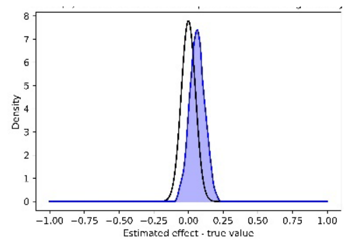


True GATE: 1.24

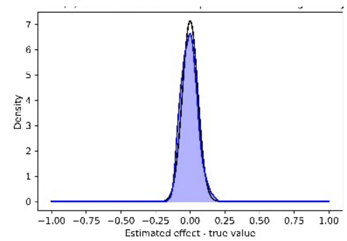


N = 10’000

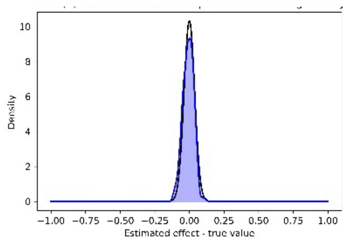
True GATE: 0.50



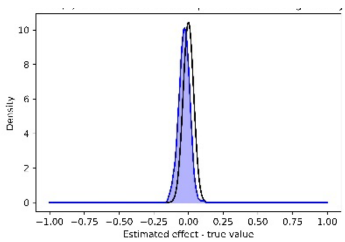
True GATE: 0.77



True GATE: 1.20



True GATE: 1.24



True GATE: 1.24

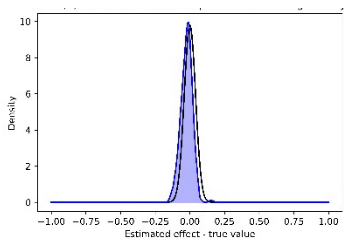


Table C.29: More GATE groups, medium selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2’500					
GATE	*5*	*mcf*	0.175	0.178	0.090	0.211	0.06	-0.04	0.008	53	32
GATE	*10*		0.174	0.178	0.093	0.212	0.02	-0.06	-0.003	55	33
GATE	*20*		0.174	0.178	0.094	0.213	0.03	-0.06	-0.003	56	34
GATE	*40*		0.174	0.179	0.096	0.215	0.03	-0.06	-0.005	56	35
GATE	*5*	*mcf*	0.037	0.094	0.090	0.121	0.05	0.07	-0.006	84	65
GATE	*10*	*cent*	0.036	0.097	0.094	0.126	0.02	0.02	-0.008	83	64
GATE	*20*		0.036	0.098	0.094	0.126	0.03	0.02	-0.007	83	64
GATE	*40*		0.036	0.099	0.095	0.128	0.03	0.05	-0.008	83	64
GATE	*5*	*grf*	0.097	0.113	0.096	0.137	0.03	-0.00	-0.003	81	57
GATE	*10*		0.097	0.134	0.134	0.166	0.02	0.04	-0.002	88	68
GATE	*20*		0.097	0.171	0.189	0.213	-0.01	0.05	-0.002	92	74
GATE	*40*		0.097	0.228	0.268	0.285	-0.01	0.09	-0.002	93	77
GATE	*5*	*grf*	0.037	0.081	0.093	0.102	-0.01	0.07	-0.000	93	76
GATE	*10*	*cent*	0.037	0.110	0.132	0.138	-0.02	0.07	-0.001	94	78
GATE	*20*		0.037	0.151	0.186	0.190	0.02	0.10	0.000	95	79
GATE	*40*		0.037	0.213	0.265	0.268	-0.01	0.08	-0.001	95	79
GATE	*5*	*dml*	0.019	0.081	0.099	0.101	-0.04	-0.05	0.001	95	80
GATE	*10*		0.019	0.113	0.140	0.142	-0.04	-0.06	0.001	95	79
GATE	*20*		0.019	0.160	0.200	0.201	-0.02	0.05	-0.001	95	79
GATE	*40*		0.019	0.227	0.284	0.285	-0.02	0.07	-0.005	94	79
GATE	*5*	*dml-*	0.020	0.081	0.081	0.101	-0.04	-0.05	0.001	95	79
GATE	*10*	*norm*	0.020	0.113	0.140	0.141	-0.04	-0.07	0.000	95	79
GATE	*20*		0.020	0.160	0.199	0.200	-0.02	0.06	-0.002	94	79
GATE	*40*		0.020	0.226	0.283	0.284	-0.02	0.07	-0.005	94	79
GATE	*5*	*ols*	0.007	0.080	0.094	0.100	-0.06	0.01	-0.027	81	61
GATE	*10*		0.006	0.110	0.133	0.138	-0.01	0.04	-0.039	82	62
GATE	*20*		0.006	0.154	0.191	0.194	-0.01	0.07	-0.057	82	62
GATE	*40*		0.006	0.219	0.273	0.275	0.02	0.06	-0.083	82	62

Note: Table to be continued.

Table C.29 - continued: More GATE groups, medium selectivity

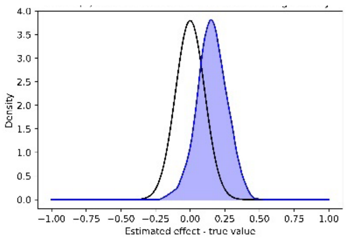
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
			Estimation of effects						**Estimation of std. errors**		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
GATE	*5*	*mcf*	0.144	0.144	0.048	0.158	0.16	-0.03	-0.001	21	7
GATE	*10*		0.143	0.143	0.050	0.159	0.05	-0.06	-0.002	25	8
GATE	*20*		0.142	0.143	0.051	0.158	0.07	-0.07	-0.001	26	10
GATE	*40*		0.142	0.143	0.52	0.159	0.09	-0.08	-0.002	26	11
GATE	*5*	*mcf*	0.013	0.049	0.045	0.063	0.22	0.17	0.004	86	69
GATE	*10*	*cent*	0.013	0.054	0.049	0.069	0.09	0.10	-0.002	84	63
GATE	*20*		0.012	0.054	0.050	0.068	0.10	0.06	-0.001	85	64
GATE	*40*		0.012	0.055	0.051	0.069	0.12	0.03	-0.002	84	64
GATE	*5*	*grf*	0.064	0.068	0.047	0.080	-0.03	-0.15	-0.001	71	46
GATE	*10*		0.064	0.076	0.067	0.093	0.00	-0.07	-0.001	84	60
GATE	*20*		0.064	0.092	0.094	0.114	-0.04	-0.04	-0.001	89	69
GATE	*40*		0.064	0.119	0.133	0.148	-0.03	-0.00	-0.001	92	75
GATE	*5*	*grf*	0.016	0.041	0.048	0.051	-0.04	-0.03	-0.001	92	75
GATE	*10*	*cent*	0.016	0.055	0.065	0.068	-0.05	-0.09	-0.000	94	78
GATE	*20*		0.016	0.075	0.092	0.094	-0.01	0.02	0.000	95	79
GATE	*40*		0.016	0.106	0.131	0.132	-0.01	-0.04	0.000	95	79
GATE	*5*	*dml*	0.010	0.041	0.050	0.051	0.14	0.04	0.000	94	80
GATE	*10*		0.010	0.056	0.069	0.070	0.05	0.05	0.001	94	80
GATE	*20*		0.010	0.081	0.100	0.101	-0.04	0.01	-0.001	94	79
GATE	*40*		0.010	0.113	0.140	0.141	-0.01	-0.08	-0.001	95	79
GATE	*5*	*dml-*	0.010	0.041	0.050	0.051	0.13	0.04	0.000	94	80
GATE	*10*	*norm*	0.010	0.056	0.069	0.070	0.05	0.05	0.001	94	80
GATE	*20*		0.010	0.081	0.100	0.101	-0.04	0.01	-0.001	94	79
GATE	*40*		0.010	0.113	0.140	0.141	-0.01	-0.08	-0.001	95	79
GATE	*5*	*ols*	0.007	0.046	0.047	0.058	0.09	0.22	-0.014	74	53
GATE	*10*		0.007	0.059	0.066	0.075	0.05	0.02	-0.019	75	58
GATE	*20*		0.007	0.081	0.095	0.101	-0.03	-0.03	-0.029	80	60
GATE	*40*		0.007	0.111	0.134	0.139	0.00	-0.05	-0.040	81	61

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 250 replications used for 10'000 obs.

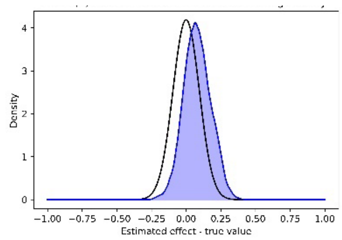
Table C.29 – continued: More GATE groups, medium selectivity - Distribution of GATE minus true value of the centred mcf

N = 2'500

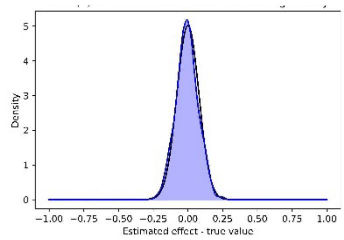
True GATE: 0.50



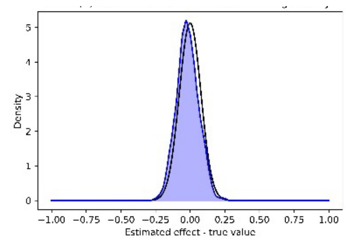
True GATE: 0.77



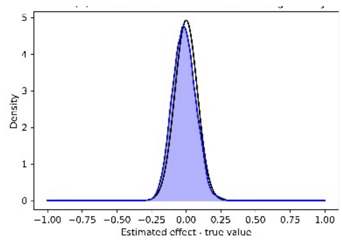
True GATE: 1.20



True GATE: 1.24

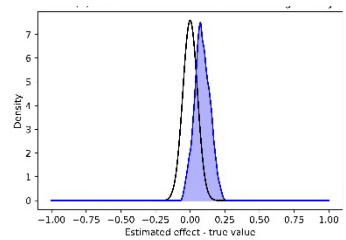


True GATE: 1.24

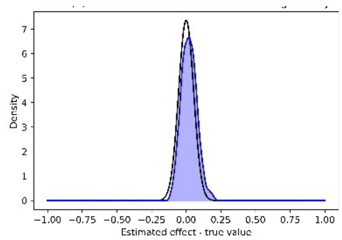


N = 10'000

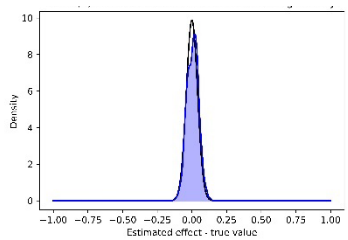
True GATE: 0.50



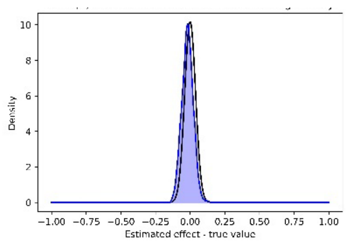
True GATE: 0.77



True GATE: 1.20



True GATE: 1.24



True GATE: 1.24

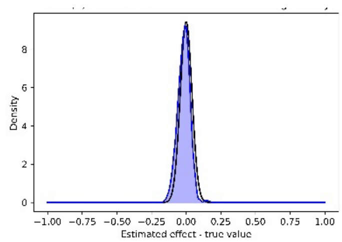


Table C.30: More GATE groups, strong selectivity

			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 2'500					
GATE	*5*	*mcf*	0.436	0.436	0.010	0.455	0.00	0.06	0.001	1	0
GATE	*10*		0.435	0.435	0.097	0.455	0.00	0.03	0.000	2	0
GATE	*20*		0.434	0.434	0.097	0.439	0.02	-0.03	-0.001	0	0
GATE	*40*		0.435	0.435	0.064	0.439	0.02	-0.04	-0.001	0	0
GATE	*5*	*mcf*	0.046	0.112	0.085	0.146	-0.03	0.11	0.009	81	62
GATE	*10*	*cent*	0.044	0.117	0.089	0.151	-0.03	0.07	0.007	80	61
GATE	*20*		0.044	0.117	0.089	0.152	-0.01	0.05	0.008	80	62
GATE	*40*		0.043	0.119	0.090	0.153	0.00	0.06	0.007	80	61
GATE	*5*	*grf*	0.284	0.284	0.101	0.309	-0.06	0.06	-0.011	19	8
GATE	*10*		0.283	0.287	0.136	0.322	-0.03	0.02	-0.009	42	21
GATE	*20*		0.284	0.297	0.187	0.346	-0.04	0.03	-0.006	63	39
GATE	*40*		0.283	0.324	0.260	0.391	-0.06	0.09	-0.004	79	55
GATE	*5*	*grf*	0.090	0.119	0.096	0.148	0.00	0.09	-0.007	75	56
GATE	*10*	*cent*	0.090	0.138	0.130	0.173	-0.02	0.06	-0.005	84	64
GATE	*20*		0.090	0.171	0.181	0.214	0.00	0.09	-0.003	89	71
GATE	*40*		0.090	0.223	0.255	0.280	-0.03	0.10	-0.003	92	75
GATE	*5*	*dml*	0.147	0.165	0.127	0.195	-0.61	4.66	-0.014	66	46
GATE	*10*		0.147	0.187	0.173	0.228	-0.78	8.17	-0.017	78	57
GATE	*20*		0.147	0.226	0.241	0.284	-0.99	12.63	-0.025	84	64
GATE	*40*		0.148	0.286	0.338	0.372	-1.07	14.96	-0.040	87	69
GATE	*5*	*dml-*	0.140	0.157	0.123	0.187	-0.13	0.01	-0.005	74	52
GATE	*10*	*norm*	0.140	0.182	0.170	0.221	-0.15	0.08	-0.005	82	62
GATE	*20*		0.140	0.225	0.239	0.278	-0.21	0.34	-0.010	87	68
GATE	*40*		0.140	0.293	0.338	0.367	-0.29	0.76	-0.020	89	71
GATE	*5*	*ols*	0.111	0.112	0.054	0.123	-0.02	0.01	-0.015	25	13
GATE	*10*		0.109	0.153	0.143	0.188	-0.04	-0.04	-0.042	69	49
GATE	*20*		0.109	0.190	0.202	0.236	0.00	0.00	-0.060	75	55
GATE	*40*		0.109	0.250	0.288	0.313	0.01	0.08	-0.087	79	58

Note: Table to be continued.

Table C.30 - continued: More GATE groups, strong selectivity

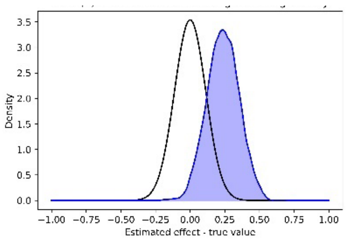
			Estimation of effects						Estimation of std. errors		
	# of groups	Estim-ator	Bias	Mean abs. error	Std. dev.	RMSE	Skew-ness	Ex. Kurt-osis	Bias (SE)	CovP (95) in %	CovP (80) in %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(12)
						N = 10'000					
GATE	*5*	*mcf*	0.375	0.375	0.053	0.384	-0.10	-0.10	-0.002	0	0
GATE	*10*		0.374	0.374	0.056	0.383	-0.09	-0.09	-0.002	0	0
GATE	*20*		0.373	0.373	0.056	0.383	-0.09	-0.08	-0.002	0	0
GATE	*40*		0.373	0.373	0.057	0.383	-0.08	-0.06	-0.002	0	0
GATE	*5*	*mcf*	-0.014	0.085	0.045	0.100	0.01	0.03	0.007	65	41
GATE	*10*	*cent*	-0.016	0.091	0.049	0.104	-0.05	-0.04	0.005	62	35
GATE	*20*		-0.016	0.090	0.049	0.104	-0.03	0.02	0.006	63	37
GATE	*40*		-0.016	0.091	0.050	0.104	-0.02	-0.02	0.005	63	37
GATE	*5*	*grf*	0.211	0.211	0.052	0.220	0.00	-0.04	-0.005	3	0
GATE	*10*		0.211	0.211	0.071	0.225	0.01	0.02	-0.006	15	5
GATE	*20*		0.211	0.212	0.096	0.234	-0.01	0.01	-0.004	39	19
GATE	*40*		0.211	0.219	0.134	0.252	-0.06	0.04	-0.004	63	37
GATE	*5*	*grf*	0.043	0.063	0.051	0.080	-0.09	0.03	-0.006	75	56
GATE	*10*	*cent*	0.043	0.073	0.068	0.092	-0.06	0.17	-0.004	83	63
GATE	*20*		0.043	0.089	0.093	0.112	-0.08	0.08	-0.002	89	71
GATE	*40*		0.043	0.115	0.130	0.145	-0.02	-0.03	-0.002	92	74
GATE	*5*	*dml*	0.088	0.099	0.091	0.119	0.10	6.78	-0.015	62	42
GATE	*10*		0.088	0.110	0.104	0.138	0.13	8.59	-0.017	74	53
GATE	*20*		0.088	0.132	0.144	0.173	-0.21	10.79	-0.024	83	61
GATE	*40*		0.088	0.164	0.198	0.224	-0.62	11.75	-0.032	87	68
GATE	*5*	*dml-*	0.078	0.090	0.073	0.108	-0.08	0.04	-0.007	71	49
GATE	*10*	*norm*	0.079	0.103	0.098	0.127	-0.09	0.30	-0.006	81	60
GATE	*20*		0.079	0.129	0.138	0.160	-0.22	1.02	-0.010	87	67
GATE	*40*		0.079	0.164	0.192	0.209	-0.35	1.75	-0.015	89	71
GATE	*5*	*ols*	0.113	0.118	0.052	0.134	0.05	-0.07	-0.015	27	17
GATE	*10*		0.110	0.121	0.071	0.142	-0.02	-0.13	-0.021	42	27
GATE	*20*		0.110	0.132	0.100	0.158	0.02	-0.07	-0.029	56	38
GATE	*40*		0.110	0.153	0.141	0.187	-0.01	-0.06	-0.042	67	47

Note: For GATE and IATE, results are averaged over all effects. *CovP (95, 80)* denotes the (average) probability that the true value is part of the 95% / 80% confidence interval. 250 replications used for 10'000 obs.

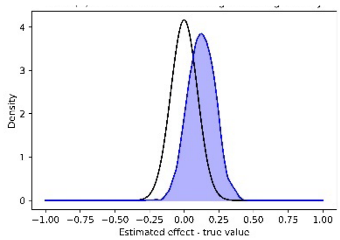
Table C.30 – continued: More GATE groups, strong selectivity - Distribution of GATE minus true value of the centred mcf

N = 2'500

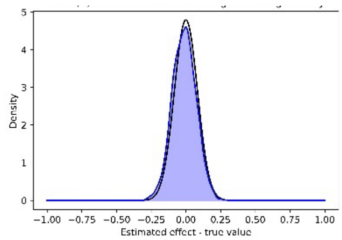
True GATE: 0.50



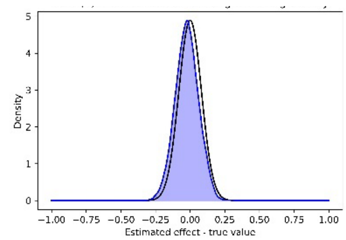
True GATE: 0.77



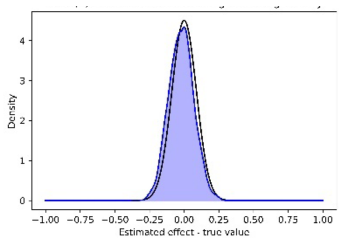
True GATE: 1.20



True GATE: 1.24

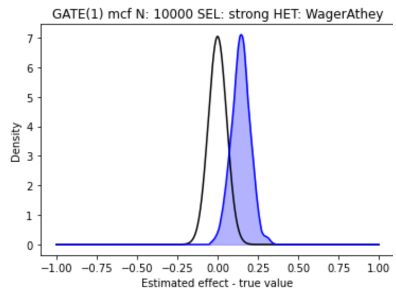


True GATE: 1.24

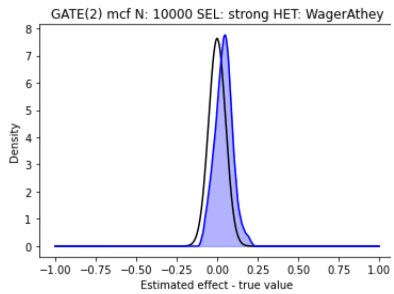


N = 10'000

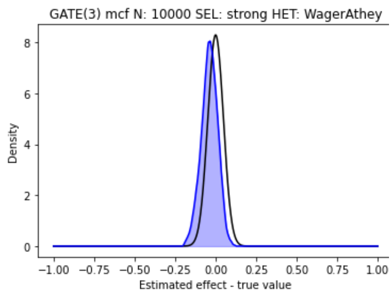
True GATE: 0.50



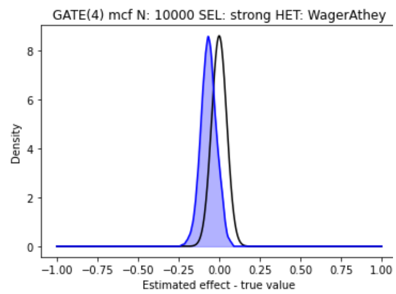
True GATE: 0.77



True GATE: 1.20



True GATE: 1.24



True GATE: 1.24