I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10532

Practical Procedures to Deal with Common Support Problems in Matching Estimation

Michael Lechner
Anthony Strittmatter

JANUARY 2017

IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 10532

Practical Procedures to Deal with Common Support Problems in Matching Estimation

Michael Lechner
SEW, University of St. Gallen, CEPR, PSI, CESifo, IAB and IZA

Anthony Strittmatter
SEW, University of St. Gallen

JANUARY 2017

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

Practical Procedures to Deal with Common Support Problems in Matching Estimation[1]

This paper assesses the performance of common estimators adjusting for differences in covariates, such as matching and regression, when faced with so-called common support problems. It also shows how different procedures suggested in the literature affect the properties of such estimators. Based on an Empirical Monte Carlo simulation design, a lack of common support is found to increase the root mean squared error (RMSE) of all investigated parametric and semiparametric estimators. Dropping observations that are off support usually improves their performance, although the magnitude of the improvement depends on the particular method used.

JEL Classification: C21, J68

Keywords: Empirical Monte Carlo Study, matching estimation, regression, common support, outlier, small sample performance

Corresponding author:
Michael Lechner
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen
Varnbüelstrasse 14
CH-9000 St. Gallen
Switzerland
E-mail: Michael.Lechner@unisg.ch

1 This project is part of the project "Regional Allocation Intensities, Effectiveness and Reform Effects of Training Vouchers in Active Labor Market Policies", IAB project number 1155. This is a joint project of the Institute for Employment Research (IAB) and the University of Freiburg. We gratefully acknowledge financial and material support by the IAB. We thank Lorenzo Camponovo, Bernd Fitzenberger, and Andreas Steinmayr for helpful comments on a previous draft of the paper. The usual disclaimer applies.

1 Introduction

It is a common task in applied econometrics to compare moments or distributions of random variables of two subsamples that are free of differences due to some observed variables. One example that received much attention in the recent applied literature is the evaluation of active labour market programmes (ALMP) based on very informative and large administrative data (see the meta study by Card, Kluve, and Weber, 2010, for a comprehensive summary of this literature). Usually, the main goal in this literature is to compare expected future reemployment and earnings ('outcomes') of participants in a programme ('treatment') with non-participants. In many cases, identification is based on a selection-on-observable identification strategy exploiting the rich data available (e.g., see Imbens, 2004, for an overview).

Essentially, all estimation procedures used in this context are based on predicting average outcomes in one treatment state (e.g., for non-participants who are called 'non-treated' from now on) based on the distribution of exogenous variables in the other treatment state (e.g., the programme participants who are called 'treated' from now on). Apparently, in this example, if the values of the characteristics observed for the treated are not observable among the non-treated, the mean outcome for the non-treated who would have such characteristics cannot be estimated without strong assumptions. These assumptions allow some extrapolation for such values of characteristics. This is the so-called *no common support problem*. A related (finite sample) problem appears when there are only a few observations in some relevant part of the covariate space in the particular sample at hand. This is called the *thin common support problem*. Such areas of no or thin support may increase biases and variances of estimators (e.g., Kahn and Tamer, 2010, Crump, Hotz, Imbens, and Mitnik, 2009). Surprisingly, and different from other aspects of treatment effect models, these issues have received little attention in the literature.

Here, we investigate the impact of these problems on commonly used estimators, and analyse procedures proposed to address it. Similar to Huber, Lechner, and Wunsch (2013), we do this in the context of a large-scale active labour market policy evaluation and use Empirical Monte Carlo methods. One of the contributions of this paper is to show that common support problems as well as the 'remedies' chosen matter for the results obtained (see also Busso, DiNardo, and McCrary, 2014a, Dehejia and Wahba, 1999, Smith and Todd, 2005). When there are support problems, they are addressed in different ways in applied studies. The most convenient way is to ignore the problem and take as given the way the specific estimators (as implemented in the software used) deal with it. If support problems are explicitly addressed, then one possibility is to change the population for which the effects are estimated to the one for which the distributions of characteristics overlap. This matters if the effects are heterogeneous. Alternatively, one may use a parametric model to predict mean outcomes in no-support regions. A third alternative is to give up point-identification and confine oneself to a set-identified parameter (e.g., Lechner, 2008).

We stick to the more popular point-identified case and analyse several practical adjustment procedures proposed in the literature. For example, Rosenbaum and Rubin (1983) suggest dropping treated observations for which there are no non-treated observations with the same covariate values. However, this procedure may lead to drastic reductions in sample size if the covariate space is large or if there are (almost) continuous covariates. Thus, most procedures used in practice, and analysed in this paper, focus directly on the propensity score, i.e. the conditional probability of treatment given the values of the covariates.

There are also suggestions to address simultaneously problems of common support as well as (too) important non-treated observations. These issues are related because in regions of thin support, the predictions are based on only few observations. This potentially leads to finite-sample bias and increased variance. Therefore, Imbens (2004) and Huber, Lechner, and Wunsch (2013) develop ways to address this issue, which appear to be effective in the simulation study conducted by the latter authors. Finally, Crump, Hotz, Imbens, and Mitnik (2009) suggest explicitly focussing the estimation on a subsample of the data with 'strong common support' to maximise the precision of the estimators.

Recently, Busso, DiNardo, and McCrary (2014a) investigated some of those procedures. They find that some of the approaches (i.e., those by Crump, Hotz, Imbens, and Mitnik, 2009, Dehejia and Wahba, 1999, Ho, Imai, King, and Stuart, 2007, and Smith and Todd, 2005, to be explained in detail below) have the potential to reduce the bias and partly increase the efficiency of the estimators. However, their simulation study is based on rather artificial distributional assumptions, which are unlikely to be observed in reality.

To address the issue of using a realistic design in a simulation study, recently, Lechner and Wunsch (2013), and Huber, Lechner, and Wunsch (2013) advocated using what they call an Empirical Monte Carlo Study (EMCS). The main idea is to use a large data set, which is similar to the data typically used in relevant applied work. In this large data set, considered to be the 'population' in the simulation exercise, a propensity score is estimated. Then, random samples are drawn from the subpopulation of the non-treated. For this group, the effect of treatment is known to be zero. Next, the estimated propensity score is used to assign a (pseudo-) treatment status to these non-treated. Such a procedure reflects the same selectivity observed in the population while insuring that the true effect is known (and zero) and does not require a priori specifying the joint distribution (or conditional moments) of outcomes, confounders, and treatment.[2]

Our results suggest that dropping observations that are off support improves the performance of many estimators, mainly by increasing their precision. For matching estimators, this im-

[2] There are alternatives in the literature that share the goal of making simulation studies more relevant but do not share this feature. For example, Abadie and Imbens (2011) and Busso, DiNardo, and McCrary (2014b) propose to apply more structural empirical simulation designs. The dependence structures between the control, treatment and outcome variables are estimated with real data. Afterwards, the treatment and outcome variables are simulated using the distribution of control variables from real data and the coefficients estimated in the first step. This approach has the advantage that the size of the treatment effect can be restricted, but it requires assumptions about the distribution of the 'error terms'.

provement can exceed 20 standard deviations in some specifications. Even for parametric estimators, the performance improvements may be non-trivial and worth pursuing. The procedures of Dehejia and Wahba (1999), Grzybowski et al. (2003), and Vincent et al. (2002) appear to improve the performance of different estimators in (almost) all specifications. We suggest dropping treated observations with propensity score values above a specific threshold. Our findings suggest specifying the cut-off value at the maximum or the 99%-quantile of the propensity score in the non-treated subpopulation. This procedure might be combined with an adjustment among the non-treated. Non-treated observations with a high importance (in the matching) should be dropped in the first place. Afterwards, treated observations with propensity score values above a threshold are dropped, with the threshold value being specified in the remaining non-treated sample (two-step procedure suggested by Huber, Lechner, and Wunsch, 2013).

The remainder of this study is organized as follows. In the next section, we introduce the econometrics framework. The underlying data and the simulation design are presented in Section 3. The results are presented in Section 4, and conclusions are drawn in Section 5. Furthermore, we provide a discussion about the detection of support problems and possibly remedies in Online Appendix A. Some details of the data, the simulation procedures, and supplementary results are provided in Online Appendices B-H.

2 The econometric model

2.1 Parameter of interest

Consider a setup with a binary treatment D, $d \in \{0,1\}$ (e.g., participation in a program) and an outcome variable Y (e.g., post-programme employment or earnings). X is a K-dimensional vector of covariates with support $\mathcal{X} \subset \mathbb{R}^K$.[3] The goal of the empirical analysis is to obtain mean comparisons of the outcome variable in the subsamples defined by D that are free of any differences due to the covariates X. If the covariate space is sufficiently rich, such a comparison will uncover parameters that have a causal interpretation, e.g., the average treatment effect on the treated (ATET, for details, see Imbens, 2004, or Rubin, 1974). More specifically, the focus is on the following estimand:

$$\gamma = E[Y|D = 1] - E[E(Y|X = x, D = 0)|D = 1].$$

In analogy to this definition, one can define similar parameters for other subpopulations as well. Here, for simplicity, the focus is only on γ, because (i) this parameter is of interest in most evaluation studies, and (ii) because support problems appearing for most other parameters can be addressed in a symmetric way.

Assume that an i.i.d. sample of size N is available containing measurements of y_i, d_i, and x_i. Thus, under usual regularity conditions, γ can be non-parametrically identified. However, when the dimension of X is high, the so-called curse of dimensionality makes reliable non-parametric estimation difficult to impossible. In this case, the balancing score property introduced by Rosenbaum and Rubin (1983) obtains practical relevance. It implies the following equality, which also holds in all subpopulations defined by X:

[3] We use the convention that capital letters denote random variables, while small letters denote particular values. If small letters are subscripted with i, such values are observed in a random sample.

$$E[E(Y|X=x,D=0)|D=1]=E[E(Y|p(X)=p(x),D=0)|D=1].$$

p(x) = P(D=1|X=x) denotes the *propensity score*. In most applications, the propensity score is approximated by a parametric model so that the resulting non-parametric estimation problem becomes essentially one-dimensional.

2.2 The common support assumption

Implicitly, the estimand defined above requires that for every unit with *d=1*, there should be a unit with the same (or a similar) value of *p(x)* among the group of units with *d=0*. Let $f(p(x)|D=d)$ be the density of the propensity score $p(x)$ conditional on the treatment status d. The density $f(p(x)|D=1)$ can be consistently estimated in the treated subsample under some regularity conditions. The same is true for $f(p(x)|D=0)$ in the non-treated subsample. There are 'support problems' (in the population) when $f(p(x)|D=0)=0$ and $f(p(x)|D=1)>0$. The set $\mathcal{W}_d=\{p(x)\in(0,1): f(p(x)|D=d)>0\}$ represents the support of $p(x)$ for d being zero or one.

Common support assumption (CS)

$\mathcal{W}_1 \subseteq \mathcal{W}_0$.

This assumption is automatically satisfied in the population when $p(x)<1$. However, even if this assumption holds in the population, the actual sample available may still be plagued by a lack of overlap. A comprehensive summary of different ways to detect support problems as well as a description of possible identification and estimation approaches in the presence of support problems is in Online Appendix A.

2.3 Balance of propensity score distributions

Even if the common support assumption is satisfied, an imbalance in the sample propensity score distributions could still lead to regions with only few (or none) treated or non-treated observations. Imbens (2004) suggests generating measures for the importance of each observation. A high importance indicates a high probability for a large imbalance in the conditional propensity score distributions with respect to the treatment status (for a given sample size). Following Huber, Lechner, and Wunsch (2013), we drop non-treated observations with a high importance weights, which we denote by $\widehat{\omega}_i$ in the following. For details about the implementation, please, see Online Appendix A.4.

2.4 Estimation

We assess the impact of support problems and their remedies on three different classes of estimators. In the following, we give a brief overview of the applied estimators. The details of the implementation can be found in Section C.4 of Online Appendix C.

The first class consists of parametric regressions. We use ordinary least squares (OLS) regressions for continuous and discrete non-binary outcome variables. For binary outcome variables, we specify parametric probit models. These estimators also 'work' without common support. However, since parametric models are only approximations of the true model, one may suspect that their out-of-support predictions might be particularly unreliable.

The second class of estimators considered is based on inverse probability weighting (IPW) using the propensity score. The estimated propensity score is obtained from a parametric probit model. Therefore, the common support assumption is, again, not required for this estimator either. However, as before, the results outside the common support depend on the correct specification of the parametric model for the propensity score.

The third class of estimators considered are propensity score matching estimators. In particular, we investigate the popular modified and bias adjusted propensity score radius-matching estimator suggested by Lechner, Miquel, and Wunsch (2011).

3 Empirical Monte Carlo study

3.1 Empirical Monte Carlo study

The analysis of the common support issues is based on an Empirical Monte Carlo Study (EMCS), as suggested by Huber, Lechner, and Wunsch (2013) and Lechner and Wunsch (2013). It combines the advantages of a ‘classical’ Monte Carlo study with the advantages of using real data. The idea is to use a large data set, which is similar to data in the respective field of study. To obtain appropriate random samples for the Monte Carlo study, the first step consists of estimating the propensity score in the full data. This estimated model will reflect the ‘true’ selectivity. Subsequently, random samples from the non-treated subpopulation are drawn. A (pseudo) treatment is assigned according to the ‘true’ selection model. Since there are only non-treated individuals in the samples used in the simulations, the true effect of the assigned treatment is known to be zero. The key advantages of this procedure are that the selectivity in the samples reflects the selectivity in the larger population, and that there is no need for stipulating a model on how the outcomes depend on treatment and covariates. Adding some more structure allows varying various components of the data generating process that are deemed relevant in the particular analysis. We vary the sample size (N = 500, 2,000, 8,000), the share of treated (10%, 50%, 90%), the outcome variable (earnings, months employed, employment), the degree of effect heterogeneity, the propensity score specification, and the severity of the support problem. The details of the simulation process are described in Online Appendix C. In the next section, we discuss the administrative data used for the simulation.

3.2 Empirical bases for the simulations: German administrative data

3.2.1 Data base

The German Federal Employment Agency generated the data for this study from their social security records. It contains information regarding individuals having received a training voucher in 2003 or 2004 and those who did not. Training vouchers certify the eligibility for public sponsored further training (see Doerr et al., 2014, for more details). Unemployed awarded with a voucher may redeem it at certified training providers.

The data contain extensive daily information on employment subject to social security contributions, receipt of transfer payments during unemployment, job search, and participation in different active labour market programs as well as rich individual information. It is a combination of two populations: A 3% random sample of those individuals in Germany who experienced at least one switch from employment to non-employment in 2003 and did not receive any voucher are merged with *all* voucher recipients. We account for the treatment-based sampling scheme by using sampling weights when necessary. This type of data has been frequently used to evaluate German active labour market policies (e.g., Biewen, Fitzenberger, Osikominu, and Paul, 2014, Lechner, Miquel, and Wunsch, 2011, Rinne, Uhlendorff, and Zao, 2013). It is comparable to many administrative data sets in Europe (see Lechner and Wunsch, 2013).

The treatment consists of receiving a voucher for further training during the first twelve months of unemployment. Training vouchers indicate the particular type of course for which they may be redeemed. We only consider vouchers awarded to obtain skills for manufacturing or service workers (in the following: *vouchers for manufacturing and service workers, VMSW*) and vouchers to obtain skills for technicians (in the following: *vouchers for technicians, VTEC*). These are interesting because their selection processes show considerable heterogeneity: The *award of VTEC* is far more selective than the *award of VMSW*.

3.2.2 Control variables and common support

The choice of control variables follows Lechner and Wunsch (2013). We consider all variables identified as important confounders in their study, i.e., baseline personal characteristics, timing of program start, regional dummies, benefit and unemployment insurance claims, pre-program outcomes, and the short-term labour market history. Measurements regarding individual characteristics refer to the time of inflow into unemployment. Further information about the control variables and the support of the propensity score in the different samples can be found in Online Appendix B. The full list of variables used for the propensity scores (including interaction terms) is given in Tables B.1 and B.2.

Individuals with high educational or occupational degrees who work as technicians, professionals, or managers are very likely to obtain a *VTEC* (but not necessarily a *VMSW*). For simulation and modelling purposes, it is useful to relate the support problems to particular variables. Thus, three interactions between occupation and educational/vocational degree are included in the propensity score probit model: (i) being a technician or having an associate profession interacted with a higher secondary schooling degree (S_1); (ii) being a professional or manager interacted with a higher secondary schooling degree (S_2); and (iii) being a professional or manager interacted with a university or college degree (S_3).

The three interaction terms, which are binary, are collected in the 'support variable' *S*:

$$S_i = \max(S_{1i}, S_{2i}, S_{3i}).$$

The support variable S_i equals one for 66% of all individuals who are awarded with *VTEC*, but only for 21% of all individuals who are awarded with *VMSW* and for 20% of those not receiving a voucher at all.

4 Results

4.1 General remarks

There are results for 252 different DGPs, 46 procedures to deal with support problems, three outcomes, three estimators, and up to three different model specifications.[4] It is not possible to report (and understand!) all of these results without reducing their dimensionality. Therefore, linear regressions provide summary measures. In these regressions, the dependent variables consist of measurements of the quality of the estimators, such as the root mean squared error (RMSE), the absolute bias, and the standard errors. The independent variables reflect different features of the DGPs (treatment shares, types of effect heterogeneity, sample size, and type of voucher), of the model specifications, of the estimators, and of the various rules to tackle the support problem (see Table 1 for the list of all procedures used; a more comprehensive description is provided in Online Appendix A).

Controls for model specifications, treatment shares, and types of effect heterogeneity are interacted with each other (but not with the dummies for the different procedures to handle support problems). Furthermore, separate regressions are reported for different types of vouchers, estimators, sample sizes, and types of support reductions as, a priori, considerable heterogeneity is expected with respect to those features.

---

[4] The DGPs vary by sample size, treatment share, type of treatment (*VMSW* or *VTEC*), type of effect heterogeneity, and type of support restrictions. Overall, there are 90,720 different results.

Table 1: Procedures for handling support problems

Procedure	Description	Rule	Assumption	References
	Drop treated based on fixed value of propensity score			
A	Drop no observations			
WA	Drop non-treated if weights high ($\hat{\omega} \geq 0.04$)		D	Imbens (2004)
B1	Drop if $.1 < p(x) < .9$			CHIM (2009)
B2	Drop if $p(x) < .9$			
WBx	Bx & WA			
	Drop treated based on density of propensity score			
C1	Drop treated if $f(p(x)) < q_2$	1	D	HIST (1998), HIT (1997), Smith, Todd (2005)
C2	Drop treated if $f(p(x)) < q_{10}$	1	D	
C3	Drop treated if $f(p(x)) < q_\tau$ with $\tau = (100/N)*100\%$			
WCx	Cx & WA	1	D	
	Drop treated based on lack of non-treated neighbours in terms of radius of propensity score			
D1	Drop treated if no close match in 0.01 radius (u=0.01)	2	D	Grzybowski et al. (2003), Vincent et al. (2002)
D2	Drop treated if no close match in 0.1 radius (u=0.1)	2	D	
WDx	Dx & WA	2	D	
	Drop treated based on upper limits of distribution of propensity score among non-treated			
E1	Drop treated above highest non-treated p-score	3	D	Dehejia, Wahba (1999)
E2	Drop treated above 99% highest non-treated p-score	3	D	
E3	Drop treated above 95% highest non-treated p-score	3	D	HLW (2013)
WEx	Ex & WA	3	D	
	Procedures that extrapolate into lack of support region			
F1	E1 & estimate $Y_{1\|N}$ - $Y_{0\|N}$ at highest non-treated p-score	3	E	
F2	E2 & estimate $Y_{1\|N}$ - $Y_{0\|N}$ at 99% highest non-treated p-score	3	E	
F3	E3 & estimate $Y_{1\|N}$ - $Y_{0\|N}$ at 95%highest non-treated p-score	3	E	
WFx	Fx & WA	3	E	
G1	E1 + lin. approx. of $Y_{1\|N}$ - $Y_{0\|N}$ above highest non-treated p-score	3	F	
G2	E2 + lin. approx. of $Y_{1\|N}$ - $Y_{0\|N}$ above 99% highest non-treated p-s.	3	F	
G3	E3 + lin. approx. of $Y_{1\|N}$ - $Y_{0\|N}$ above 95%highest non-treated p-s.	3	F	
WGx	Gx & WA	3	F	
H1	Estimate Y0|N at highest non-treated p-score	3	B	
H2	Estimate Y0|N at 99% highest non-treated p-score	3	B	
H3	Estimate Y0|N at 95% highest non-treated p-score	3	B	
WHx	Hx & WA	3	B	
I1	Lin. approx. of $Y_{0\|N}$ above the highest non-treated p-score	3	C	
I2	Lin. approx. of $Y_{0\|N}$ above the 99% highest non-treated p-score	3	C	
I3	Lin. approx. of $Y_{0\|N}$ above the 95%highest non-treated p-score	3	C	
WIx	Ix & WA	3	C	

Note: WA indicates that non-treated observations with $\hat{\omega} \geq 0.04$ are dropped (see Section 2.3 and Online Appendix A.4). This procedure is used in combination with other procedures (two-step procedure). CHIM: Crump, Hotz, Imbens, Mitnik; HIST: Heckman, Ichimura, Smith, Todd; HIT: Heckman, Ichimura, Todd; HLW: Huber, Lechner, Wunsch. A complete description of all support procedures can be found in Online Appendix D.

In the regressions, three different outcome variables (earnings, months employed, employment) are pooled. Because they are measured on different scales, they are normalized by their standard deviation in the baseline specification (see Online Appendix D for details). Therefore, for a given estimation procedure, the regression coefficients of the binary control variables indicate by how many standard deviations the performance measure changes if this dummy turns on in comparison with the omitted category. Additionally, Tables H.1-H.12 in Online Appendix H report heterogeneous results by the type of outcome.

The next section reports regression results for the case without additional support restrictions. It is followed by the case with restricted support. Detailed results are reported in Online Appendix G.

4.2 Regression results for DGPs not restricting common support

The DGPs for which the treatment is based on the *award of VMSW* are expected not to be subject to severe support problems because of their low degree of selectivity (see Tables B.1 and B.2 in Online Appendix B). Although the *award of VTEC* is more selective, even in this case, the population propensity scores exceed the level of 0.9 only when the treatment share is 90%. Even then, it remains below one (see Figures B.1 and B.2 in Online Appendix B) so that the DGPs considered in this section show no asymptotic support problems. Nevertheless, issues of thin support may still be relevant. Tables 2 and 3 report the results from the regressions of the normalized RMSE on different procedures to handle support problems (the reference category is ignoring the problem, i.e., Procedure A). Negative coefficients indicate improvements, and positive ones indicate impairments in RMSE. We report the results for selected procedures only. The results for the complete set of common support procedures (including the standard errors of the estimated coefficients and R^2s), as well as for other performance measures (normalized absolute bias, standard deviation) can be found in Online Appendix E (Tables E.1 to E.6).

Table 2: Effect of support-adjustment procedures on normalized RMSE in DGPs without support restrictions for the treatment 'award of VMSW'

Sample size		500			2,000			8,000	
Support	Param.	IPW	Match.	Param.	IPW	Match.	Param.	IPW	Match.
procedures	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				No adjustment (A: reference)					
WA	.002	-.012	.003	.005	-.017	.004	0	0	0
				Drop treated based on fixed value of propensity score					
B2	-.007	-.015*	-.007	.220***	.187***	.188***	.701***	.640***	.611***
				Drop treated based on density of propensity score					
C2	.017	.019**	.011	.001	0	.002	0	0	0
WC2	.018	.007	.013	.007	-.017	.005	0	0	0
C3	.441	-.020*	.486	.209	-.021	-.021***	.104	-.015	.176
				Drop treated based on lack of non-treated neighbours in terms of radius of propensity score					
D1	-.016	.007	-.025	-.001	0	-.001	0	0	-.001
WD1	-.016	-.006	-.025	.001	-.017	0	0	0	-.001
D2	-.001	0	-.002	0	0	0	0	0	0
WD2	0	-.012	.001	.005	-.017	.003	0	0	0
				Drop treated based on upper limits of distribution of propensity score among non-treated					
E1	-.013	.003	-.019***	-.013	.001	-.013**	-.003	-.001	-.004
WE1	-.013	-.009	-. 020	-.015	-.018	-.016***	-.003	-.001	-.004
E2	-.010	-.025***	-.018	-.006	-.021	-.004	.028	-.001	.019
WE2	-.009	-.025***	-.018	-.008	-.026*	-.005	.028	-.001	.019
E3	.037	-.004	.017	.070***	.027*	.058***	.242***	.182***	.186***
WE3	.038	-.004	.018	.072***	.028*	.059***	.242***	.182***	.186***
				Procedures that extrapolate into lack of support region					
F2	.633	.031***	.249	.047	.259***	.006	.078	.255*	.157
WF2	.770*	.034***	.251	.050	.221*	.011**	.078	.255*	.157
G2	.042	.013*	-.002	.032	.014	.017***	.020	.004	.013
WG2	.046	.015**	-.001	.036	.013	.019***	.020	.004	.013
H2	1.039*	.023***	-.033	-.001	.448**	.004	-.062**	.234	-.052**
WH2	.792**	.026***	0	.004	.544***	.009*	-.062**	.234	.052**
I2	.037	.005	-.009	.022	-.003	.001	.022	-.003	.010
WI2	.042	.007	-.007	.026	-.004	.001	.022	-.003	.010
# of obs.	828	1,242	1,242	2,484	3,726	3,726	2,484	3,726	3,726

Note: The results are obtained from OLS regressions. The dependent variable is the normalized RMSE. The covariates contain a full set of dummy variables for the different procedures to handle support problems. The reference category is to drop no observations (Procedure A, a full description of the different procedures is given in Table 1). Further control variables are the tuning parameters of the different DGPs in a fully interacted way (see description in main text). Only selected coefficients for the different procedures are reported in this table. A complete set of results, including standard errors and R^2s, is shown in Online Appendix E. ***, **, and * indicate significance at the 1-, 5-, and 10-percent levels, respectively (based on robust standard errors). There are fewer observations for the small *N*, because only the case of 50% treatment shares is considered. The same holds for the parametric estimations, because only 2 specifications are considered (instead of 3, see Online Appendix C).

Table 3: Effect of support-adjustment procedures on normalized RMSE in DGPs without support reduction for the treatment 'award of VTEC'

Sample size	500			2,000			8,000		
Support procedures	Param. (10)	IPW (11)	Match. (12)	Param. (13)	IPW (14)	Match. (15)	Parametr. (16)	IPW (17)	Match. (18)
No adjustment (A: reference)									
WA	.014	-.097***	.037***	.018*	-.092***	.018***	.001	-.037	.003
Drop treated based on fixed value of propensity score									
B2	-.038*	-.112***	-.069***	.163***	.036	.118***	.576***	.374***	.461***
Drop treated based on density of propensity score									
C2	-.007	.014	-.036***	.002	0	.003	0	0	0
WC2	.006	-.089***	-.006	.020*	-.091***	.021***	.001	-.037	.003
C3	-.060**	-.035**	-.087***	-.029**	-.028	-.054***	-.097***	.207	-.012
Drop treated based on lack of non-treated neighbours in terms of radius of propensity score									
D1	-.047**	.017	-.113***	-.011	.001	-.020***	-.001	0	-.004
WD1	-.045**	-.095***	-.112***	-.023**	-.102***	-.029***	0	-.037	-.001
D2	-.002	0	-.004	-.001	0	0	0	0	0
WD2	.006	-.099***	.021**	.017	-.092***	.018***	.001	-.037	.003
Drop treated based on upper limits of distribution of propensity score among non-treated									
E1	-.039*	.004	-.082***	-.039***	0	-.048***	-.016	-.003	-.023
WE1	-.046**	-.101***	-.096***	-.039***	-.106***	-.054***	-.018	-.043*	-.026
E2	-.039*	-.123***	-.098***	-.019	-.097***	-.049***	.073**	-.040*	.020
WE2	-.037	-.132***	-.098***	-.014	-.121***	-.050***	.075**	-.040*	.021
E3	.032	-.106***	-.059***	.132***	-.013	.053***	.489***	.312***	.363***
WE3	.040	-.101***	-.054***	.140***	-.007	.059***	.490***	.313***	.364***
Procedures that extrapolate into lack of support region									
F2	.094***	.264	-.013*	.048***	.156	.005	-.019	.072	.013
WF2	.305*	.205	.001	.065***	.141	.030***	-.017	.072	.015
G2	.091***	-.010	-.026***	.077***	-.012	.008	.085***	-.012	.022
WG2	.125***	.002	-.013	.100***	-.019	.021***	.086***	-.012	.023
H2	.086***	-.006	-.026***	.041***	-.028	-.003	.020	-.049**	-.023
WH2	.125***	.007	.-.010	.059***	-.037**	.022***	.022	-.048**	-.021
I2	.048***	-.059***	-.066***	.044***	-.055***	-.027***	.067**	-.052*	-.009
WI2	.067***	-.057***	-.061***	.061***	-.065***	-.019***	.068**	-.051*	-.008
# of obs.	828	1,242	1,242	2,484	3,726	3,726	2,484	3,726	3,726

Note: See note below Table 2.

Both tables show that for most procedures, the case with a reduction of the normalized RMSE (in comparison with the omitted category of dropping no observations) most likely occur in the smallest samples. Interestingly, it appears that almost all procedures either have no effect on the normalized absolute bias of the estimators (see Tables E.3 and E.5 in Online Appendix E), or increase it (somewhat). However, such increases are usually (over-) compensated for by a reduction in the variability of the estimators (see Tables E.4 and E.6 in Online Appendix E). Accordingly, the performance of the estimators can be improved, even in DGPs where the support conditions are not violated in the population. This is in line with the arguments of Busso, DiNardo, and McCrary (2014a), Crump, Hotz, Imbens, and Mitnik (2009), and Kahn and Tamer (2010) suggesting that thin support issues lead to a loss of precision. Finally, note that for the treatment *award of VTEC*, with a strong selection into treatment, the potential performance improvements are considerably larger than for the treatment *award of VMSW*. Next, the performance of the single procedures is discussed in more detail.

Dropping observations with high importance (*WA*) improves only the performance of IPW estimators. However, when this approach is combined with other procedures, then these (joint) procedures (beginning with *W*) have the potential to improve the performance of the single procedure with which it is combined. This is consistent with the findings of Huber, Lechner, and Wunsch (2013).

The procedure of Crump, Hotz, Imbens, and Mitnik (2009) is applied in two different specifications. First, we drop all observations with a propensity score below 0.1 and above 0.9 (*B1*; see Online Appendix E). Second, only observations with a propensity score above 0.9 (*B2*) are dropped. Although *B1* and *B2* seem to work for the small samples, for the larger sample, they lead to biases that are large enough to dominate the RMSE.

The results for the procedure dropping treated observations with a low marginal density (*C*) are not encouraging either. While there appears to be some possibility of improvements on the performance of estimators if the selectivity is strong enough (Table 3), in the case of weak selectivity (Table 2) the small sample performance deteriorates. For procedure C3, which is adaptive to the sample size, we find improvements in RMSE particularly for smaller sample sizes. Reduction in the standard deviations and not improvements in the biases drive this result.

Procedure *D* drops treated observations for which the distance to the closest one-to-one match is below u, i.e., all treated observations with a propensity score difference to the closest control observation above u are dropped, with $u \in \{0.01; 0.001\}$ (see Grzybowski et al., 2003, Vincent et al., 2002). Procedure *E* drops treated observations with a propensity score above a cut-off value $\bar{p}$ (Dehejia and Wahba, 1999). Generally, D and E improve the performances of the estimators equally well. Unlike most other procedures, only in very rare cases do these procedures hurt the performances of the estimators in terms of normalized RMSE. The largest improvement (originating from the standard deviations) can most often be obtained from *E*. This procedure works better than *D*, particularly in larger samples, when $\bar{p}$ is specified as being either the maximum or the value of the 99%-quantile of the propensity score in the non-treated sample. In the smallest sample, Procedure *D* appears to have slightly better properties than *E*. In most cases, D works best with $u = 0.01$ for parametric and matching estimators. Using D with $u = 0.01$ or $u = 0.1$ performs about equally well for IPW estimators. Both procedures improve somewhat when combined with *W*.

All procedures that aim to estimate the conditional treatment effect $\gamma_N = Y_{1|N} - Y_{0|N}$ or $Y_{0|N}$ off support (*F*, *G*, *H*, *I*) do not perform well. In most cases, there are no performance improvements. However, if any improvements show up, then they are, with rare exceptions, smaller than for the other procedures discussed above. A noticeable exception is support procedure *H*, which seems to be one of the few procedures that may improve the RMSE even in the absence of support problems.

4.3 Regression results for DGPs with reductions of common support

Next, DGPs are considered for which the support with respect to S is restricted in a way that causes (serious) support problems. For this purpose, we restrict $P(D = 1|S = 1)$. In particular, we replace $P(D = 1|S = 1)$ by $P(D = 1|S = 1)' = 1 - \phi$, where the prime indicates the restricted conditional treatment probability. If $\phi = 0$, then even asymptotically, there is no common support. When $0 < \phi < 1$, there is common support in the population, but it may be thin when ϕ is small. Remember, the share of observations with $S = 1$ is much larger for individuals awarded with *VTEC* than with *VMSW*.[5] Accordingly, the treatment *award of VTEC* is considerably more affected by the support reductions than the treatment *award of VMSW*.

Figure F.1 in Online Appendix F shows the performance of parametric, IPW, and matching estimators when no observations are dropped (*A*). In this figure, $P(D = 1|S = 1)'$ is gradually increased based on the grid $\{0.9, 0.91, \ldots, 1\}$. To be precise, the figures show the coefficients of the indicator variables for the different points of the grid based on the same regressions reported before but pooled for the two treatments. We find an increase in the average normalized RMSE between 0.1 and 1 standard deviation if $P(D = 1|S = 1)' = 0.9$ in comparison with the specifications with no support restrictions. However, if $P(D = 1|S = 1)' = 1$, the average normalized RMSE can be increased by up to eight standard deviations. Support reductions have the strongest impact on the matching estimator. Interestingly, matching estimators appear to have on average lower normalized absolute biases than parametric and IPW estimators. This is in line with the findings of Busso, DiNardo, and McCrary (2014b), who report that the biases of matching estimators are less affected by overlap problems than the bias of IPW estimators. The normalized absolute bias of matching estimators exceeds those of parametric and IPW estimators, only for very strong support reductions. However, the average normalized standard deviation of matching estimators becomes very large under strong support restrictions. This is the main reason for the bad performance of matching in the specifications with strong support reductions. For parametric and IPW estimators the performance in terms of average normalized absolute bias and average normalized standard deviation is more balanced in this situation. However,

---

[5] From Section 3.2, we obtain $P(S = 1)' = \delta P(D = 1)$ with $\delta = 0.21$ for VMSW and $\delta = 0.66$ for VTEC.

the average normalized RMSE, average normalized absolute bias, and average normalized standard deviation of these estimators may increase by up to one standard deviation when support reductions are substantial. The performance of these estimators is almost linearly decreasing when the support is reduced.

Tables 5 and 6 report the results from regression for the normalized RMSE for the case of such support reductions. For the sake of computation time, only specifications with $\phi = 0$ and $\phi = 0.01$ are included because these two scenarios lead to the most serious support problems. Thus, the specification of the OLS regressions is similar to the one used in the previous section. The only difference is that an additional dummy variable for the case $P(D = 1|S = 1)' = 0.99$ is included. As before, Tables 4 and 5 report only a subset of the results. The complete set of results can be found in Online Appendix F (Tables F.1 to F.6).

Not surprisingly, the performance improvements for the different estimators are more pronounced than for the DGPs without support restrictions. Furthermore, parametric, IPW, and matching estimators have very different properties for the DGPs with restricted support (see Tables 4 and 5 as well as Tables F.1 and F.2 in Online Appendix F). The greatest improvements occur for matching estimators for which all procedures handling support problems work well. However, note that various procedures improve the performance of parametric and IPW estimators as well: their normalized RMSE improves by up to 0.7 standard deviations.

Noticeably, the performance of the matching estimator improves when the sample size increases, while the performance of the parametric and IPW estimator is much less affected by the sample size. This finding suggests that the matching estimator can address support problems appropriately when the sample size is increasing and thin support regions are filled. Under the strong support restrictions we simulate, the performance improvements of parametric and IPW estimators does not increase with sample size.

As before, Procedure *WA* used alone only improves the performance of IPW estimators, but when combined with other procedures, such two-step procedures may improve the performance of all estimators.

Procedure *B* strongly reduces the normalized standard deviation of all estimators (the parametric estimators in the largest sample are an exception; see Tables F.5 and F.6 in Online Appendix F). However, as the normalized absolute bias increases (see Tables F.3 and F.4 in Online Appendix F), the overall effect on the normalized RMSE is ambiguous. It is only for the matching estimators that this procedure always leads to such performance improvements.

In contrast, Procedure *C* has again only little impact on the performance of the different estimators. With few exceptions, *C* neither harms nor improves their performance. Additionally, the data-adaptive procedure C3, which adjusts to the sample size, does (with few exceptions) not affect the performance of the estimators.

Procedures *D* and *E* have the largest positive effect on the performance of the estimators when there are strong support problems. They improve the normalized standard deviation of all estimators, particularly in the smaller samples. These improvements are largest for matching estimators. For the *award of VTEC* they can exceed 20 standard deviations. *D* and *E* also reduce the normalized RMSE and normalized absolute bias, particularly when the treatment is *award of VTEC*. In some specifications, the normalised RMSE can be improved by up to 20 and the absolute bias by up to 5 standard deviations. In rare cases when the treatment is *award of VMSW*, *E* increases the normalized absolute bias of the parametric and IPW estimators,[6] while *D* does not affect them. However, *E* leads to larger improvements in the normalized standard

[6] This appears only very rarely when the threshold $\overline{p}$ is specified at the highest propensity score value of the non-treated subpopulation.

deviations than does *D* in these specifications. Both procedures work better when combined with *W*.

Table 5: Effect of support-adjustment procedures on normalized RMSE in DGPs with support restrictions for the treatment 'award of VMSW'

Sample size		500			2,000			8,000	
Support procedures	Param. (1)	IPW (2)	Match. (3)	Param. (4)	IPW (5)	Match. (6)	Param. (7)	IPW (8)	Match. (9)
				No adjustment (A: reference)					
WA	.019	-.096***	.070	.017	-.081***	.155	0	-.009	4.590
				Drop treated based on fixed value of propensity score					
B2	-.110***	-.109***	-1.624***	.112***	.020	-3.709***	.563***	.394***	-7.292***
				Drop treated based on density of propensity score					
C2	-.091***	.001	.002	.003	.001	.005	0	0	0
WC2	-.077***	-.102***	.049	.021	-.080***	.160	0	-.009	4.590
C3	-.106***	-.024***	-.403	-.006	-.013	-.016	-.021	-.026	6.114
				Drop treated based on lack of non-treated neighbours in terms of radius of propensity score					
D1	-.124***	.009	-1.641***	-.023	-.014	-3.768***	.015	-.051*	-7.524***
WD1	-.124***	-.095***	-1.648***	-.050***	-.110***	-3.903***	.015	-.061**	-7.675***
D2	-.012	-.003	-1.493***	0	-.001	-3.516***	.002	-.009	-7.112***
WD2	-.016	-.103***	-1.501***	.018	-.082***	-3.611***	.003	-.018	-7.184***
				Drop treated based on upper limits of distribution of propensity score among non-treated					
E1	-.109***	.001	-1.625***	-.067***	-.017	-3.875***	.009	-.071**	-7.764***
WE1	-.121***	-.100***	-1.641***	-.075***	-.114***	-3.919***	.009	-.084***	-7.901***
E2	-.116***	-.113***	-1.635***	-.040	-.102***	-3.910***	.177***	-.020	-7.910***
WE2	-.114***	-.127***	-1.636***	-.0421	-.115***	-3.913***	.177***	-.021	-7.910***
E3	-.064***	-.119***	-1.593***	.043	-.052	-3.840***	.376***	.201***	-7.730***
WE3	-.060**	-.115***	-1.591***	.047	-.050	-3.838***	.376***	.201***	-7.729***
				Procedures that extrapolate into lack of support region					
F2	.266***	.267***	-1.281***	.180***	.270***	-3.539***	.221***	.202***	-7.664***
WF2	.248***	.277***	-1.273***	.190***	.266***	-3.543***	.222***	.203***	-7.671***
G2	.149***	.142***	-1.404***	.106***	.063**	-3.732***	.110***	.039	-7.856***
WG2	.199***	.177***	-1.388***	.121***	.069**	-3.723***	.110***	.038	-7.851***
H2	.131***	.188***	-1.387***	.083***	.079***	-3.646***	.118***	.051	-7.681***
WH2	.150***	.187***	-1.385***	.094***	.075***	-3.649***	.119***	.052*	-7.697***
I2	.041***	.042***	-1.571***	.033*	-.008	-3.844***	.051	-.015	-7.956***
WI2	.059***	.047***	-1.568***	.043**	-.008	-3.842***	.051	-.015	-7.953***
# of obs.	792	792	1,584	2,376	2,376	4,752	2,376	2,376	4,752

Note: The results are obtained from OLS regressions. The dependent variable is the normalized RMSE. The covariates contain a full set of dummies for the different procedures handling support problems. The reference category is to drop no observations (Procedure A). Further covariates are the tuning parameters of the different DGPs fully interacted (see description in Online Appendix B). Only selected coefficients for the different procedures are reported in this table. A complete set of results, including standard errors and R^2s, are found in Online Appendix F. ***, **, and * indicate significance at the 1-, 5-, and 10-percent levels, respectively (based on robust standard errors).

Table 6: Effect of support-adjustment procedures on normalized RMSE in DGPs with support restrictions for the treatment 'award of VTEC'

Sample Size		500			2,000			8,000	
Support procedures	Param. (10)	IPW (11)	Match. (12)	Param. (13)	IPW (14)	Match. (15)	Param. (16)	IPW (17)	Match. (18)
				No adjustment (A: reference)					
WA	.257***	-.654***	3.642**	.130	-.393*	.797	.0291	-.173	7.955
J1	.014	-.112	.494	-.041	-.092	2.268	.006	.004	-1.211
				Drop treated based on fixed value of propensity score					
B2	-.702***	-.629***	-6.017***	-.538**	-.649**	-22.42***	-.501**	-.770***	-19.57***
				Drop treated based on density of propensity score					
C2	-.048	-.013	-.031	.011	.001	.018	0	0	0
WC2	.207***	-.675***	3.663**	.143	-.391*	.819	.029	-.173	7.955
C3	.014	-.112	.494	-.041	-.092	2.268	.006	.004	-1.211
				Drop treated based on lack of non-treated neighbours in terms of radius of propensity score					
D1	-.594***	-.005	-5.327***	-.124	-.008	-22.80***	-.182	-.186	-19.93***
WD1	-.733***	-.650***	-6.113***	-.396*	-.495***	-23.32***	-.157	-.368**	-20.23***
D2	-.038	-.007	-4.112***	-.003	.030	-22.01***	-.036	-.027	-18.51***
WD2	-.527***	-.708***	-5.721***	.127	-.362*	-21.68***	-.007	-.200	-18.15***
				Drop treated based on upper limits of distribution of propensity score among non-treated					
E1	-.576***	-.018	-5.554***	-.336*	-.038	-23.10***	-.334**	-.279	-20.28***
WE1	-.746***	-.670***	-6.117***	-.466***	-.505***	-23.45***	-.379**	-.552***	-20.75***
E2	-.712***	-.581***	-6.069***	-.447**	-.458**	-23.61***	-.507**	-.626***	-21.86***
WE2	-.739***	-.701***	-6.109***	-.495***	-.537***	-23.67***	-.518***	-.636***	-21.88***
E3	-.683***	-.704***	-6.042***	-.445***	-.533***	-23.63***	-.531***	-.661***	-21.90***
WE3	-.654***	-.689***	-6.026***	-.442***	-.534***	-23.63***	-.531***	-.661***	-21.90***
				Procedures that extrapolate into lack of support region					
F2	1.450***	1.261***	-3.332***	3.017***	2.740***	-19.27***	2.474***	2.169***	-19.18***
WF2	1.861***	1.540***	-3.388***	2.997***	2.682***	-19.29***	2.471***	2.161***	-19. 17***
G2	1.499***	1.275***	-3.888***	2.043***	1.800***	-21.49***	1.167***	.898***	-20.51***
WG2	3.515***	3.089***	-2.975***	2.226***	1.947***	-21.34***	1.201***	.927***	-20.48***
H2	1.017***	.895***	-4.431***	1.351***	1.141***	-20.78***	2.034***	1.736***	-19.69***
WH2	1.005***	.803***	-4.533***	1.334***	1.092***	-20.79***	2.034***	1.732***	-19.69***
I2	.144***	.072	-5.494***	.188	.018	-23.21***	.279	.015	-21.33***
WI2	.446***	.338***	-5.353***	.263	.075	-23.13***	.304*	.037	-21.31***
# of obs.	792	792	1,584	2,376	2,376	4,752	2,376	2,376	4,752

Note: See note below Table 5.

All procedures aiming to estimate $\gamma_{ATET|N}$ or $Y_{0|N}$ off support (*F*, *G*, *H*, *I*) work only for matching estimators, if they work at all. However, even for matching estimators they do not outperform procedures *D* and *E*, which are much easier to implement.

In Online Appendix G, we provide a detailed description of further results.

5 Conclusions

This paper studies the performance of different parametric and semiparametric estimators adjusting observable characteristics in no- or thin-support situations and the performance of remedies suggested in the literature. An Empirical Monte Carlo Study is used to obtain performance measures using simulation designs based on particular real data. Therefore, the many different data generating processes investigated should be close to what is encountered in empirical applications in the context of program evaluation.

Our findings suggest that almost all procedures proposed in the literature to mitigate support problems have the potential to improve the performance of the estimators investigated, in particular when support problems become severe. Although the largest improvements can be achieved for matching estimators, parametric estimators benefit as well. However, not surprisingly, some procedures are more effective than others are.

The results obtained for support problems of different severity, different estimators, specifications, and other features of the data generating process suggest that procedures based on trimming observations in the treated group with propensity scores larger than the maximum value (or the 99% quantile) of the propensity score in the non-treated group consistently outperform the other procedures considered. Furthermore, when these procedures are combined with further trimming non-treated observations that receive a 'too-large' weight, additional improvements were observed. When support problems are mild, the gains usually come from an improvement in the precision of the estimators. When support problems are strong, biases are generally reduced and precision increases. Therefore, we recommend using these methods in applied work independent of the type of estimator used.

Clearly, due to the empirical design of the Monte Carlo approach used, the results of this study should be valid for similar programme evaluation studies based on large administrative databases. Furthermore, many features of the data generating processes have also been varied. Thus, we are tempted to claim that our simulation designs contain many different cases relevant in practise and have external validity beyond such programme evaluation studies. Whether this claim can be confirmed in another Empirical Monte Carlo study based on a different applied field remains speculative and deserves further research. For a general argument on how to address support problems, one has to provide technical comparisons among different estimators in finite or large samples. So far, only Crump et al. (2009), who aimed to find the most precise estimator, have provided such an investigation.

References

Abadie, A., and G. W. Imbens (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics*, 29 (1), 1–11.

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

Busso, M., J. DiNardo, and J. McCrary (2014a): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *Journal of Business and Economic Statistics*, forthcoming.

Busso, M., J. DiNardo, and J. McCrary (2014b): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, 96 (5), 885-896.

Card, D., J. Kluve, and A. Weber (2010): "Active Labour Market Policy Evaluations: A Meta-Analysis," *The Economic Journal*, 120 (548), 452–477.

Crump, R. K., J. V. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96 (1), 187–199.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94 (448), 1053–1062.

Doerr, A., B. Fitzenberger, T. Kruppe, M. Paul, and A. Strittmatter (2014): "Employment and Earnings Effects of Awarding Training Vouchers," *Industrial and Labor Relations Review*, forthcoming.

Grzybowski, M., E. A. Clements, L. Parsons, R. Welch, A. T. Tintinalli, M. A. Ross, and R. J. Zalenski (2003): "Mortality Benefit of Immediate Revascularization of Acute ST-Segment Elevation Myocardial Infarction in Patients with Contraindications to Thrombolytic Therapy: A Propensity Analysis," *The Journal of the American Medical Association*, 290 (14), 1891–1898.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66 (5), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64 (4), 605–654.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71 (4), 1161– 1189.

Ho, D., K. Imai, G. King, and E. Stuart (2007): "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15 (3), 199–236.

Huber, M., M. Lechner, and C. Wunsch (2013): "The Performance of Estimators Based on the Propensity Score," *Journal of Econometrics*, 175 (1), 1–21.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86 (1), 4–29.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions and Inverse Weight Estimation," *Econometrica*, 78 (6), 2021-2042.

Lechner, M. (2008): "A Note on the Common Support Problem in Applied Evaluation Studies," *Annales d'Economie et de Statistique*, 91-92, 217–234.

Lechner, M., R. Miquel, and C. Wunsch (2011): "Long-run Effects of Public Sector Sponsored Training," *Journal of the European Economic Association*, 9 (4), 742– 784.

Lechner, M., and C. Wunsch (2013): "Sensitivity of Matching-Based Program Evaluations to the Availability of Control Variables," *Labour Economics*, 21, 111–121.

Rinne, U., A. Uhlendorff, and Z. Zhao (2013): "Vouchers and Caseworkers in Training Programs for the Unemployed," *Empirical Economics*, 45 (3), 1089-1127.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70 (1), 41–55.

Rubin, D. B. (1973): "Matching to Remove Bias in Observational Studies," *Biometrics*, 29 (1), 159–183.

Smith, J. A., and P. E. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?", *Journal of Econometrics*, 125 (1-2), 305–353.

Vincent, J. L., J.-F. Baron, K. Reinhart, L. Gattinoni, L. Thijs, A. Webb, A. Meier-Hellmann, G. Nollet, and D. Peres-Bota (2002): "Anemia and Blood Transfusion in Critically Ill Patients," *The Journal of The American Medical Association*, 288 (12), 1499–1507.