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

**Working Paper**

# Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed

IZA Discussion Papers, No. 12526

**Provided in Cooperation with:**
IZA – Institute of Labor Economics

*Suggested Citation:* Goller, Daniel; Lechner, Michael; Moczall, Andreas; Wolff, Joachim (2019) : Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed, IZA Discussion Papers, No. 12526, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/207352

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12526

# Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed

Daniel Goller
Michael Lechner
Andreas Moczall
Joachim Wolff

AUGUST 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12526

# Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed

**Daniel Goller**
*University of St.Gallen*

**Andreas Moczall**
*IAB*

**Michael Lechner**
*University of St.Gallen, CEPR, CESIfo, IAB, IZA and RWI*

**Joachim Wolff**
*IAB*

AUGUST 2019



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed*

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

**JEL Classification:** J68, C21

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

**Corresponding author:**
Michael Lechner
Professor of Econometrics
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen
Varnbüelstrasse 14
CH-9000 St. Gallen
Switzerland

E-mail: Michael.Lechner@unisg.ch

* Support of the IAB under grant for the project "Estimating heterogeneous effects of the Schemes for Activation and Integration on welfare recipients' outcomes: Enhanced analyses by the application of machine learning algorithms" is gratefully acknowledged. A previous version of the paper was presented at the University of St. Gallen. We thank participants, in particular Michael Zimmert, as well as Michael Knaus and Gabriel Okasa for helpful comments and suggestions. The usual disclaimer applies.

# 1 Introduction

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

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

Although off-the-shelf machine learning methods have many well-documented advantages in prediction and classification, it is not obvious that using them for propensity score

1 For an overview of statistical learning methods, see e.g. Hastie, Tibshirani, and Friedman (2009).

estimation in a matching framework will improve the estimation of causal effects. One potential reason is that they aim at a different target (compare Athey and Imbens (2019)). The goal of using a PS in matching estimation is to balance the covariate distribution of treated and non-treated units to obtain a quasi-experimental situation. Machine learning algorithms, if used for PS estimation, follow the goal to predict treatment participation given the covariates, as good as possible, by trading off bias and variance in out-of-sample comparisons. One example for why this may be a bad idea are covariates that are very good predictors of outcome but only weakly correlated with treatment assignment (compare e.g. Belloni, Chernozhukov, and Hansen, 2014). Since machine learning algorithms try to maximize predictive power (in a mean square error sense), they may omit such variables as they do not help much to predict the treatment, accepting a somewhat larger bias in the propensity score that is dominated by the resulting variance reduction. However, since these now-omitted variables are important predictors of the outcome, the small bias in the propensity may translate to a large one in the ATET estimation.

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

in PS matching. As these four studies used the same data generating process based on only ten covariates and a treatment share of 50 percent, they might be less informative for microeconometric applications in which the dimension of confounders is usually much higher and the treatment shares very likely to deviate from 50 percent.

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

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

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

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

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

We evaluate the effect of a training programme and simulate the performance of different PS estimators, using the radius matching on the propensity score with bias adjustment (RMBA) algorithm developed in Lechner, Miquel, and Wunsch (2011), which performed best in the simulation of Huber, Lechner, and Wunsch (2013). To be more precise, we use two different machine learning techniques, namely Random Forest and LASSO to estimate the PS. We choose these two as they use very different approaches in a non-parametric sense: Random Forest approximate the PS locally, similar to non-parametric regression, while LASSO with many polynomials and interaction term will approximate the PS with a flexible global function, e.g. similar to series estimation. In that sense, they represent two very different types of approaches. A large literature discusses both methods, establishing theoretical properties (e.g. Hastie, Tibshirani, and Friedman, 2009), as well as modifying them for usage in other types of causal inference problems (e.g. Belloni, Chernozhukov, and Hansen (2014), Wager and Athey (2018), Lechner (2018), Athey, Tibshirani, and Wager (2019)). We compare these two estimators to the true, a random, and a PS based on a parametric ad-hoc (Probit) model, which we then use for estimating the ATET in the RMBA estimator.

Our findings are mixed. LASSO performs well as PS estimator for the usage in radius matching especially in situations in which using Probit and Random Forest do not deliver credible estimates. When there are many covariates compared to observations, Probit does not work well; once the number of observations increases sufficiently, Probit and LASSO perform equally well. Random Forest tends to predict the treatment in sample well, but does not work

properly as balancing score estimator. If the share of treated units is low, the Random Forest cannot manage to split deep enough to estimate a PS flexible enough to remove the selection bias. In fact, we find that PS estimated with Random Forest may lead to comparing control units and treated units, which are not sufficiently similar. Thus, whether using specific off-the-shelf machine learning algorithms does help depends on the context of the application. Since knowing which of the methods works a priori appears to be difficult, a plausible alternative is to use Causal Machine Learning methods instead, e.g. 'double machine learning' suggested by Chernozhukov et al. (2018) or the Modified Causal Forest suggested by Lechner (2018), that are optimized specifically for treatment effect estimation (for an overview see e.g. Knaus, Lechner, Strittmatter, 2018).

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

The structure of the rest of the paper is as follows: In Sections 2 and 3, we describe the institutional background and the database used for the simulation and application in detail. Section 4 introduces the EMCS, as well as the estimators used. Sections 5 and 6 present the results of the simulations and the empirical application. Section 7 concludes. Additional results can be found in the Appendices.

# 2 Institutional background

We analyse these methodological questions with regard to the effects of a German short-term training programme named Determining, Reducing and Removing Employment Impediments (DRR). It is a sub-programme of the Schemes for Activation and Integration (SAI)

that consist of different training programmes as well as placement services by private providers.[2]

The SAI programmes, introduced in 2009, replaced a number of earlier programmes with similar basic objectives. They differed from its predecessors in providing greater flexibility to local service providers to better suit their services to the particular needs of different unemployed persons. While there are many sub-programmes within SAI differing in their target groups and detailed goals, we focus only on the "Determining, Reducing and Removing Employment Impediments" sub-programme in order to analyse a rather homogeneous treatment type. The DRR sub-programme focuses on finding out which particular attributes define the individual's disadvantage, improving participants' skills, and providing them with knowledge about suitable occupational fields and individual opportunities on the labour market. The target group is both unemployment insurance and unemployed welfare recipients. The latter are usually long-term unemployed with some prospects of labour market integration. Among the various types of Schemes for Activation and Integration, the relative importance of the "Determining, Reducing and Removing Employment Impediments" sub-programme is considerable. It represents 15 percent of the 428'000 persons entering any type of Schemes for Activation and Integration (SAI) programme in our observation period January to March 2010.[3][4] Due to the flexible programme design, there is no programme duration defined à priori; the average duration is slightly less than two months.

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

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

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

# 3 Data

## 3.1 Dataset

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

The data on individuals are based on employer reports to the German social security administration as well as internal records of job centres and labour agencies. They contain socio-demographic characteristics, information on the last job, and almost complete employment and unemployment histories.[5] Moreover, these data include welfare benefit receipt, welfare benefit sanctions, ALMP participation, household composition and income information. The variables are available for the unemployed welfare recipients themselves as well as for their partners.

We augment this dataset with characteristics of the local labour market. They include the unemployment rate, the long-term-unemployment rate, the vacancy-unemployment ratio, the number of registered unemployed people and of unemployment benefit II recipients, and the inflow into various active labour market programmes. Finally, we add information on the staff structure of the job centres. Job centre employee data is available as full-time equivalents. The most important piece of information in this context is the average number of welfare recipients for which a job centre employee is responsible. It provides a measure of the intensity of activation. Other available measures in this context are, e.g. the gender distribution of job centre employees, the distribution of contract types, e.g. fixed-term versus open ended or employee versus civil servant, the presence of equal opportunity officers, and the wage distribution among the job centre employees.

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

## 3.2 Treatment and sample selection

Our sample design is similar to the one used by Harrer, Moczall, and Wolff (2019), which analysed the effectiveness of the entire SAI. Our treatment group consists of the total inflow from January to March 2010 into the "Determining, Reducing and Removing Employment Impediments" sub-type of the SAI who were unemployed and receiving means-tested benefits on December 31st, 2009. The control group represents a 20 percent random sample of persons likewise unemployed and receiving means-tested welfare benefits on December 31st, 2009, who did not enter any SAI programme from January to March 2010 but may have entered other programme types.

For data quality reasons, we restrict the sample to individuals administered jointly by the Federal Employment Agency and municipalities.[6] Moreover, we only include individuals aged 25 to 55 who are not disabled. For younger welfare recipients, various special rules and group-specific programmes exist so that they are subject to more intense activation than older welfare recipients are. Finally, we dismissed observations from our sample due to missing or obviously wrong values in some of the variables. The remaining final sample of 276'637 observations is analysed in the application in Section 6 and our EMCS described in Section 4.

## 3.3 Descriptive statistics

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

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

*Table 1: Descriptive statistics of selected covariates*

| Variable | Treated | | Controls | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Cumulated duration in regular employment 3-36 months after treatment *(Outcome)* | 218 | (314) | 162 | (282) |
| Female | 0.44 | | 0.46 | |
| Age at sampling date in years | 38 | (8) | 40 | (9) |
| Receives some income from employment (<15h/week) | 0.17 | | 0.23 | |
| Cumulated number of days in welfare receipt in year before sampling date | 312 | (100) | 325 | (86) |
| Participated in Schemes by Providers[1)] | 0.14 | | 0.06 | |
| Participated in classroom training | 0.61 | | 0.45 | |
| Job centre district: Inflow into Schemes by Providers[1)] relative to jobseeker stock in 2009q4 | 0.02 | (0.01) | 0.01 | (0.01) |
| Job centre district: Inflow into In-Firm Training[2)] relative to jobseeker stock in 2009q4 | 0.004 | (0.002) | 0.004 | (0.002) |
| Days since last employment[3)] | 1904 | (1877) | 2262 | (2045) |
| Cumulated days in regular employment in five years before sampling date | 230 | (372) | 183 | (334) |
| Secondary schooling degree: None | 0.12 | | 0.15 | |
| Secondary schooling degree: Lower | 0.47 | | 0.44 | |
| Secondary schooling degree: Intermediate | 0.29 | | 0.28 | |
| Secondary schooling degree: University of applied science qualification | 0.04 | | 0.04 | |
| Secondary schooling degree: University qualification (A-level equivalent) | 0.07 | | 0.08 | |
| Vocational degree: None | 0.50 | | 0.51 | |
| Vocational degree: Non-college | 0.47 | | 0.44 | |
| Vocational degree: College | 0.03 | | 0.04 | |
| Family status: Single | 0.42 | | 0.38 | |
| Family status: Married | 0.26 | | 0.30 | |
| Family status: Divorced/widowed | 0.23 | | 0.25 | |
| Family status: Cohabiting | 0.08 | | 0.08 | |
| Has partner | 0.34 | | 0.36 | |
| Partners vocational degree: None | 0.20 | | 0.20 | |
| Partners vocational degree: Non-college | 0.13 | | 0.13 | |
| Partners vocational degree: College | 0.01 | | 0.01 | |

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

However, in contrast to the sample studied in Lechner and Wunsch (2013) our sample consists to a far higher extent of people who did not work for various years. Therefore, we included in more detail covariates on the labour market history of the last five years.

As Table 1 shows, treatment and control units, with 218 versus 162 days in regular employment in a three-year period after treatment, differ in terms of our outcome variable of interest. There are also considerable differences in pre-treatment characteristics.

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

The mean values of education and family status and partner characteristics included in Table 1 in most cases do not differ remarkably between treated and control individuals. Nevertheless, these descriptive statistics show that selection into treatment is non-random with respect to some variables. The rest of this paper is therefore concerned with modelling selection on these observable characteristics based on our extensive set of potential confounders.

# 4 Methodology

## 4.1 Target, notation and identification

In the following, we will use the notation for treatment effects estimation using the potential outcome framework of Rubin (1974). Participation in a training programme, as discussed in Section 3.2, is indicated with $D_i$ as the binary treatment variable, while $D_i = 1$

indicates that individual $i$ $(i = 1,...,N)$ takes part in a training programme and $D_i = 0$, otherwise. The outcome variable $Y_i$ denotes accumulated days in employment of individual *i* three years after the treatment. Let $Y_i^d := Y_i(D_i = d)$ denote the potential outcome if individual *i* receives treatment $d \in \{0,1\}$.[7] Since each individual can only receive either treatment or non-treatment one potential outcome is observable, the other remains counterfactual: $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$. While this implies that individual treatment effects are not directly observable, imposing assumptions may make it possible to identify treatment effects at various aggregation levels, e.g. the average treatment effect (ATE): $\tau = E(Y_i^1 - Y_i^0)$. The focus of this work is on the ATET, i.e. $\theta = E(Y_i^1 - Y_i^0 \mid D_i = 1)$.

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

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

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

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

A.1: $Y_i^1, Y_i^0 \perp D_i \mid X_i = x, \quad \forall x \in \chi$, Conditional Independence Assumption (CIA)

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

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

A.4: $Y_i = Y_i^1 D_i + Y_i^0 (1 - D_i)$, Stable Unit Treatment Value Assumption (SUTVA)

A.1 might be relaxed to $Y_i^0 \perp D_i \mid X_i = x$ for the case of ATET estimation. This assumption ensures that all confounders are observed and rules out the existence of further (unobserved) confounders jointly influencing the treatment and the potential outcome under non-treatment conditional on the observed *X*, or in this case conditional on the PS. A.2 ensures common support by bounding the treatment probability away from 0 and 1, and can also be relaxed in ATET estimation to $p(x) < 1$. The two latter assumptions require that covariates are not affected by the treatment (A.3) and that there are no spill over effects between the treatment groups (A.4). Under A.1-A.4, we have:

$$\begin{aligned} \theta &= E[Y_i^1 \mid D_i = 1] - E[Y_i^1 \mid D_i = 0] \\ &= E[Y_i \mid D_i = 1] - E[E[Y_i \mid D_i = 0, p(x)] \mid D_i = 1] \end{aligned}$$

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

## 4.2 Empirical Monte Carlo Simulation

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

To be as close as possible to applications in the empirical research literature, Huber, Lechner, and Wunsch (2013), and Lechner and Wunsch (2013) developed a so-called Empirical

Monte Carlo Study (EMCS). The idea is to use a DGP that exploits the structure of an empirical dataset to its full extent. For example, outcomes and covariates of real data are used. Of course, there are limitations, since the researcher needs to control some features to allow for generalizations, like the sample size or the share of treated in our case. Further, the empirical dataset must be large enough to plausibly presume that the random samples come from an infinite population. This is the case for our data as described in Section 3, which is a typical large-scale administrative dataset.

Every EMCS used to evaluate a treatment effects model consists of three basic steps. First, a true PS is estimated in the full population.[8] Second, a sample is drawn from the control units, a placebo treatment is simulated according to the true PS and the effects are estimated in this sample. Last, this is repeated many times and the performance is evaluated.

*Table 2: Empirical Monte Carlo Study*

| | |
|---|---|
| 1) | The PS is estimated in the full data. The true score is constructed as a combination of the separately estimated scores using the Probit, LASSO and Random Forest as: $$\hat{p}^{true}(x) = \frac{1}{3}\left(\hat{p}_{\mathrm{Pr}obit}(x) + \hat{p}_{LASSO}(x) + \hat{p}_{RandomForest}(x)\right)$$ |
| 2) | Remove all the treated observations from the population.[9] |
| 3) | Draw a sample of N units from the (remaining) population of control observations and simulate a placebo treatment in this draw, for which the treatment effect is zero by definition, as: $$d \sim Bernoulli(\hat{p}^{true}(x) \times \phi),$$ where $\phi \in \{2,5\}$ is to modify the share of (placebo) treated.[10] |
| 4) | Estimate the PS in the sample using the different estimation techniques described in Section 4.4 and use those respective PS to estimate the ATET using the RMBA estimator described in Section 4.3. |
| 5) | Repeat step 3&4 *R* times. |
| 6) | Calculate performance measures. |

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

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

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

We look at various performance measures, when evaluating the performance. First, the bias is calculated as mean of the deviation from the true effect, i.e. $bias = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta_0)$. $\hat{\theta}_r$ is the estimated ATET of the matching step in repetition *r*, $\theta_0$ the true effect (which is equal to zero since we discard all treated units). Most important is the mean squared error (MSE) of the ATET, calculated as $MSE = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta_0)^2$. Other measures we look at are the mean absolute deviation (MAD), kurtosis, skewness, the mean of the estimation (standard) error in the matching step, as well as the variance of $\hat{\theta}_r$. Further common support statistics are reported, as the mean share of observations, as well as the mean share of treated observations remaining in the common support. To investigate the performance of the first-stage estimation, we look at how well the various methods do in the PS estimation. Here we report the mean correlation of the estimated with the true PS, as well as the (in-sample) prediction MSE. Since radius matching compares treated and non-treated units, which are close to each other in terms of PS, the correct ordering of the estimated PS is important. We show two statistics for this, namely the (mean of) Kendall's Tau and the (mean of the) Spearman Rank Correlation coefficient.[11]

According to the procedure presented in Table 2, we simulated four different scenarios with two different treatment shares and two different sample sizes (see Table 3). We use 10 and 25 percent as treatment shares, because the number of treated is usually much smaller than the number of controls in active labour market programme evaluations. Similarly, samples smaller than our minimum sample size of 4'000 observations rarely occur in observational studies in the labour market context. The maximum of 16'000 observations is chosen due to the increasing computational burden of larger samples.

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

Another parameter to determine in simulations is the number of repetitions, *R*. Ideally, one would like to set this parameter as large as possible to minimize simulation noise. Since this noise depends on the variance of the estimators, which declines with sample size, we repeated each estimation for the smaller sample 1000 times and the larger sample 250 times. In case of $\sqrt{N}$ -convergence, this will keep the simulation error approximately constant.

*Table 3: Summary of DGP's*

| Scenario | Treatment share | Sample size (N) | Repetitions (R) |
|---|---|---|---|
| A | 10 % | 4000 | 1000 |
| B | 25 % | 4000 | 1000 |
| C | 10 % | 16000 | 250 |
| D | 25 % | 16000 | 250 |

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

## 4.3 Matching estimator

While there are several different matching algorithms available, we use the bias-adjusted-radius-matching-on-the-propensity-score estimator (RMBA) of Lechner, Miquel, and Wunsch (2011). This estimator combines the features of distance-weighted radius matching with bias adjustment to remove biases due to mismatches and performed well in Huber, Lechner, and Wunsch (2013).[12]

It has been shown by Lechner and Strittmatter (2019), among others, that trimming treated observations may be important if there is thin or even lacking support to guard against bias and excessive importance of specific control variables. In the setup of this work trimming does not change the ATET, since the true treatment effect is homogenous (and zero) by construction. The trimming rule used follows the recommendation of Lechner and Strittmatter (2019) and

[12] The radius is determined data-driven as 1.5 times the maximum pair matching distance as suggested by Lechner, Miquel, and Wunsch (2011).

removes too important, i.e. control units with a weight larger than 5 percent, and off-support observations jointly for treated and controls.

## 4.4 Propensity score estimation

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

The other three approaches are choices researchers might use in their work, namely a Probit, a Random Forest, and a LASSO-based estimator. While those methods are known to be good prediction techniques there is little knowledge how they perform in empirical labour market evaluation studies for estimating a causal effect in matching estimators. We describe each of the estimation techniques used in the following in more detail, as well as how they are implemented in the EMCS.

### 4.4.1 Probit

Since the PS is the probability of receiving the treatment conditional on the confounders, the Probit estimation, especially in the past, was the usual choice for this first step estimation.[13] $\hat{p}(x) = \Phi(x\hat{\beta})$ is estimated for each individual, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. This parametric, non-linear technique is well suited for those kinds of prediction problems if the following four conditions are satisfied. 1) The true selection equation is well approximated by the Probit link function. 2) The set of confounding characteristics and their relevant measurement (i.e. logs, particular polynomials, etc.) is known. 3) The required functional flexibility of the covariates, in particular with respect to interactions of the variables can be well approximated by the researcher. 4) The final set of covariates (incl.

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

all terms that enter the linear index in the probit link function) is not too large with respect to sample size.

Usually in observational studies ensuring conditions 1) to 3) is subject to a credible line of argumentation, and in most cases, even with a strong intuition, hard to specify correctly. Further, including every variable and functional transformation thereof contradicts the fourth condition in most settings. Too many covariates may decrease the precision of the estimator or may make estimation numerically infeasible.[14]

### 4.4.2 LASSO

The LASSO as proposed by Tibshirani (1996) is a shrinkage estimator, which works like an OLS estimator with penalized coefficients. Since we are estimating a probability-like quantity, we oppose this potential issue of predicting values below 0 or above 1 by using a Logit version of the LASSO[15]. Therefore, the following minimization function is used:

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

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

The last term in equation (1) is to penalize the size of the $j=1,...,k$ coefficients, with $k$ being the number of covariates. $\lambda$ represents the penalty term. The larger this penalty term, the more the coefficients are pushed towards zero and variable selection takes place, i.e. coefficient become exactly zero. The idea behind this procedure is to shrink the coefficients of those covariates to zero that contain no or little predictive information about the dependent variable.[16]

Determining the size of the penalty term is therefore crucial. This choice represents a trade-off between bias, which $\lambda$ increases, and variance, which decreases when $\lambda$ increases.

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

15 Compare Hastie, Tibshirani, and Friedman (2009, p. 125).

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

Here, the penalty term is chosen by 5-fold cross-validation minimizing the out-of-sample mean squared error (MSE).

### 4.4.3 Random Forest

In the machine learning literature, the Random Forests algorithm developed by Breiman (2001) is a widely used non-parametric and non-linear estimation technique. It is built as an ensemble of Regression Trees, which are to some extent randomly constructed. A Regression Tree recursively splits the covariate space into separate non-overlapping areas as it minimizes the MSE of the prediction of the outcome. The resulting structure is reminiscent of a rotated tree, as one observes the trunk with all the observations in the beginning, which is split into finer branches the further you go down. The tree predictions are the average of the outcome of those observations falling into the same end-nodes, so called leaves.

Like in LASSO, there is a trade-off between bias and variance: Deeply grown trees have lower bias and higher variance compared to shallow trees. This trade-off is controlled by specifying the minimum number of observations in each leaf.[17] For a Random Forest several deep, low-bias trees are estimated on random subsamples and the predictions are averaged over those trees.[18] In our simulations, 600 trees are built for each forest. The more trees are estimated, the smoother the prediction become, but computation time increases. Further, to de-correlate the trees only a random subset of covariates is considered at every split point within the tree building process. [19]

Finally, we use the so-called honest splitting rule, as proposed by Athey and Imbens (2016). Using independent samples for building the tree and for making the predictions contributes to higher prediction accuracy. This comes with the price to pay in terms of reduced

17 In our simulations, we used a minimum leaf size of five observations.
18 The random subsamples can be generated by either bootstrapping or subsampling. We follow the recommendation of Wager and Athey (2018) to use subsampling. In the simulations and application, the subsampling size is a share of 50 percent of the sample size.
19 In the simulations and application, the number of covariates is chosen to be 50.

sample sizes. As an example, in the N=4'000 setting only 1'000 observations are used to build the tree structures, another 1'000 to do the predictions.[20]

### 4.4.4 Sets of covariates

The Methods described above are able to work with different kinds of variables (as described in Section 3) in other ways, and therefore the sets of covariates in the PS estimation differ for each method. Probit and LASSO cannot distinguish between ordered and unordered categorical variables. Unordered variables are therefore split into binary variables for each category.[21] This results in 309 covariates for the Probit estimation.

Since the LASSO has a variable selection property, it is able to solve the objection of including too many covariates up to a certain degree.[22] To be more flexible, we are able to increase the set of potential confounding variables by including second-order polynomials and interactions of all continuous variables resulting in a full set of 1'011 covariates available for the LASSO. Of course, ideally, one would like to include interactions up to a higher degree, to be as flexible as possible, but since the potential set of covariates increases exponentially computational resources are quickly exhausted.

The Random Forest is able to work with unordered categorical variables, while in the other methods dummies are used instead.[23] Further, there is no need to include transformation of variables, like polynomials and interactions in the LASSO, as the tree structures are able to incorporate any interactive and non-linear nature of the covariate structure. Therefore, this method ensures a very large degree of flexibility, as it can, at least asymptotically, pick-up any non-linearity. The set of covariates is therefore substantially lower, i.e. 109 covariates, compared to the other methods. Still, this is only another way to work with the same information

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

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

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

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

and there should be no advantage or disadvantage compared to the other methods. To reduce the computational burden binary variables representing less than 2 percent of the observations are removed in all the methods, as well as we only keep one if there are multiple covariates, which show correlations of more than $\pm 0.98$ in the respective sample.

# 5 Simulation

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

Before discussing the results, as this may be an important issue in applied research, we like to point out convergence problems of the Probit estimation in the small sample. We report the results for all repetitions in the main results. Further, we report the results for only converged approaches in the Appendix, as common practice in the literature is rather to modify the specification of the Probit instead of using a non-converged PS in applied work. The results do slightly differ, but the general conclusions are equivalent, however, this points to difficulties in using the Probit in settings with low number of observations and a large set of confounders, especially if the share of treated is low.[24]

*Table 4: Summary of Simulation Results, Propensity Score Estimation*

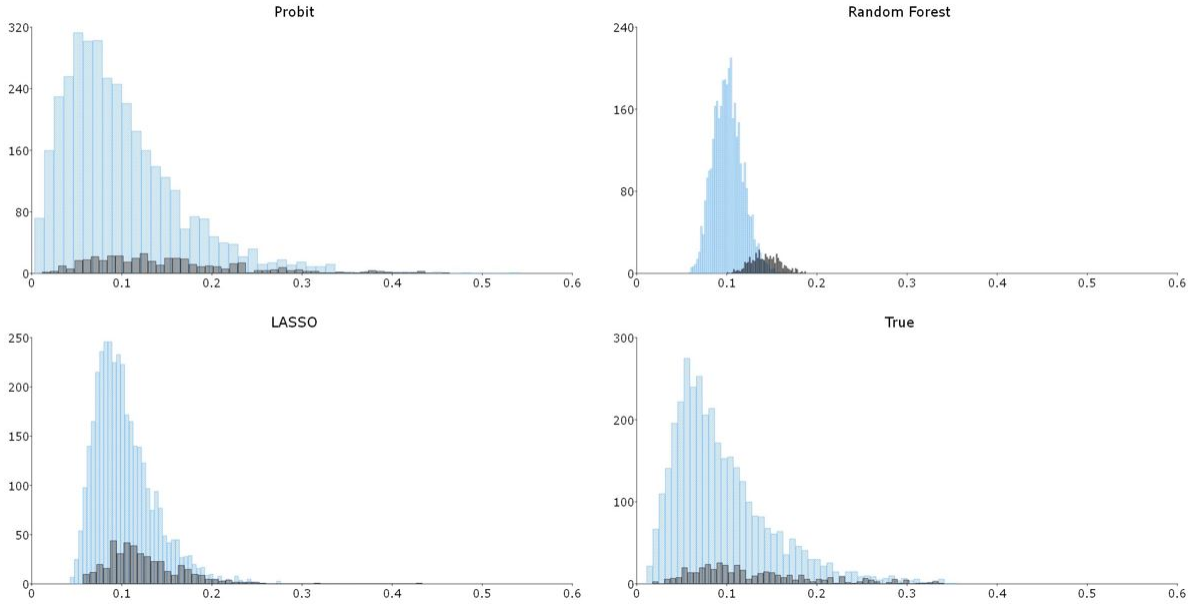
| | Spearman Rank Correlation | | | | MSE | | | |
|---|---|---|---|---|---|---|---|---|
| | N = 4000 | | N = 16000 | | N=4000 | | N=16000 | |
| Treatment share | 10% | 25% | 10% | 25% | 10% | 25% | 10% | 25% |
| Probit | 0.36 | 0.60 | 0.73 | 0.87 | 8.50 | 17.25 | 8.56 | 16.60 |
| Random Forest | 0.72 | 0.82 | 0.81 | 0.86 | 8.19 | 16.72 | 8.16 | 15.92 |
| LASSO | 0.77 | 0.86 | 0.87 | 0.92 | 8.62 | 16.58 | 8.64 | 16.56 |
| True | - | - | - | - | 8.60 | 16.53 | 8.63 | 16.54 |
| Random | 0.00 | 0.00 | 0.00 | 0.00 | 9.30 | 20.90 | 9.33 | 20.90 |

*Notes:* Figures shown are the mean of the Spearman Rank Correlation of the estimated PS compared to the true PS, as well as the (in-sample) MSE (times 100) of the prediction over 1'000, respectively 250 simulation repetitions. The full results can be found in Tables A.1.2, A.2.2, A.3.2 & A.4.2 in the Appendix. True and Random indicates the true, respectively the randomized PS.

[24] For N=4'000 and 10 percent treated about 35 percent of replications, for 25 percent treated about 4 percent of replications did not converge. In the larger samples, this problem is not present. Compare Tables A.1.1, A.1.2, A.2.1 and A.2.2 in the Appendix.

To investigate the performance of the PS estimation in Table 4 we find the (in-sample) prediction MSEs to show the Random Forest predicting best. More important is the ordering of the PS determining which control units are matched to the respective treated units. The results of the Spearman Rank Correlation with the true PS are depicted in Table 4. We find every method to perform better in those settings with higher treatment shares and/or more observations. The Random Forest and the LASSO both reach the highest rank correlations, while the Probit is doing rather poor in the small samples. With more observations, i.e. effectively a lower number of covariates relative to observations, the Probit becomes more competitive and reaches a higher Spearman Rank Correlation compared to the Random Forest in the higher treatment share. This may indicate that the underlying model is well approximated by the probit functional form. Further, as expected, the random PS obtains values of (close to) zero.

*Figure 1: Propensity scores by treatment status, N=4000, 10% treated*

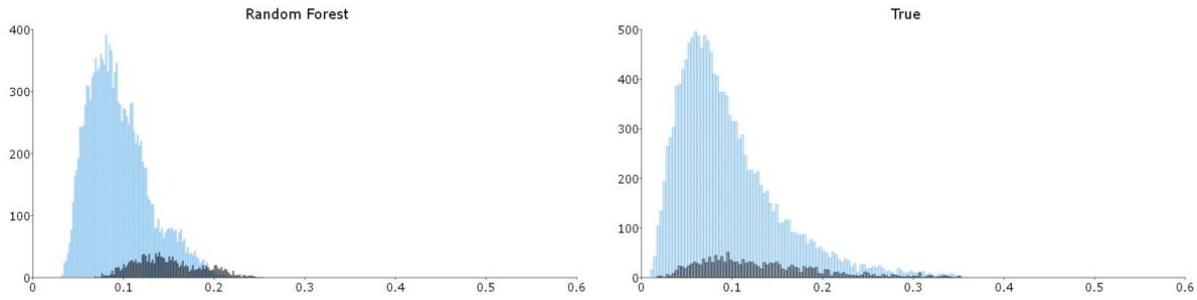



Notes: Histograms with PS on the horizontal axis. Top left is the Probit PS, top right Random Forest, bottom left and right the LASSO estimated and true PS. Each from the same one simulation with N=4'000 and 10% treatment share. Control units are light, treated units dark shaded.

For the Random Forest, having a low treatment share may contribute to splitting less deeply than it should.[25] Therefore having a higher treatment share enables the growing of deeper trees, which might be necessary for balancing the covariates in the matching estimator. Figure 1 provides some insights into the estimated PS of the respective methods, as well as the true PS for the small sample and low treatment share.[26]

First to note is that the Random Forest in the top right graph looks quite different to the other estimates, as well as the true PS. Not being able to split deep enough leads to a narrower distribution of estimated PS and treated and controls are more separated compared to the other methods. On the one hand, this reduced overlap leads to lower common support. On the other hand, this might lead to matching "wrong" control to the respective treated units. Despite there might be a tendency towards a wider spread of the Random Forest in the larger sample, Figure 2 shows generally a similar pattern.

*Figure 2: Propensity scores by treatment status, N=16000, 10% treated*




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

To investigate this further, we provide matching quality measures in Table 5 showing which quantiles of the distribution of control units' PS are matched in the simulation to the respective quantiles of the distribution of the treated PS.

[25] Having a low share of treated, i.e. a large number of zeros and a low number of ones, in the outcome variables makes it more likely that there cannot be any improvement in terms of MSE by splitting a certain leaf, leading to large final leaves after few splits. In fact, the average leaf size for the Random Forest is larger in both simulations with the low treatment share compared to the higher treatment share.

[26] Figures for the other simulations scenarios can be found in Appendix B.

*Table 5: Matching-Quality*

| | $q_{0.1}$ | $q_{0.3}$ | $q_{0.5}$ | $q_{0.7}$ |
|---|---|---|---|---|
| Panel A: | | N=4000, 10% treated | | |
| Probit | 0.31 (0.05) | 0.59 (0.05) | 0.76 (0.03) | 0.89 (0.02) |
| Random Forest | 0.80 (0.54) | 0.91 (0.36) | 0.96 (0.22) | 0.99 (0.11) |
| LASSO | 0.27 (0.03) | 0.55 (0.03) | 0.74 (0.02) | 0.88 (0.01) |
| True | 0.26 (0.00) | 0.54 (0.00) | 0.74 (0.00) | 0.88 (0.00) |
| Panel B: | | N=4000, 25% treated | | |
| Probit | 0.27 (0.02) | 0.56 (0.04) | 0.74 (0.04) | 0.88 (0.04) |
| Random Forest | 0.46 (0.17) | 0.69 (0.10) | 0.83 (0.04) | 0.94 (0.03) |
| LASSO | 0.30 (0.02) | 0.61 (0.02) | 0.79 (0.01) | 0.91 (0.01) |
| True | 0.29 (0.00) | 0.59 (0.00) | 0.79 (0.00) | 0.91 (0.00) |
| Panel C: | | N=16000, 10% treated | | |
| Probit | 0.28 (0.05) | 0.56 (0.06) | 0.73 (0.06) | 0.85 (0.05) |
| Random Forest | 0.66 (0.40) | 0.82 (0.27) | 0.91 (0.17) | 0.97 (0.10) |
| LASSO | 0.26 (0.01) | 0.55 (0.01) | 0.74 (0.01) | 0.88 (0.01) |
| True | 0.26 (0.00) | 0.54 (0.00) | 0.74 (0.00) | 0.88 (0.00) |
| Panel D: | | N=16000, 25% treated | | |
| Probit | 0.30 (0.01) | 0.60 (0.01) | 0.79 (0.00) | 0.91 (0.00) |
| Random Forest | 0.47 (0.18) | 0.69 (0.09) | 0.85 (0.07) | 0.94 (0.02) |
| LASSO | 0.30 (0.01) | 0.60 (0.01) | 0.79 (0.00) | 0.91 (0.00) |
| True | 0.29 (0.00) | 0.59 (0.00) | 0.79 (0.00) | 0.91 (0.00) |

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

As can be seen in Table 5 in every panel the Random Forest estimates lead to the most distinct matching of quantiles. This is most pronounced in the scenarios of low treatment shares. Of course, the matching quantiles of the true PS is not necessarily the best, but a valid benchmark. While the LASSO is in most situations the closest to the true PS matching quantiles, the Random Forest is, especially in the 10 percent quantile far away from the true PS results. Despite the "matching-quality" becoming closer to the true PS for the higher quantiles, the estimates of the Random Forest does not seem to work well, especially in low treatment shares, in the context of matching-type estimators. Table 6 shows the observed final performance of the estimated PS in the RMBA estimator.

*Table 6: Summary of Simulation Results, Matching*

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Bias | MSE | Variance | CS (%) | CS (%), treated | SB |
| Panel A: | | | N=4000, 10% treated | | | |
| Probit | 21.95 | 885.52 | 403.54 | 92.8 | 64.7 | 8.20 |
| RF | -26.15 | 2258.05 | 1574.16 | 56.7 | 90.2 | 28.18 |
| LASSO | 5.03 | 398.29 | 372.95 | 98.1 | 99.4 | 5.49 |
| True | -0.39 | 341.25 | 341.10 | 98.3 | 99.6 | 5.33 |
| Random | 20.47 | 773.07 | 353.89 | 99.6 | 99.9 | 16.34 |
| Panel B: | | | N=4000, 25% treated | | | |
| Probit | 11.68 | 310.55 | 174.05 | 98.0 | 95.6 | 3.13 |
| RF | -2.18 | 275.33 | 270.57 | 94.2 | 97.4 | 9.41 |
| LASSO | 3.63 | 213.93 | 200.73 | 98.9 | 99.1 | 4.03 |
| True | -0.32 | 226.28 | 226.18 | 99.0 | 99.0 | 4.06 |
| Random | 24.29 | 762.48 | 172.80 | 99.9 | 99.9 | 19.46 |
| Panel C: | | | N=16000, 10% treated | | | |
| Probit | 1.56 | 109.86 | 107.45 | 99.1 | 95.1 | 2.47 |
| RF | -12.40 | 440.31 | 286.46 | 74.9 | 96.8 | 17.89 |
| LASSO | 1.40 | 86.05 | 84.08 | 99.4 | 99.9 | 2.67 |
| True | -0.19 | 95.90 | 95.86 | 99.5 | 99.9 | 2.70 |
| Random | 20.63 | 507.72 | 82.22 | 99.9 | 99.2 | 16.09 |
| Panel D: | | | N=16000, 25% treated | | | |
| Probit | 2.63 | 49.80 | 42.87 | 99.7 | 99.3 | 1.56 |
| RF | 1.10 | 72.73 | 71.50 | 95.3 | 98.6 | 8.50 |
| LASSO | 1.15 | 42.34 | 41.03 | 99.7 | 99.8 | 2.19 |
| True | -0.72 | 53.62 | 53.10 | 99.7 | 99.4 | 2.04 |
| Random | 24.52 | 641.45 | 40.42 | 99.9 | 99.9 | 19.39 |

*Notes:* Figures shown are the mean of the respective measure over 1'000 (in Panel A&B) or 250 (in Panel C&D) replications. *RF* stands for Random Forest. *Random* indicates the randomized PS. *Bias* is the mean bias over all simulation repetitions. *MSE* is the mean squared error. *CS* and *CS, treated* is the common support (for the treated) and *SB* is the (mean) absolute standardized bias in covariate balancing of the ten most important confounders. The full results can be found in Tables A.1.1, A.2.1, A.3.1 & A.4.1 in the Appendix.

In column (6) of Table 6, we observe the absolute mean standardized bias in covariate balancing (SB), which is one rough measure of how well the covariates are balanced using the respective PS estimate.[27] While the balancing ability of the Probit increased considerably in Panels B-D compared to Panel A, the seemingly good Random Forest prediction led to rather worse covariate balancing. For the higher treatment shares the balancing statistics is acceptable. The true and the LASSO PS showed good balancing properties throughout the results.

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

Although it is not clear how low the SB should be and if this translates directly into good final ATET estimates this is indicative for the bad performance of the Random Forest in the matching step with 10 percent treated as can be seen in Panels A and C of Table 6, columns (1)-(3). The LASSO PS is only slightly biased and the resulting MSE is the lowest despite the true PS results in Panel A and even lower than the true PS in Panel C (compare Abadie and Imbens (2016) for this phenomenon). Panels B and D are giving some insights into the simulations with the higher treatment share. All estimation techniques performed better compared to the lower treatment share, with the LASSO outperforming the other PS in terms of MSE and MAD. More observations, as can be seen in Panels C and D, generally improves the performance of every method. Estimating the PS with the Probit is benefiting from the larger sample especially by reducing the mean bias compared to the low observations scenarios. The Random Forest PS works decently well with 25 percent treated units, i.e. the bias is closest to zero, but is biased with a lower share of treated and has the highest variance in every scenario.

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

# 6 Empirical application

We evaluate the effect of participating in the training programme, “Determining, Reducing and Removing Employment Impediments”, using the full sample of 14’817 treated and 261’820 control units as described in Section 3. The ATET is estimated using the three PS methods, i.e. Random Forest, LASSO and Probit, in the RMBA estimator. The results can be found in Table 7.

*Table 7: Empirical Treatment Effect Estimation, Matching*

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

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

Although LASSO (and Probit) performed well in our simulation exercise as PS estimation technique for 16'000 observations, this gives us only little indication how this performance translates into this larger sample. Having an even lower treatment share as in the simulations of about five percent, but a larger sample, the expected performance of the Random Forest is unclear.[28]

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

[28] In Appendix B.4, we show the distributions of the PS are very similar for the Probit and the LASSO, while the Random Forest estimates a slightly narrower distribution.

*Table 8: Covariate balancing in application*

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

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

In Table 8, we provide the covariate balancing statistics for the ten most important confounders. While the Probit is balancing every covariate well, the Random Forest PS shows deficits in balancing some of the, especially non-binary, variables.[29] In conclusion, the choice of the first stage estimator does matter in practical research and choosing a non-appropriate method could lead to wrong policy conclusions.

# 7 Conclusion

In this work, we investigated through simulations and an application whether predicting the PS by machine learning methods helps to increase credibility of programme evaluation studies based on propensity score matching. Having an arguably realistic DGP using a rich, high-dimensional administrative dataset for German long-term unemployed, we simulated the finite sample performance of various PS estimation techniques in a matching-type estimator estimating the ATET. We considered two very different methods from the machine learning

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

literature, namely the Random Forest and the LASSO. We compared their performance to a "classical" Probit approach with an ad-hoc specification of covariates, as well as to the true and a randomized PS.

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

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

The conclusion of these exercises is that estimating the propensity score by machine learning is not clearly beneficial compared to current conventional matching methods. Instead, the methods of the new causal machine literature that are directly optimized for treatment effect

estimation may be a more promising alternative, although this is beyond the scope of this paper (see Knaus, Lechner, and Strittmatter, 2018, and Lechner, 2018, for various proposals and comparisons).

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

# References


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

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

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

Athey, S., & Imbens, G. (2016). Recursive Partitioning for Heterogeneous Effect. *Proceedings of the National Academy of Sciences, 113*(27), 7353-7360.

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

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

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

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

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

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

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

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

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

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

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

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

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

Goller, D., & Krumer, A. (2019). Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues. *SEPS Discussion Paper, 2019-07*, 1-35.

Harrer, T., Moczall, A., & Wolff, J. (2019). Free, free, set them free? Are programmes effective that allow job centres considerable freedom to choose the exact design? *forthcoming in International Journal of Social Welfare*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data mining, inference, and prediction.* 2nd. ed. New York: Springer.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research, 46*(3), 477-513.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*, 199-236.

Huber, M., Lechner, M., & Steinmayr, A. (2015). Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour. *Empirical Economics, 49*(1), 1-31.

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

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(1), 4-29.

Imbens, G. (2015). Matching Methods in Practice: Three Examples. *Journal of Human Resources, 50*(2), 373-419.

Imbens, G., & Rubin, D. (2015). *Causal inference: For statistics, social, and biomedical sciences an introduction.* Cambridge University Press.

Knaus, M., Lechner, M., & Strittmatter, A. (2018). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *arXiv:1810.13237v2*.

Krumer, A., & Lechner, M. (2018). Midweek effect on soccer performance: Evidence from the German Bundesliga. *Economic Inquiry, 56*(1), 193-207.

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. *IZA Discussion Paper Series, No. 12040*.

Lechner, M., & Strittmatter, A. (2019). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews, 38*(2), 193-207.

Lechner, M., & Wunsch, C. (2009). Are Training Programs More Effective When Unemployment Is High? *Journal of Labor Economics, 27*(4), 653-692.

Lechner, M., & Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics, 21*, 111-121.

Lechner, M., Miquel, R., & Wunsch, C. (2011). Long-run effects of public sector sponsored training in West Germany. *Journal of the European Economic Association, 9*(4), 742-784.

Lee, B., Lessler, J., & Stuart, E. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346.

Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryan, T., Cleary, P., & McNeil, B. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*(4), 387-398.

Pirracchio, R., Petersen, M., & Van Der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using Super Learner. *American Journal of Epidemiology, 181*(2), 108-119.

Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrica, 70*(1), 41-55.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*, 20-36.

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., & Cook, E. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*(6), 546-555.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*(1-2), 305-353.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B, 58*(1), 267-288.

van der Laan, M., Polley, E., & Hubbard, A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1), 1-21.

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

Wunsch, C., & Lechner, M. (2008). What did all the money do? On the general ineffectiveness of recent west German labour market programmes. *Kyklos, 61*(1), 134-174.

# Appendices

## Appendix A: Full result tables

In this Appendix, we show the full results tables of the EMCS presented in Section 5. The following four subsections refer to the four simulation scenarios. Summaries of those tables are found in the text.

### A.1 Scenario A: N = 4000, 10% treated

*Table A.1.1: Simulation results for N=4000 and ~10% share of treated*

| Measures | Probit | Probit (conv.) | Random Forest | LASSO | True | Random |
|---|---|---|---|---|---|---|
| | | | Treatment effects | | | |
| Mean treatment effect / bias | 21.95 | 16.60 | -26.15 | 5.03 | -0.39 | 20.47 |
| Mean SE of matching[1)] | 19.65 | 21.76 | 31.03 | 20.07 | 20.17 | 19.41 |
| MAD | 25.11 | 21.44 | 36.12 | 15.91 | 14.83 | 23.12 |
| MSE | 885.52 | 706.23 | 2258.05 | 398.29 | 341.25 | 773.07 |
| SE | 20.09 | 20.75 | 39.68 | 19.31 | 18.47 | 18.81 |
| Variance | 403.54 | 430.75 | 1574.16 | 372.94 | 341.10 | 353.89 |
| Skewness | -0.27 | 0.002 | -0.78 | -0.16 | -0.02 | 0.08 |
| Kurtosis | 3.05 | 2.98 | 4.75 | 2.95 | 2.77 | 3.32 |
| | | | Common support | | | |
| Mean share remaining in CS | 0.93 | 0.91 | 0.57 | 0.98 | 0.98 | 0.99 |
| Mean share treated remaining in CS | 0.65 | 0.99 | 0.90 | 0.99 | 0.99 | 0.99 |
| | | | Balancing of covariates as standardized differences | | | |
| Mean abs. stand. mean bias | 8.20 | 3.74 | 28.18 | 5.49 | 5.33 | 16.34 |
| Mean abs. stand. max. bias | 19.14 | 8.65 | 106.29 | 12.47 | 12.51 | 37.54 |
| Sample size | 4000 | 4000 | 4000 | 4000 | 4000 | 4000 |
| Replications | 1000 | 653 | 1000 | 1000 | 1000 | 1000 |
| Share of treated | 0.0993 | 0.0935 | 0.0993 | 0.0993 | 0.0993 | 0.0993 |

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

*Table A.1.2: Propensity score estimation results for N=4000 and ~10% share of treated*

| Measure | Probit | Probit (conv.) | Random Forest | LASSO | Random |
|---|---|---|---|---|---|
| Mean correlation | 0.36 | 0.56 | 0.70 | 0.75 | 0.00 |
| Mean Kendall's Tau | 0.26 | 0.39 | 0.53 | 0.58 | 0.00 |
| Mean Spearman Rank | 0.36 | 0.56 | 0.72 | 0.77 | 0.00 |
| Sample size | 4000 | 4000 | 4000 | 4000 | 4000 |
| Replications | 1000 | 653 | 1000 | 1000 | 1000 |
| Share of treated | 0.0993 | 0.0935 | 0.0993 | 0.0993 | 0.0993 |

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

## A.2 Scenario B: N = 4000, 25% treated

*Table A.2.1: Simulation results for N=4000 and ~25% share of treated*

| Measures | Probit | Probit (conv.) | Random Forest | LASSO | True | Random |
|---|---|---|---|---|---|---|
| Treatment effects | | | | | | |
| Mean treatment effect / bias | 11.68 | 11.27 | -2.18 | 3.63 | -0.32 | 24.29 |
| Mean SE of matching[1)] | 14.51 | 14.66 | 16.63 | 14.84 | 15.02 | 13.10 |
| MAD | 14.52 | 14.22 | 13.14 | 11.50 | 12.10 | 24.63 |
| MSE | 310.55 | 299.99 | 275.33 | 213.92 | 226.28 | 762.48 |
| SE | 13.19 | 13.15 | 16.45 | 14.17 | 15.04 | 13.15 |
| Variance | 174.05 | 172.99 | 270.57 | 200.73 | 226.18 | 172.80 |
| Skewness | -0.08 | -0.05 | -0.07 | -0.17 | 0.009 | 0.002 |
| Kurtosis | 3.17 | 3.21 | 3.03 | 3.37 | 2.89 | 2.87 |
| Common support | | | | | | |
| Mean share remaining in CS | 0.98 | 0.98 | 0.94 | 0.99 | 0.99 | 0.99 |
| Mean share treated remaining in CS | 0.96 | 0.99 | 0.97 | 0.99 | 0.99 | 0.99 |
| Balancing of covariates as standardized differences | | | | | | |
| Mean abs. stand. mean bias | 3.13 | 2.66 | 9.41 | 4.03 | 4.06 | 19.46 |
| Mean abs. stand. max. bias | 7.90 | 6.36 | 31.75 | 10.07 | 9.73 | 46.47 |
| Sample size | 4000 | 4000 | 4000 | 4000 | 4000 | 4000 |
| Replications | 1000 | 961 | 1000 | 1000 | 1000 | 1000 |
| Share of treated | 0.2493 | 0.2485 | 0.2493 | 0.2493 | 0.2493 | 0.2493 |

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

*Table A.2.2: Propensity score estimation results for N=4'000 and ~25% share of treated*

| Measure | Probit | Probit (conv.) | Random Forest | LASSO | Random |
|---|---|---|---|---|---|
| Mean correlation | 0.61 | 0.64 | 0.80 | 0.86 | 0.00 |
| Mean Kendall's Tau | 0.43 | 0.44 | 0.62 | 0.67 | 0.00 |
| Mean Spearman Rank | 0.60 | 0.62 | 0.82 | 0.86 | 0.00 |
| Sample size | 4000 | 4000 | 4000 | 4000 | 4000 |
| Replications | 1000 | 961 | 1000 | 1000 | 1000 |
| Share of treated | 0.2493 | 0.2485 | 0.2493 | 0.2493 | 0.2493 |

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

## A.3 Scenario C: N = 16000, 10% treated

*Table A.3.1: Simulation results for N=16000 and ~10% share of treated*

| Measures | Probit | Random Forest | LASSO | True | Random |
|---|---|---|---|---|---|
| | | Treatment effects | | | |
| Mean treatment effect / bias | 1.56 | -12.40 | 1.40 | -0.19 | 20.63 |
| Mean SE of matching[1)] | 9.98 | 13.39 | 10.04 | 10.06 | 9.70 |
| MAD | 8.12 | 16.82 | 7.63 | 7.71 | 20.63 |
| MSE | 109.86 | 440.31 | 86.05 | 95.90 | 507.72 |
| SE | 10.37 | 16.93 | 9.17 | 9.79 | 9.07 |
| Variance | 107.45 | 286.46 | 84.08 | 95.86 | 82.22 |
| Skewness | 0.48 | -0.24 | -0.03 | 0.07 | 0.17 |
| Kurtosis | 3.65 | 3.50 | 2.49 | 2.97 | 2.67 |
| | | Common support | | | |
| Mean share remaining in CS | 0.99 | 0.75 | 0.99 | 0.99 | 0.99 |
| Mean share treated remaining in CS | 0.95 | 0.97 | 0.99 | 0.99 | 0.99 |
| | | Balancing of covariates as standardized differences | | | |
| Mean abs. stand. mean bias | 2.47 | 17.89 | 2.67 | 2.70 | 16.09 |
| Mean abs. stand. maximum bias | 5.80 | 71.11 | 6.70 | 6.33 | 37.62 |
| Sample size: | | 16000 | | | |
| Replications: | | 250 | | | |
| Mean share of treated: | | 0.0997 | | | |

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

*Table A.3.2: Propensity score estimation results for N=16000 and ~10% share of treated*

| Measure | Probit | Random Forest | LASSO | Random |
|---|---|---|---|---|
| Mean correlation | 0.73 | 0.79 | 0.86 | 0.00 |
| Mean Kendall's Tau | 0.54 | 0.62 | 0.68 | 0.00 |
| Mean Spearman Rank | 0.73 | 0.81 | 0.87 | 0.00 |
| Sample size: | | 16000 | | |
| Replications: | | 250 | | |
| Mean share of treated: | | 0.10 | | |

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

## A.4 Scenario D: N = 16000, 25% treated

*Table A.4.1: Simulation results for N=16000 and ~25% share of treated*

| Measures | Probit | Random Forest | LASSO | True | Random |
|---|---|---|---|---|---|
| | | Treatment effects | | | |
| Mean treatment effect / bias | 2.63 | 1.10 | 1.15 | -0.72 | 24.52 |
| Mean SE of matching[1)] | 7.37 | 8.59 | 7.53 | 7.59 | 6.55 |
| MAD | 5.55 | 6.76 | 5.14 | 5.80 | 24.52 |
| MSE | 49.80 | 72.73 | 42.34 | 53.62 | 641.45 |
| SE | 6.55 | 8.46 | 6.41 | 7.29 | 6.36 |
| Variance | 42.87 | 71.50 | 41.03 | 53.10 | 40.42 |
| Skewness | 0.28 | 0.01 | -0.21 | 0.03 | 0.29 |
| Kurtosis | 4.18 | 2.98 | 3.37 | 3.31 | 2.86 |
| | | Common support | | | |
| Mean share remaining in CS | 0.99 | 0.95 | 0.99 | 0.99 | 0.99 |
| Mean share treated remaining in CS | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | | Balancing of covariates as standardized differences | | | |
| Mean abs. stand. mean bias | 1.56 | 8.50 | 2.19 | 2.04 | 19.39 |
| Mean abs. stand. maximum bias | 3.70 | 27.94 | 6.14 | 4.78 | 46.35 |
| Sample size: | | 16000 | | | |
| Replications: | | 250 | | | |
| Mean share of treated: | | 0.25 | | | |

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

*Table A.4.2: Propensity score estimation results for N=16'000 and ~25% share of treated*

| Measure | Probit | Random Forest | LASSO | Random |
|---|---|---|---|---|
| Mean correlation | 0.86 | 0.86 | 0.92 | 0.00 |
| Mean Kendall's Tau | 0.69 | 0.67 | 0.76 | 0.00 |
| Mean Spearman Rank | 0.87 | 0.86 | 0.92 | 0.00 |
| Sample size: | | 16000 | | |
| Replications: | | 250 | | |
| Mean share of treated: | | 0.25 | | |

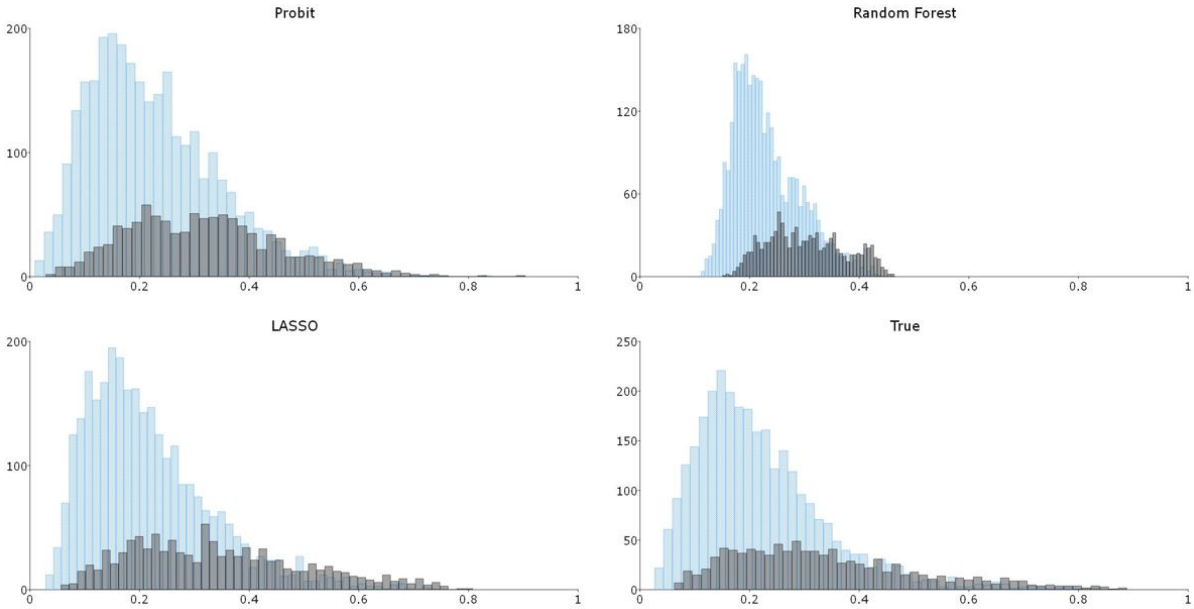
*Notes*: The formulas for Kendall's Tau and the Spearman Rank Correlation can be found in the main text.

# Appendix B: Estimated propensity score by treatment status

The distributions of the same one PS estimation for each scenario from the EMCS in Section 5 is presented in the appendices B.1 – B.3. Scenario A can be found in the main text. The distribution of the PS of the Probit, Random Forest and LASSO from the application in Section 6 are depicted in B.4.

## B.1 Scenario B: N=4000, 25% treated

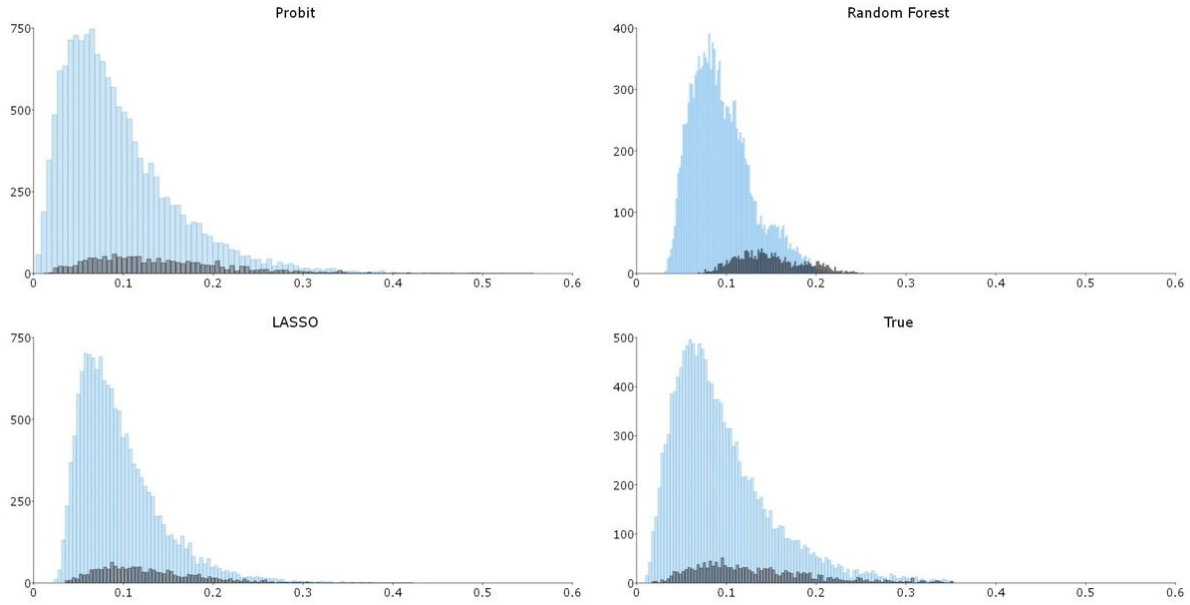
*Figure B.1: Propensity scores by treatment status*




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

## B.2 Scenario C: N=16000, 10% treated

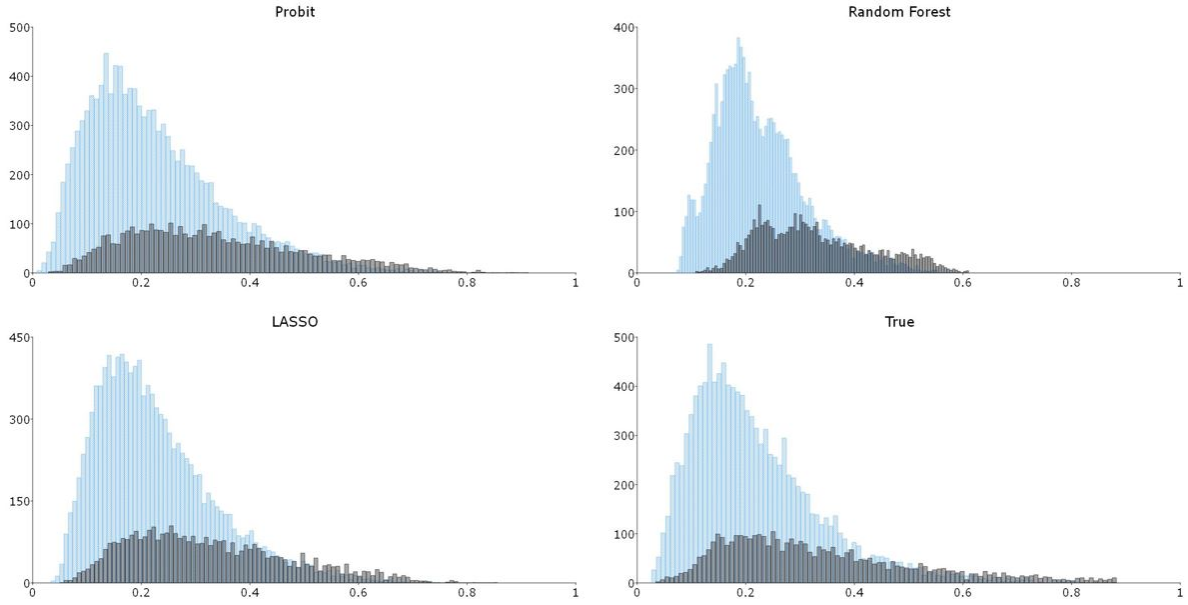
*Figure B.2: Propensity scores by treatment status*




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

## B.3 Scenario D: N=16000, 25% treated

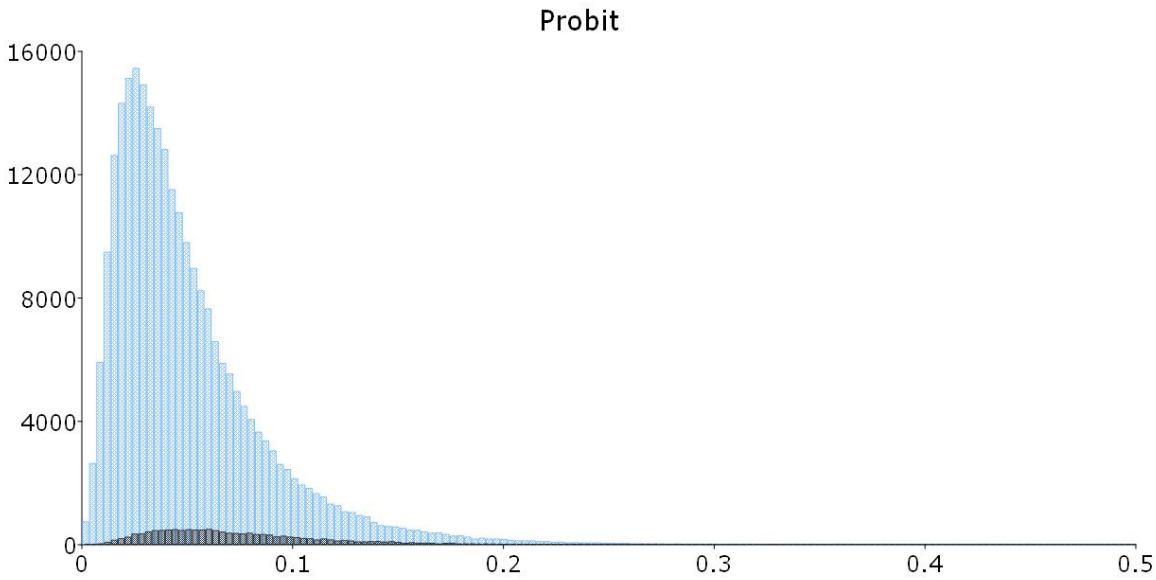
*Figure B.3: Propensity scores by treatment status*




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

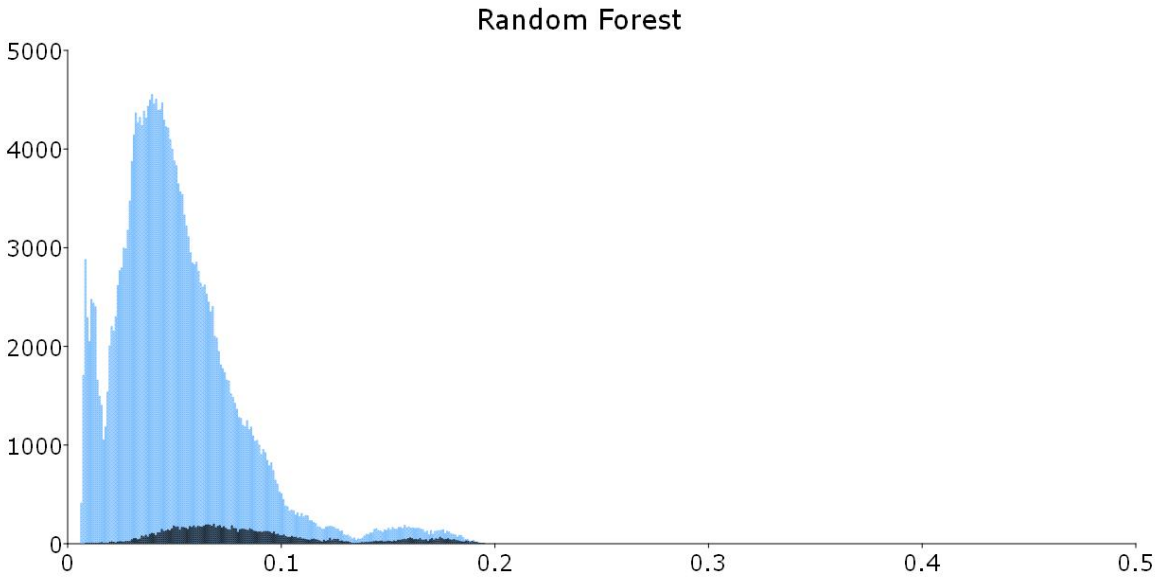
## B.4 Application

*Figure B.4.1: Propensity score by treatment status, Probit*



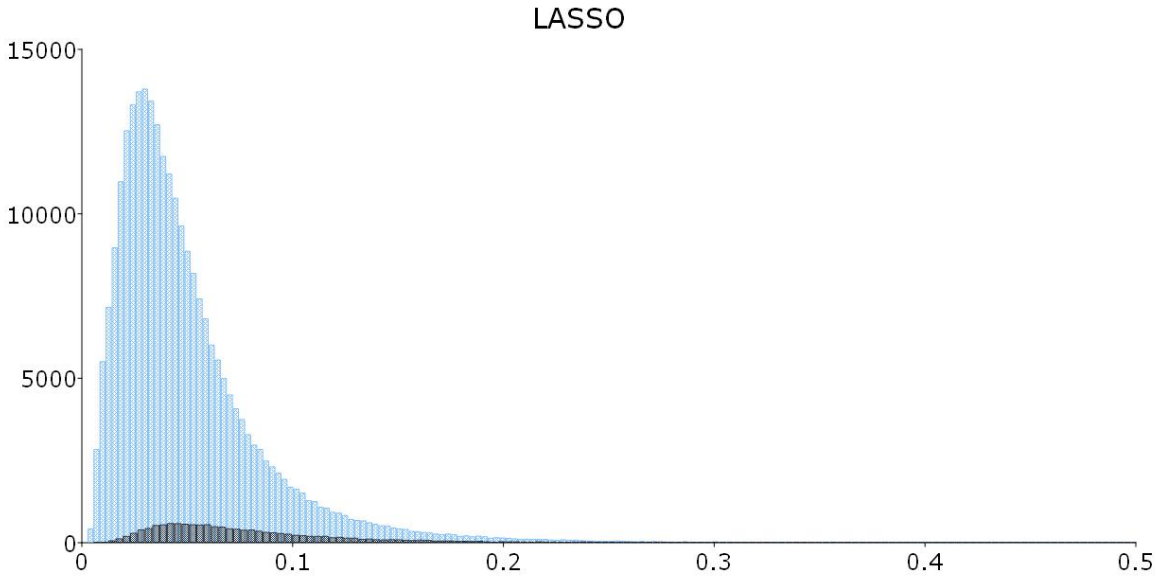

Notes: Histogram with PS on the horizontal axis estimated using the Probit. From the application in Section 6 with N=276'637 and about 5% treatment share. Control units are light, treated units dark shaded.

*Figure B.4.2: Propensity score by treatment status, Random Forest*




Notes: Histogram with PS on the horizontal axis estimated using the Random Forest. From the application in Section 6 with N=276'637 and about 5% treatment share. Control units are light, treated units dark shaded.

*Figure B.4.3: Propensity score by treatment status, LASSO*




Notes: Histogram with PS on the horizontal axis estimated using the LASSO. From the application in Section 6 with N=276'637 and about 5% treatment share. Control units are light, treated units dark shaded.