The Role of Sample Size to Attain Statistically Comparable Groups : A Required Data Preprocessing Step to Estimate Causal Effects With Observational Data

Kolar, Ana

2021-10

Kolar , A & Steiner , P M 2021 , ' The Role of Sample Size to Attain Statistically Comparable Groups : A Required Data Preprocessing Step to Estimate Causal Effects With Observational Data ' , Evaluation Review , vol. 45 , no. 5 , 0193841X211053937 , pp. þÿ 1 9 5 2 2 7 . h t t p s : / / d o i . o r g / 1 0 . 1 1 7 7 / 0 1 9 3 8 4 1 X 2 1 1 0 5 3 9 3 7

http://hdl.handle.net/10138/341009
https://doi.org/10.1177/0193841X211053937

The role of sample size to attain statistically comparable groups - a required data preprocessing step to estimate causal effects with observational data

Ana Kolar* **and Peter M. Steiner**

Tarastats Statistical Consultancy, Fredrikinkatu 61A, Helsinki, Finland
Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
e-mail: ana@tarastats.com*; url:* http://www.tarastats.com

University of Maryland, Department of Human Development and Quantitative Methodology, 1233 Benjamin Building, College Park, USA
e-mail: psteiner@umd.edu

Abstract: Propensity score methods provide data preprocessing tools to remove selection bias and attain statistically comparable groups – the first requirement when attempting to estimate causal effects with observational data. Although guidelines exist on how to remove selection bias when groups in comparison are large, not much is known on how to proceed when one of the groups in comparison, e.g., a treated group, is particularly small, or when the study also includes lots of observed covariates (relative to the treated group's sample size). This article investigates whether propensity score methods can help us to remove selection bias in studies with small treated groups and large amount of observed covariates. We perform a series of simulation studies to study factors such as sample size ratio of control to treated units, number of observed covariates and initial imbalances in observed covariates between the groups of units in comparison, i.e., selection bias. The results demonstrate that selection bias can be removed with small treated samples, but under different conditions than in studies with large treated samples. For example, a study design with 10 observed covariates and eight treated units will require the control group to be at least 10 times larger than the treated group, whereas a study with 500 treated units will require at least, *only*, two times bigger control group. To confirm the usefulness of simulation study results for practice, we carry out an empirical evaluation with real data. The study provides insights for practice and directions for future research.

Keywords and phrases: causal inference, bias removal, propensity score methods, matching, experimental and observational study designs.

1. Introduction

Propensity score (PS) methods (Rubin, 1974, 1977, 1978, 1980; Rosenbaum and Rubin, 1983b; Holland and Rubin, 1988; Rosenbaum, 2002; Imbens and Rubin, 2015) are today one of the most widely applied methods for removing selection bias - a crucial data preprocessing step in causal inference applications. This preprocessing step is part of the PS methods' *design phase* where researchers are required to remove a sufficient amount of selection bias and obtain statistically comparable groups prior to proceeding with the *analysis phase* and

*All the correspondence goes to Dr. Ana Kolar via e-mail: ana@tarastats.com.

causal effect estimations. Groups are said to be statistically comparable when, for example, we have a treated and a control group which units are approximately identical with respect to observed (baseline) covariates and differ only with respect to an applied treatment. The statistically comparable groups are a required design framework to estimate causal effects without bias, however, it is important to note here that, if the set of observed covariates does not cover all confounding covariates, effect estimates cannot warrant causal interpretations.

Preprocessing data with PS methods is known to work well with large samples, but very little is known about their usefulness when the treated group, or even both groups in comparison are small. For example, small treated groups consisting of 10 to 50 units are common when we are interested in effect heterogeneity across small subpopulations, clusters or sites. For instance, estimating effects of educational interventions for subpopulations formed by gender, race and grade, or for each school separately. Small treated groups, are also common when a target population is small, for example, patients with rare diseases, or when the treatment is implemented at the cluster level, e.g. schools, districts, or states. Furthermore, in observational studies with 100 or less treated units, the number of covariates is often relatively large in comparison to the number of treated units. There is no evidence about the impact that the number of observed covariates has on removal of selection bias, particularly when the number of treated units is less than the number of observed covariates. For all theses reasons, an investigation of the role of sample size, when preprocessing data to remove selection bias, is of great importance.

Some within-study-comparisons (Pohl et al., 2009; Shadish, Clark and Steiner, 2008) suggest that PS methods can successfully remove selection bias when the treated group is small. However, there is little evidence about the minimum required sample sizes. A few publications have studied usefulness of PS methods when removing selection bias in small treated samples Rubin and Thomas (1996), Zhao (2004) and Luellen (2007), but, none of them studied treated samples of less than 25 units, and none of them studied the impact that the number of observed covariates has on the removal of selection bias when treated samples are small, relative to the number of observed covariates. Nonetheless, the studies suggest that the sample size of the treated group is a vital factor in the process of removing selection bias and that the guidelines provided for large samples may not be appropriate when the treated group is small, e.g., recommendations regarding a sample size ratio between a sample of control and a sample of treated units.

This article reviews the limited research available on this topic and presents a simulation study with as few as eight treated units. It is the first study that investigates the role of sample size (for the treated and control group) and the number of covariates in attaining statistically comparable groups. The study is by no means exhaustive, but it provides a general insight on important and previously unresearched topics.

We focus on treated samples of eight to 500 units and investigate the required sample size for the control group in order to obtain statistically comparable groups with respect to a set of observed covariates. In our simulation study we observe all the confounding covariates, therefore we can remove all the selection bias, but in practice creating statistically comparable groups, with respect to a set of observed covariates, removes *only* overt selection bias. We cannot remove the selection bias that is induced by unobserved covariates, a so called hidden bias. Furthermore, the observed covariates can remove the entire bias only if a set of causal

assumptions[1] is met, in particular the strong ignorability assumption.

We show that studies consisting of small treated samples require different guidelines to remove selection bias than studies with moderately large treated samples.[2] The main requirement in studies with small treated samples is that the control group has to be significantly larger than the treated group, meaning that the group ratio, $R = n_c/n_t$, with n_c and n_t denoting the sample sizes of the control and treated group, has to be large. A large group ratio guarantees that each of the few treated units will have a close match in the control group, and thus enables a satisfactory balance of the groups' covariate distributions. We show that the minimum required group ratio, R, depends on: (i) the size of the treated group, (ii) the number of observed covariates, and (iii) the level of initial covariate imbalances, i.e., an initial selection bias. The initial selection bias is a consequence of a systematic, nonrandom selection procedure or a broken randomised experiment. When the initial selection bias is large, a larger pool of control units is required. The need for control units increases further with large number of observed covariates and even more so when the treated sample gets smaller.

The article is organised as follows. Section 2 summarises past research on PS methods when groups of treated units are small and moderately large. Section 3 provides a brief overview of PS methods, introduces the notation and provides description of used acronyms (Table 1). Section 4 introduces the factors investigated in our simulation study, describes the simulation design, and presents simulation study results. Section 5 empirically evaluates the simulation results by using real observational data (the within-study comparison of Lalonde (LaLonde, 1986)). Section 6 provides a summary of the obtained insights, recommendations for practitioners and future research.

2. Past Research on Removing Selection Bias With Small Treated Samples Using Propensity Score Methods

In reviewing publications on PS methods for small and moderately large treated samples, we focus on publications that investigated removal of selection bias theoretically, or by carrying out simulation studies. Our intention is to learn what impacts the removal of selection bias in studies with small treated samples. We are particularly interested in the role that different methods to estimate propensity scores and different PS adjustment methods (PS matching, PS subclassification and PS weighting) have in the process of removing selection bias. Furthermore, we are interested in knowing the impact that the number of observed

[1] Assumptions required for *unbiased causal effect estimation*:(i) the observed covariates are measured before units are assigned to a treated or control group; (ii) the observed covariates are simultaneously related to both: the outcome variable and the indicator variable, i.e., the variable indicating to which group (treated or control) a unit belongs; (iii) all the relevant covariates are observed regardless of their statistical significance (Rubin and Thomas, 1996; Rubin, 1997); (iv) probability of being in a treatment group conditional on observed covariates has to be grater than 0 - often denoted as a positivity assumption; (v) there should be no interference among the units between the groups in comparison (Cox, 1958; Rubin, 1978); and (vi) only one form of treatment and control status is applied to each unit (Rubin, 1980). The third and fourth point present the so called *strong ignorability assumption*, often denoted also as the *unconfoundedness assumption*. The fifth and sixth point present the SUTVA - Stable Unit Treatment Value Assumption (Rubin, 1990).

[2] Moderately large samples of treated units consist of more than 100 units; small samples of less than 100 treated units.

covariates has in studies with small treated groups, for example, is removal of selection bias possible when the number of observed covariates is larger than the number of units in a treated group, and what role does the size of the control group play.

To our knowledge, only few publications have investigated implications of small and moderately large treated samples on removing selection bias with PS methods. There are three simulation studies that studied, to some extend, small and moderately large sample properties of PS methods: Rubin and Thomas (1996), Zhao (2004) and Luellen (2007). These studies indicate that guidelines established for large data sets might not be appropriate when treated samples are small. These studies show that not all of the available PS estimation and PS adjustment methods work well with small treated samples, and that in general, small treated samples require a much larger control group; meaning, that the group ratio, R, is required to be much larger than in studies with moderately large treated samples. Although, the influence of the number of observed covariates on R has never been investigated, their studies indicate that PS methods are able to remove selection bias when treated samples are small. The same view is shared by some within-study-comparisons (Pohl et al., 2009; Shadish, Clark and Steiner, 2008) where causal effect estimates from a randomised experiment are compared to the one obtained from a corresponding non-randomised study design, i.e., an observational study design. Their estimates have shown to be capable to approximate results from randomised experiments, indicating that even with small treated samples, selection bias can be removed (their studies consists of $n_t > 70$).

The most comprehensive study on sample size and PS methods was performed by Luellen (2007). He investigated: (i) treated samples of size $n_1 = 100$ and $n_2 = 500$, with a group ratio of $R = 1$ and 20 observed covariates; (ii) different methods for estimating PSs, i.e., logistic regression, classification trees, and ensemble methods such as bootstrap aggregating, boosted regression and random forest; and (iii) PS adjustment methods such as PS matching, PS subclassification, and PS weighting, performed independently and in combination with an additional covariate regression adjustment. His simulation results show, that all the investigated factors impact the removal of selection bias with sample size affecting the performance of both: the PS estimation and PS adjustment method. With small treated samples, only logistic regression performed well and only one-to-one PS matching removed all the selection bias. PS subclassification did not perform well with small treated samples, however, this is expected, because PS subclassification is meant to be used with large samples. Luellen also showed that PS weighting performs the worst of all the adjustment methods, regardless of the sample size. It is important to note here that Luellen's study simulates real observational data (drawing samples from a real data set), thus, he was unable to know whether the model, that he used to estimate PSs is correctly specified. PS weighting requires a correctly specified PS model in order to estimate treatment effects unbiasedly (Waernbaum (2012), Kang and Schafer (2007), Stuart (2010)). Luellen further showed that a combination of a PS adjustment method with an additional covariate regression adjustment performed better than any of the adjustment methods alone. Similar discovery was made already by Rubin (1973), later confirmed by Rubin (2001, 173-174), Rubin (2006) and Hirano and Imbens (2001).

Rubin and Thomas (1996) theoretically and analytically investigated the role of the group ratio, R, when using PS matching with one-to-one matching. Their results are based on

moderately large treated samples. Their findings show that with an initial bias, IB[3], of 0.5, 1.0, and 1.5, group ratios of 2, 3, and 6 are required to eliminate differences in covariate distributions between the groups of units in comparison. Therefore, the greater the difference in the treated and control groups' covariate distributions, i.e., the greater the initial selection bias, the more control units per treated unit are required. Rubin and Thomas noted that smaller treated samples require even larger group ratios, but without suggesting how large. They further tested their findings with a simulation study of ellipsoidal data using treated samples of 25 and 50 units with 5 and 10 observed covariates, group ratios, R, of 2, 5 and 10, and different levels of initial bias, IB of 0.0, 0.25, 0.5, 0.75, 1.0 and 1.5. For the $IB = 0.5$ their simulation results show that with $R = 5$ or $R = 10$ all the initial bias is removed, whereas for larger IB, these group ratios were not sufficient in removing initial bias.

Zhao's simulation study (2004) investigated small and moderately large treated samples in combination with one-to-one PS matching, covariate-Mahalanobis distance matching, covariate-and-PS matching and covariate-and-outcome matching. The smallest investigated sample had 100 treated and 400 control units, indicating a group ratio of four. Zhao's results show that the one-to-one PS matching removes selection bias the most effectively.

Based on published research and PS theory, we conclude the following: (i) treated samples smaller than 100 have not yet been sufficiently investigated whereas treated samples consisting of less than 25 units have not been investigated at all; (ii) the importance of the group ratio, R, in studies with small treated samples, where the amount of observed covariates is often relatively large compared to the number of treated units, has not been thoroughly investigated; (iii) estimating PSs with binomial regression methods, such as logistic regression, is a sensible choice in cases of small samples; (iv) success of PS weighting depends on knowing either the true model to estimate PSs or the true outcome model which is rare in practice; (v) PS subclassification requires large data sets, thus it is not suitable for treated samples consisting of less than 100 units; (vi) one-to-one PS matching is one of the most promising PS adjustment methods in small treated sample studies; (vii) the combination of PS adjustment with an additional covariate regression adjustment is highly recommended to further remove any remaining covariate imbalances, i.e., a remaining selection bias.

3. Propensity Score Methods

PS methods comprise of two separate parts: (i) *a design phase* which is outcome-free; and (ii) an *analysis phase*. In the *design phase* we use *only* observed covariates, X, and an indicator variable, mostly denoted as an assignment mechanism variable, W, to remove selection bias. The W indicates a group to which a sample unit belongs, e.g., a treated or a control group. The outcome variable, Y, must be excluded from this phase! Once selection bias is removed and statistically comparable groups are attained, we enter into the *analysis phase*. In the *analysis phase* we use data on the outcome variable, Y, to estimate desired statistical quantities, perform additional statistical adjustments, e.g., the covariate regression adjustment, and sensitivity analyses of causal claims.

[3]Initial bias, IB, is a measure of a selection bias, calculated by using Mahalanobis distance equation on the power of two: $((\mu_t - \mu_c)' \sum_c^{-1} (\mu_t - \mu_c))^2$ where μ_t and μ_c denote the covariate mean values of the treated and control group, respectively, with $\sum_c$ denoting the variance-covariance matrix of the control group.

This article primarily focuses on the *design phase* of PS methods, that is, studying factors that impact removal of selection bias when samples of treated groups are small. However, we also cover the *analysis phase*, where we apply covariate regression adjustment on data of statistically comparable groups, to further remove remaining bias and to estimate an average treatment effect on the treated, ATT. For this purpose, we briefly introduce a theoretical foundation of PS methods.

The foundation of PS methods consists of: (i) the *Rubin Causal Model*, RCM (Holland, 1986); and (ii) a *propensity score* (Rosenbaum and Rubin, 1983a). The RCM comprises of the *potential outcomes approach* and an *assignment mechanism* - the mechanism that informs us on how units are assigned to the treated and control group. For more details about the RCM please refer to: Rubin (2005, 2008); Imbens and Rubin (2015)). The role of the *propensity score* (PS) is to summarise information of all the observed (baseline) covariates, X, with respect to the group status, into a single value between zero and one, i.e., $0 < PS < 1$. Accordingly, the PS is defined as a conditional probability of being treated, i.e., belonging to one of the groups in comparison, given the observed covariates, X. Such a PS is a balancing score, $e(X)$, because the conditional distribution of covariates given the PS is the same for treated and control units (Rosenbaum and Rubin, 1983a). As a result, treated and control units with (approximately) the same PS, have (approximately) identical covariate distributions. The PS thus acts as a principal element in the process of removing selection bias and making groups in comparison statistically comparable.

The true PSs are known only when dealing with randomised experiments. With observational data we have to estimate them by using: (i) the observed covariates, X, and (ii) the indicator variable, i.e., an assignment mechanism, W, denoting to which group a unit belongs, e.g., treated ($W = 1$) or control ($W = 0$). PSs can be estimated by using binomial regression methods, e.g., a linear probability model, logistic or probit regression, or classification methods, such as, classification trees, boosted regression, neural networks or random forest (Keller, Kim and Steiner (2015); McCaffrey, Ridgeway and Morral (2004); Siroky (2009); Westreich, Lessler and Jonsson-Funk (2010)). Once PSs are estimated, a PS adjustment method, such as PS matching, PS subclassification or PS weighting (inverse-propensity score weighting) is used to remove selection bias and attain statistically comparable groups.

Before we apply an adjustment method, we are required to asses a statistical comparability of groups in comparison. As mentioned earlier, this comparability is assessed only with respect to observed covariates and based on that provides us with the information about the magnitude of the initial selection bias. For example, we asses whether there is overlap between covariate distributions of the groups in comparison in terms of: (i) the *common support* [4]; and (ii) the *shape of covariate distributions* [5]. The lack of overlap in the *shape of covariate distributions* in small sample studies makes the process of attaining comparable groups more difficult because comparable units are harder to be found when only a limited number of units is available. On the other hand, the lack of *common support* reduces sample size due to deletion of units that do not share a *common support*. In case of complete absence of the *common support*, the attainment of statistically comparable groups becomes impossible and

[4]The *common support* is the area where covariate distributions of groups in comparison overlap in terms of X-axis in two-dimensional Cartesian coordinate system.

[5]An overlap in the *shape of covariate distributions* refers to the overlap in distributional forms of observed covariates of groups in comparison.

the *design phase* cannot be completed, i.e., we cannot proceed with the *analysis phase*.

When assessing statistical comparability of the treated and control group, we can assess each observed covariate, or the PSs by using (i) *the standardised mean difference* (SMD) between sample means of observed covariates or the estimated PSs. In case of PSs, the equation is the following: $(\bar{e}(x_t)_i)$ - $(\bar{e}(x_c)_i)$ $/\sqrt{(s_t^2 + s_c^2)/2}$ (Austin, 2009; Flury and Riedwyl, 1986; Rosenbaum and Rubin, 1985), where $\bar{e}(x_t)_i$ and $\bar{e}(x_c)_i$ denote estimated PSs of units in the treated and the control group with s_t^2 and s_c^2 denoting variances of the estimated PS for the treated and control group; and (ii) the *a variance ratio* between PSs of the treated and control group, s_t^2/s_c^2 (Rubin, 2001) to indicate similarity of their covariate distributions' variances.

Instead of estimated PSs, it is recommended to use their logits, $\bar{l} = log[(\bar{e}(X)_i)/(1-\bar{e}(X)_i)]$ (Rubin and Thomas, 1992; Rubin, 2001). By using PS-logits, an absolute value of SMD is calculated as $|\bar{l}_t - \bar{l}_c|$ / $\sqrt{(s_{lt}^2 + s_{lc}^2)/2}$ with $\bar{l}_t$ and $\bar{l}_c$ denoting estimated PS-logits of the treated and control group, respectively. Such calculation of SMD can also be used as a measure of a remaining bias, RB (the measure that we use for the RB in our simulation study). With PS-logits, the *variance ratio*, VR, is calculated as s_{lt}^2/s_{lc}^2 with s_{lt}^2 and s_{lc}^2 corresponding to variances of the estimated PS-logits in the treated and control group.

To assess whether a sufficient amount of initial selection bias is removed (this means that the remaining bias, RB is considered to be negligible from the perspective that it can be safely removed further with an additional statistical adjustment in the *analysis phase* of PS methods), we use the following recommendations: (i) the VR should be close to one but not smaller than 0.5 or larger than 2 (Stuart and Rubin, 2007); (ii) the RB should be less than 0.1 (Austin, 2011; Cochran and Rubin, 1973; Cochran, 1968), however, when this is the case, an additional covariate regression adjustment is recommended to remove covariate imbalances further, i.e., the RB (Hirano and Imbens, 2001; Rubin, 2006); (iii) when RB is bigger than 0.1, but it remains below 0.2, the covariate regression adjustment is *required* (Rubin, 1979); (iv) RB of 0.2 or more might be of a concern, meaning that an additional covariate regression adjustment might not be successful in removing the remaining bias and it could even introduce an additional bias (Rubin, 2001); (v) removing selection bias completely, i.e., $RB = 0$, is rarely possible, therefore, when $RB < 0.1$ and $0.5 < VR < 2$, it is typically considered that there is a negligible imbalance in observed covariates, i.e, we have attained statistically comparable groups, with respect to observed covariates.

4. Simulation Study

The aim of the simulation study is to investigate which factors influence removal of selection bias when the treatment group is particularly small and the number of observed covariates is large, relative to the size of the treated group. For example, what is the required size of the control group, n_c^* when the treated group is $n_t = 8$ and the number of observed covariates is $p = 10$. In other words, what is the minimum required group ratio, $R^* = n_c^*/n_t$?

There are three main objectives of this study. First, to determine the minimum required group ratio, R^*, that reduces selection bias to negligible levels such that $RB < 0.15$[6] and

[6]The reason for using remaining bias, RB, larger than the negligible RB of 0.10 standard deviations is the following: we aim to determine a minimum required group ratio, R^*, therefore we allow for a bit larger

TABLE 1
Description of used acronyms

$PS(s)$	Propensity score(s)
R	Group ratio, e.g., n_c/n_t
R^*	Minimum required group ratio
IB	Initial (selection) bias (in the simulation study calculated by using Mahalanobis distance equation (please refer to footnote 3 for the equation))
RB	Remaining (selection) bias, i.e., remaining covariate imbalances
VR	Variance ratio
SMD	Standardised mean difference
p	Number of observed covariates
n_c	Sample size of a control group
n_t	Sample size of a treated group
s_t^2, s_c^2	Variance of estimated PS for treated and control group, respectively
$\bar{l}_t$, $\bar{l}_c$	Estimated PS-logits of the treated and control group, respectively
MSE	Mean square error
MSS	Mean sum of square explained
DF	Degrees of freedom
method	Greedy or optimal matching algorithm
ATT	Average treatment effect on the treated
ATE	Average treatment effect
SE	Simulation standard errors of $\widehat{ATT}(SE_{\widehat{ATT}})$ designed by using standard deviation of treatment effect estimates across 1,000 iterations, $SE_{\widehat{ATT}} = s_{\widehat{ATT}}/\sqrt{1000}$

$0.5 < VR < 2$ or more formally, $R^* = min\{R : RB(R) < 0.15$ and $0.5 < VR(R) < 2\}$ where $RB(R)$ and $VR(R)$ indicate that the RB and VR are functions of R. Second, to study how R^* changes when the number of observed covariates, p, varies. Third, to study the impact that the strength of the selection mechanism has on R^*, that is, how does R^* changes with different levels of IB. Furthermore, we also examine the performance of two popular matching algorithms, i.e., greedy and optimal matching. The *greedy* algorithm is a nearest-neighbour matching algorithm concentrating only on well-matched pairs, whereas the *optimal* algorithm is concerned with pairs being well-matched also within each group of comparison (Rosenbaum, 1989). The *optimal* algorithm is particularly useful when there is lots of competition for controls (Rosenbaum, 2002), for examples in cases of large group ratios $R > 100$.

Our main objectives are addressed by studying factors that are known in the *design phase* of PS applications, such as n_t, R, p, or estimable in the *design phase*, e.g. IB. The factors are presented in Table 2.

RB than a negligible RB. The remaining RB we remove with an additional covariate regression adjustment in the *analysis phase* of PS methods, as presented in Section 4.2).

TABLE 2
Simulation study design

Factors known or estimable in the design phase of PS methods	Factor's levels for **Small** treated **sample study**	Factor's levels for **Moderately large** treated **sample study**
Treated sample size - n_t	{8,10,15,20,25,30,50,100}	{200,500}
Group ratio - R	{1:100}	{1:9}
Number of observed covariates - p	{10,15,20,30}	{10,15,20,30}
Initial covariate imbalances - IB	{0.5,1.0,1.5}	{0.5,1.0,1.5}
Method, i.e., *matching algorithm*	{greedy, optimal}	{greedy, optimal}
Full factorial design	**8x100x4x3x2 = 19200**	**2x9x4x3x2 = 432**

4.1. Data generation

The simulation design is based on four target populations, one for each investigated number of observed covariates, $p \in \{10, 15, 20, 30\}$. Each target population consists of $N_{p_i} = 1,125,000$ units from which we draw repeated samples without replacement. Such data generation design enables us to investigate the influence of p on the minimum required group ratio, R^*.

The observed covariates of each target population, X, are generated as independent, standard normally distributed variables: $X \sim N(0,1)$. The outcome variable, Y, for each target population is generated as $Y = \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$ with $\varepsilon \sim N(0,1)$. We assume a treatment effect of zero, thus, the treatment and control outcomes are identical.

Our four target populations are comparable, with regard to the correlation structure between the linear combination of X and Y, by calculating beta coefficients for each p as $\beta_p = \sqrt{Q/p}$ where Q denotes a covariance between the outcome variable and the linear combination of observed covariates: $Q = Cov(Y, \beta \sum X_i)$. The factor Q is set to 0.35 and it represents data that could be observed in practice with a coefficient of determination between Y and X of $R^2_{Y,\underline{X}} = Q/\sqrt{(Q+1)Q} = 0.51$ (derivation of $R^2_{Y,\underline{X}}$ can be found in Appendix B). The following values of the beta coefficients, β_p, are obtained based on $Q = 0.35$: $\beta_{p=10} = 0.19$, $\beta_{p=15} = 0.15$, $\beta_{p=20} = 0.13$ and $\beta_{p=30} = 0.11$. These coefficients are then used to generate our target populations, N_{p_i}.[7]

For each target population, we generate an assignment variable, W_i, by randomly drawing from a Bernoulli distribution with probabilities consisting of the true PSs, $e(X)$: $W_i \sim Bernoulli(prob = e(X_i))$. These true PSs are calculated for each strength of the selection mechanism, to create an initial bias, IB, of $0.5, 1.0, 1.5$. Consequently, each target population,N_{p_i}, consists of three such populations corresponding to the three levels of the initial bias.

We end up with three target populations within each of the four target populations, N_p.

[7]The betas of all the observed covariates in the outcome variable equation are the same. For example, if $\beta_{p=10} = 0.19$ then $\beta_1 = \beta_2 = ... = \beta_{p=10} = 0.19$. Because our assignment mechanism is ignorable, our results would not change if betas would vary.

For example, within a target population consisting of 10 observed covariates, $N_{p=10}$, we have $N_{p=10,IB=0.5}$, $N_{p=10,IB=1}$ and $N_{p=10,IB=1.5}$. The true PSs are calculated according to $e(X) = logit^{-1}\gamma(X_1 + ... + X_p)$ where γ is the coefficient determining the strength of the selection mechanism enforcing imbalances in covariate distributions between the groups of units in comparison, i.e., the initial selection bias. Table 3 displays the coefficients used for different values of IB and p.

The simulation study is programmed and analysed in R (R Core Team 2015). The R package MatchIt (Ho et al., 2011) is used for matching treated and control units.

TABLE 3
Gamma coefficient, γ to calculate true propensity score, $e(X)$, as a function of an initial bias, IB, and a number of observed covariates, p.

IB	$p = 10$	$p = 15$	$p = 20$	$p = 30$
0.5	0.24	0.19	0.17	0.14
1.0	0.35	0.29	0.25	0.21
1.5	0.46	0.37	0.33	0.27

4.2. Data simulation

For each target population, we draw 1000 repeated samples (without replacement). In each of these 1000 samples we perform one-to-one PS matching (without replacement) on the logit of the estimated PSs, $\hat{l} = log[(\hat{e}(X)_i)/(1 - \hat{e}(X)_i)]$ (Rubin, 2001). The PS logits are estimated via logistic regression according to $logit(W) = \lambda_o + \lambda_1 X_1 + ... + \lambda_p X_p$. Matching is performed with the *greedy* and *optimal* matching algorithm.

We summarise the simulated matched data across 1000 iterations by using: (i) the average RB and VR, and the average of the estimate of the $\widehat{ATT}$[8]; and (ii) the standard deviation of the 1000 simulated replications of $\widehat{ATT}$ which we use to construct confidence intervals of the ATT. The ATT is estimated from the matched data in combination with additional covariate regression adjustment and calculated as $\hat{\tau} = (\overline{y_t} - \overline{y_c}) - \hat{\beta}_d(\overline{l_t} - \overline{l_c})$, where the regression coefficient, $\hat{\beta}_d$, is obtained from the regression of $y_{d_j} = y_{t_j} - y_{c_j}$ on $\hat{l}_{d_j} = \hat{l}_{t_j} - \hat{l}_{c_j}$ with y_{d_j} being the difference in the outcomes of the matched pairs and $\hat{l}_{d_j}$ the differences in the PS logit of the matched pairs. According to Rubin (1979), such covariate regression adjustment is the most natural adjustment in pair matching settings and it produces the least biased treatment effect estimate, particularly when group ratios are large, which is the case in studies with small treated groups.

[8]The $\widehat{ATT}$ is the average treatment effect estimate for the subpopulation of those units to which treatment is applied. The ATT is beside the average treatment effect, ATE, one of the two common causal quantities of interest, frequently more useful than ATE. In cases of small treated samples, ATT can provide a more reliable results than ATE.

4.3. Analysis of Simulated data

We perform two different analyses of the simulated data: (i) an analysis of variance - ANOVA with a remaining bias RB as a dependant variable - to investigate which of the studied factors n_t, R^*, p, IB, and *matching algorithm*, has the strongest impact on removing selection bias, that is, results in the least remaining bias; and (ii) a numerical and graphical descriptive analyses to show how the studied factors interact.

Because small treated samples in general require large group ratios, we could not perform ANOVA with our original simulation design (Table 2). The estimation of PSs with very small treated samples, i.e., n_t of 8, 10, 15, 20, 25, and $R = 1$ resulted in extreme PS estimates of 0 and 1 for all sets of p and in all simulation replications. When $\hat{e}(X) = 0$ or 1, PSs are not effective balancing scores. Table 4 presents the combination of p, n_t and R where PSs are not effective balancing scores. As we can see from the table, only when $R > 7$ estimation of effective PSs is possible with almost all combinations of p and n_t except for $n_t = 8$ and $p = 30$.

TABLE 4

A combination of factors and their levels for which estimated PSs are not effective balancing scores

	$R = 1$	$R = 2$	$R = 3, 4, 5, 6$	$R = 7, 8, 9, 10, 11, 12$
p	n_t	n_t	n_t	n_t
10	8, 10, 15, 20, 25	/	/	/
15	8, 10, 15, 20, 25	/	/	/
20	8, 10, 15, 20, 25	/	/	/
30	8, 10, 15, 20, 25	8, 10, 15	8, 10	8

Note. Cells denoted with / present cases when estimated PSs are not effective balancing score. p - number of observed covariates; R - group ratio; n_t - treated sample size.

Because some combinations of p, n_t and R could not estimate PSs that are effective balancing scores, i.e., $0 < PS < 1$, we excluded those cases from further analysis of simulation data. As a result we carried out three separate analyses of variance (Table 5): (i) Small treated sample study 1; (ii) Small treated sample study 2; and (iii) Moderately large treated sample study. Although these three studies are not fully comparable, they provide reasonable insight into the most influential factors when removing selection bias in studies with small treated samples.

4.4. Simulation results

The ANOVA results are presented in Table 5 where the studied factors are sorted by the influence they have on removal of selection bias in small treated sample studies, that is, by decreasing order of the mean sum of squares explained (MSS). For example, the first three factors in Table 5 have the biggest impact on remaining bias, RB.

The summary of descriptive statistics results are presented in Figures 1 - 3, and in Table 8 and 9 in the Appendix, showing the minimum required group ratio, R^*, i.e., the smallest group ratio for which RB is smaller than 0.15 standard deviations and VR is between 0.5 and 2.

4.4.1. Analysis of Variance

The ANOVA analyses consist of the main effects and all the interactions, i.e., up to five-way interactions. We display the influential factors which have $MSS > 0$ and are rounded to two decimal places. Among the top five influential factors in the *small treated sample study 1* and *2* (Table 5) are n_t, p, IB, R and $n_t : p$. Few other interactions that are related to the remaining bias in small treated samples are the following: $n_t : p$, $n_t : IB$, $p : IB$ and $R : p$. Its interrelatedness is confirmed also with descriptive statistics presented in the next section. Success of removing selection bias thus primarily depends on the size of the treated group, the number of observed covariates, the size of initial selection bias and the number of control units relative to n_t (presented with R).

The *moderate large treated sample study* has the same top five influential factors, but with one important exception: instead of the two-way interaction $n_t : p$, it has $R : IB$. This means that removal of selection bias with moderately large n_t does not depend that much on the number of observed covariates that are studied here, i.e., $p = 10, 15, 20, 30$, but rather more on the number of control units, relative to the number of treated units, i.e., R, when initial selection bias varies. It is safely to assume that the $n_t : p$ interaction would become an important factor also for moderately large treated samples when the number of observed covariates would be bigger than what we study here.

Furthermore, the *method* (greedy vs. optimal matching algorithm) is an influential factor in *small sample studies*, but not in the *moderately large sample study*. To clarify this influence, we compare mean-squared-errors (MSE) of the $\widehat{ATT}_{optimal}$ and $\widehat{ATT}_{greedy}$ estimates. The comparison indicates that the *optimal* algorithm performs slightly better than the *greedy* algorithm, but the difference is not statistically significant. Despite the non-significant difference, the difference is bigger for small treated samples than moderately large treated samples. Such results are expected, because *optimal* algorithm is known to perform better in studies with larger R, i.e., when bigger pools of control units are required (Rosenbaum, 2002).

4.4.2. Descriptive Analyses

The descriptive analysis is performed as explained in section 4.4. The results are presented with Figures 1 - 3, and in Tables 8 - 9 that can be found in the Appendix A. The results show that smaller n_t require larger R^*, which means that with less treated units a comparatively bigger pool of control units is required. Furthermore, R^* for small treated samples with $n_t < 100$, is greatly influenced by p. For instance, when the number of observed covariates increases from $p = 10$ to $p = 30$, R^* increases by a factor of four for $n_t = 8$ whereas it barely changes for $n_t \geq 100$. The relationship between R^* and p shows a strong exponential functional form, particularly with $n_t < 20$ (Figure 1). Although initial bias, IB, can be removed with small n_t to negligible levels and such study design can produce reliable estimates in terms of selection bias removal, studies with smaller n_t produce less precise effect estimates of causal effects than studies with larger n_t. For example, with $IB = 0.5$ and p of 10 or 15, standard errors, $SE_{\widehat{ATT}}$, are almost seven times bigger (0.02/0.003=6.7 - Table 8) for $n_t = 8$ in comparison to n_t of 200 or 500. This ratio of $SE_{\widehat{ATT}}$ between small and moderately large treated samples increases further with larger IB or with more p. For instance, when $IB = 1.5$ and p is 30, the $SE_{\widehat{ATT}}$ for $n_t = 8$ is more than ten times bigger (0.027/0.002=13.5 - Table 9) than the

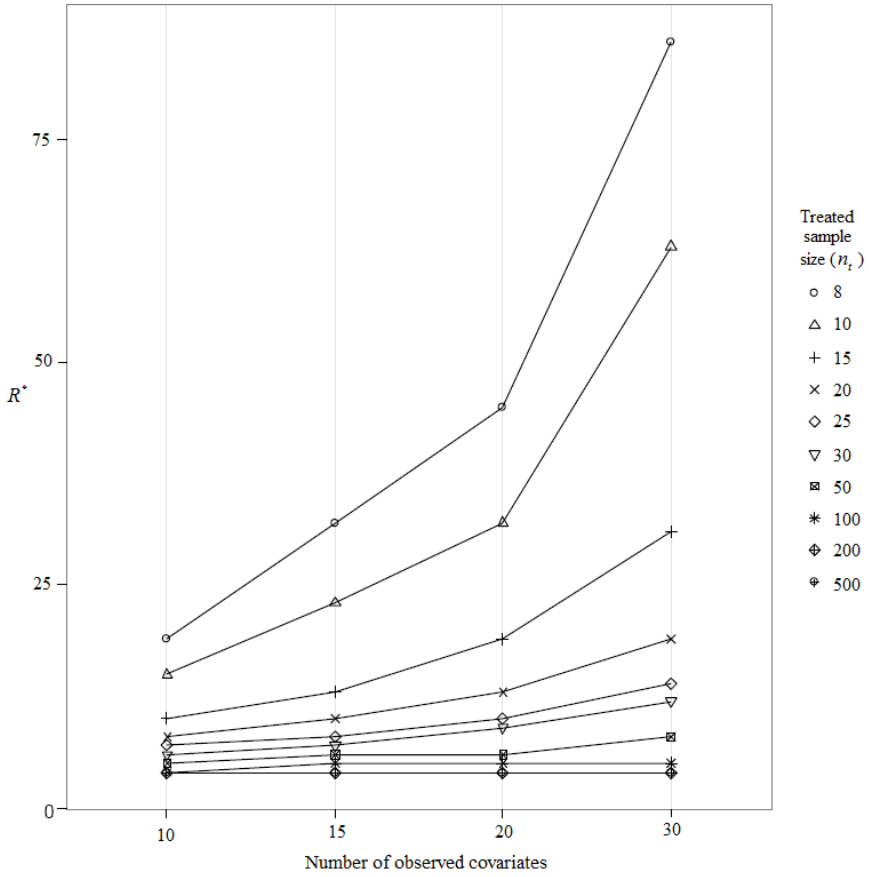
$SE_{\widehat{ATT}}$ for $n_t = 500$.

TABLE 5
Factor designs and results of ANOVA analyses

Small treated sample study 1			Small treated sample study 2			Moderately large treated sample study		
Factor design			Factor design			Factor design		
$n_t = \{8, 10, 15, 20, 25, 30, 50, 100\}$ $R = \{13 : 100\}$ $p = \{10, 15, 20, 30\}$ $IB = \{0.5, 1.0, 1.5\}$ method $= \{greedy, optimal\}$			$n_t = \{20, 25, 30, 50, 100\}$ $R = \{2 : 100\}$ $p = \{10, 15, 20, 30\}$ $IB = \{0.5, 1.0, 1.5\}$ Method $= \{greedy, optimal\}$			$n_t = \{200, 500\}$ $R = \{1 : 9\}$ $p = \{10, 15, 20, 30\}$ $IB = \{0.5, 1.0, 1.5\}$ Method $= \{greedy, optimal\}$		
Results			Results			Results		
Factor	DF	MSS	Factor	DF	MSS	Factor	DF	MSS
n_t	7	4.60	IB	2	2.26	R	7	4.60
p	3	2.53	R	98	1.48	IB	2	2.37
IB	2	2.37	n_t	4	1.29	$R : IB$	21	0.32
$\boldsymbol{n_t : p}$	21	0.32	p	3	0.96	n_t	87	0.18
R	87	0.18	$\boldsymbol{n_t : p}$	12	0.09	p	609	0.01
$\boldsymbol{n_t : IB}$	14	0.09	$R : IB$	196	0.05	$R : n_t$	174	0.01
$\boldsymbol{p : IB}$	6	0.03	$\boldsymbol{n_t : IB}$	8	0.04			
$\boldsymbol{method}$	1	0.02	$R : n_t$	392	0.04			
$R : n_t$	609	0.01	$\boldsymbol{R : p}$	294	0.03			
$R : IB$	174	0.01	$\boldsymbol{p : IB}$	6	0.02			
$\boldsymbol{R : p}$	261	0.01	$\boldsymbol{method}$	1	0.01			

Note. The difference in factor designs between *Small treated sample study 1* and *2* is explained in the paragraph after Table 4. The factors in **bold** are those that are *not* among the influential in the *moderately large treated sample study.*

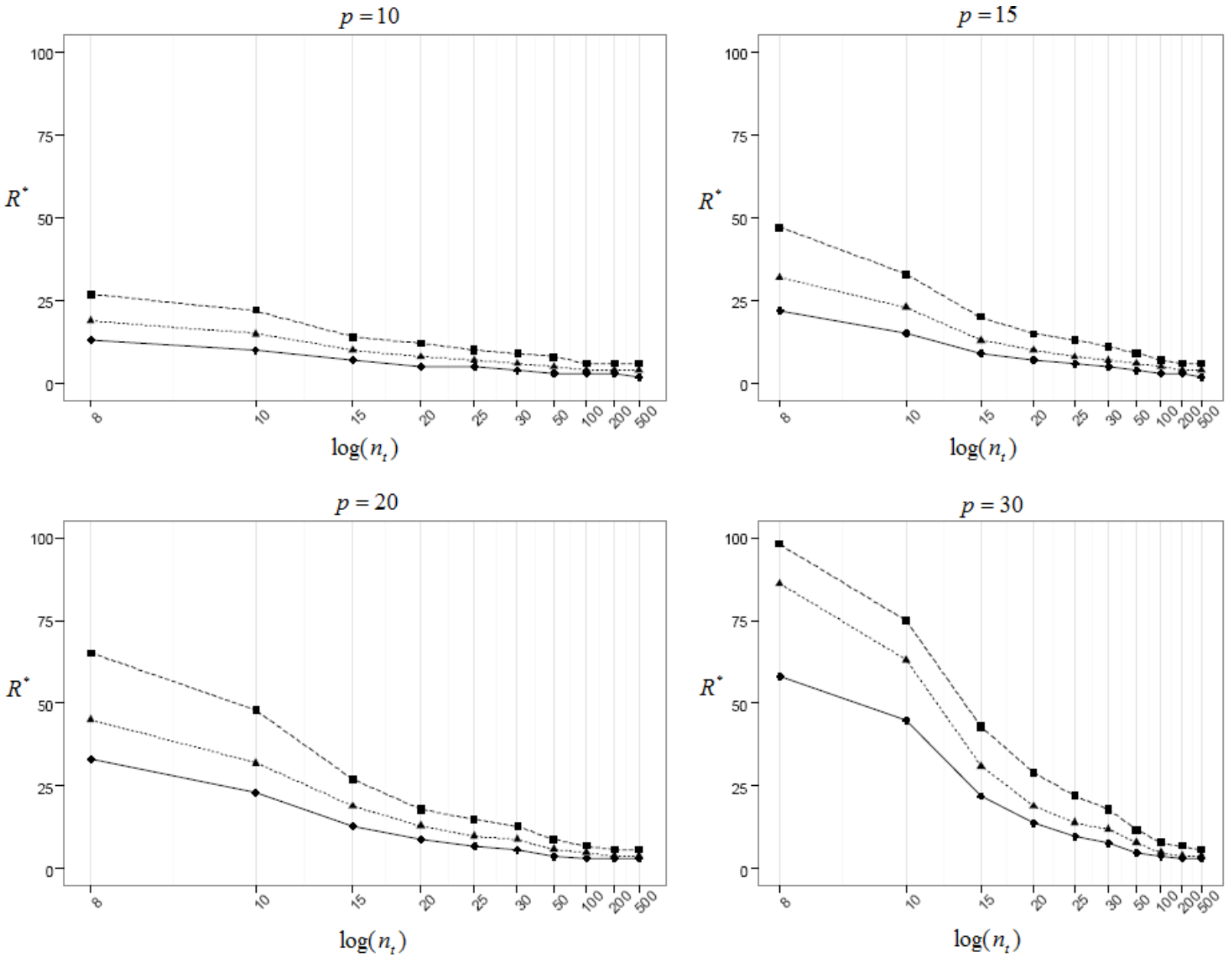
Fig 1: Relationship between the number of observed covariates, p, and the minimum required group ratio, R^*, for different treated samples (presented with lines) when initial bias, $IB = 1$.



Note. Other IB produce very similar depictions. Treated samples, n_t, of 200 and 500 are presented with the same line (the first line from bottom-top) due to the same values of R^*.

When IB increases, R^* also increases. For example, in the case of $p = 10$ and $IB = 0.5$, our results demonstrate that we need at least $R^* = 13$ for $n_t = 8$. For $IB = 1.0$ and $IB = 1.5$ the R^* should be at least 19 and 27, respectively. (Figure 2).

Fig 2: A relationship between different initial biases, IB (depicted with lines), and the required minimum group ratio, R^*, for different treated samples, n_t and different number of observed covariates, p.



4.5. Estimated versus true propensity scores

The PSs need to be estimated when *true* PSs are not known. The *true* PSs are usually known when dealing with randomised experiments, whereas when dealing with observational data, PSs need to be estimated. The above simulation results reflect situations that could be observed in real life scenarios when observational data is used and PSs are required to be estimated. Although, Rubin and Thomas (1992) argue that in settings with normally distributed covariates, matching on the *estimated* PSs is preferred even if *true* PSs are available, our aim here is to study whether this is the case also when treated samples are small. In particular, we are interested in knowing whether the required minimum group ratio changes when *true* instead of *estimated* PSs are used. For this purpose we run another simulation study which is equally framed as our previous simulations, except that now we use *true* PSs (the values that were used in our previous simulations to allocate units to the treated and control group). The results show that there are some major differences in the rankings of the most influential factors between these two simulation studies (Table 6). The factors in **bold** are influential in the study with *estimated* PSs, but have no importance in the study with *true* PSs.

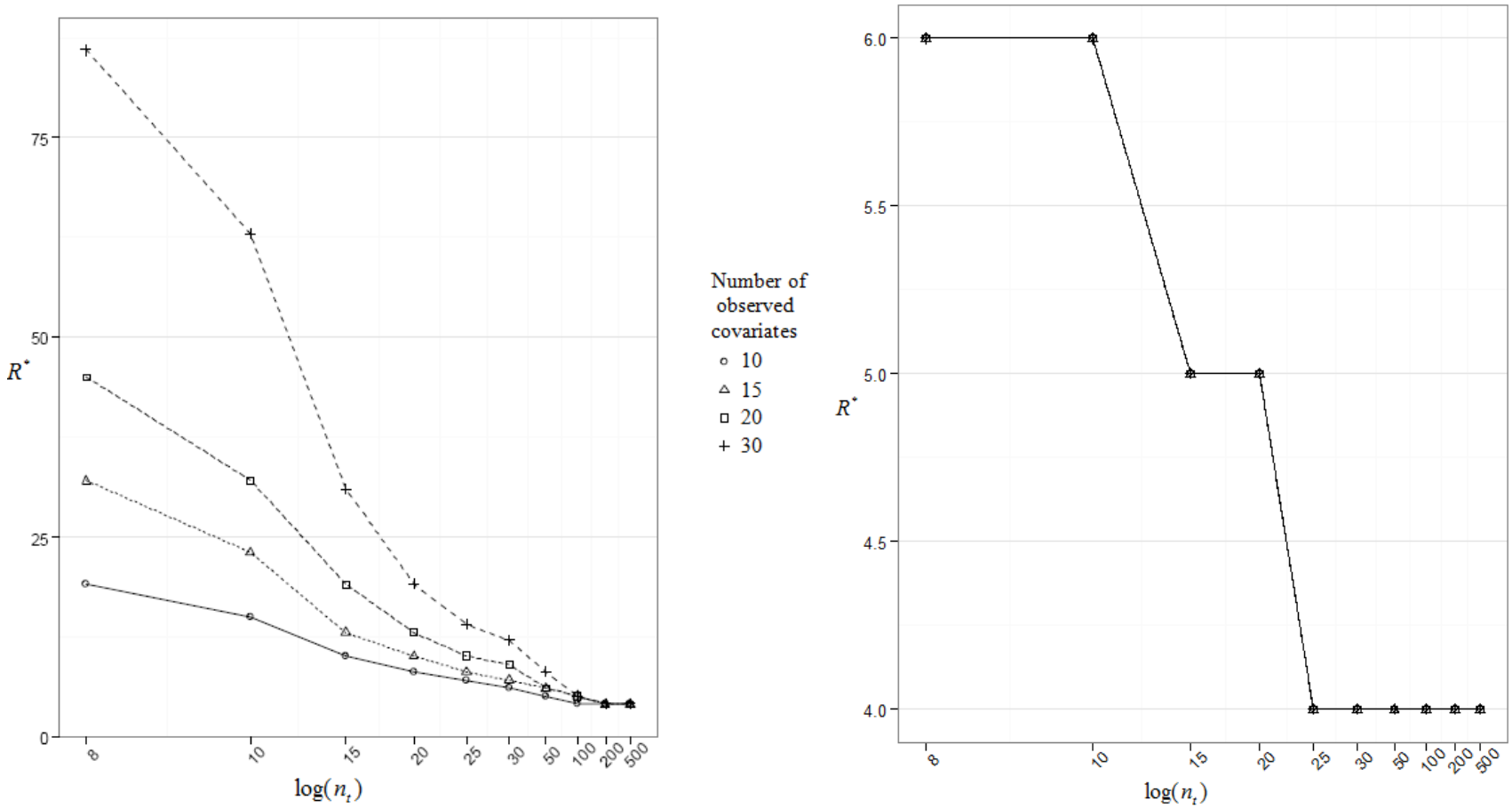
TABLE 6
ANOVA results for estimated and true PS

Estimated propensity score			True propensity score		
Factor	*DF*	*MSS*	Factor	*DF*	*MSS*
n_t	7	4.60	IB	2	0.40
$\boldsymbol{p}$	3	2.53	n_t	7	0.05
IB	2	2.37	R	87	0.02
$n_t : p$	21	0.32	$n_t : IB$	14	0.01
R	87	0.18			
$n_t : IB$	14	0.09			
$\boldsymbol{p : IB}$	6	0.03			
$\boldsymbol{method}$	1	0.02			
$\boldsymbol{R : n_t}$	609	0.01			
$\boldsymbol{R : IB}$	174	0.01			
$\boldsymbol{R : p}$	261	0.01			

The factor design for both studies is the same as in the *Small treated sample study 1* (Table 5). The reason for such discrepancy in influential factors between the study with *true* PSs and the study with *estimated* PSs (Table 6) lies in the estimation process of PSs. In small treated sample studies, *estimated* PSs are not as precise as *true* PSs. Therefore, the *estimated* PSs are not as effective as the *true* PSs in the process of removing selection bias. The combination of small treated samples and large amount of observed covariates makes the estimation process even more challenging and sometimes impossible. In other words, as we can see from Figure 3, when *true* PSs are available, much smaller group ratios are required and the number of observed covariates does not have an impact on the minimum required group ratio. What does this mean for practice? When treated samples are small and there

is no possibility for a sufficient pool of control units (as recommended in Tables 8-11 in the Appendix), than a carefully designed randomised experiment is our best option.

Fig 3: Comparison of the number of observed covariates, p, (presented with lines) versus the treated sample size, n_t, and their correspondingly required minimum group ratio, R^*, with $IB = 1$. Study with *Estimated PSs* (left) and study with *True PSs* (right).



Note. Other IB produce very similar depictions. For the graph on the right: lines which present each covariate set are on the top of each other. The p does not have an impact on the R^* when true PSs are used.

5. Empirical evaluation of simulation results

The idea of the empirical evaluation is to find out whether our theoretical simulation results and insights can be reliably applied to practice. To do so, we used an observational data set that comes with a benchmark estimate from a corresponding randomized experiment. Not many data sets fulfil this criterion, but the data of Lalonde's (1986) within-study comparison does. Lalonde compares results of a randomised experiment with results obtained from observational data. In attempt to replicate effect estimates of the randomised experiment, he uses the following approaches to remove selection bias and estimate effects from observational data: least squares regressions, an instrumental variable approach and Heckman's (1979) two-step procedure. Lalonde does not succeed in replicating results. A decade later, Dehejia and Wahba (1999) use PS matching as an approach to remove selection bias from observational data and succeed in replicating effect estimates of the randomised experiment. For more discussion on this topic refer to: Smith and Todd (2001, 2005) and Dehejia (2005).

The Lalonde data examine the effect of labour market training programmes on earnings and are available in the R package MatchIt (Ho et al., 2011). His observational data consists of 445 observations with a treated sample $n_t = 185$, a control sample $n_c = 260$ and measurements on eight observed covariates, $p = 8$, (half discrete and half continuous), a continuous earnings outcome, an assignment variable (of having participated in the labour market programme or not) and an initial selection bias, $IB = 0.19$.

In order to determine the minimum required group ratios, R^*, when initial bias is 0.19, we performed an additional theoretical simulation study with data generated as explained in section 4.1, but with $\gamma = 0.09$ in the logistic data generating model in order to obtain IB of 0.19. The generated data is simulated as explained in section 4.2, optimal PS matching algorithm is used to create comparable groups. The results of this simulation study (presented in the left side of Table 7) give us the minimum required group ratio, R^*, which we use in the simulation study with Lalonde's observational data. We evaluate whether such R^* and small n_t can remove selection bias from Lalonde's observational data so that $RB < 0.15$ and $0.5 < VR < 2$. We are also interested in seeing whether we can obtain an unbiased treatment effect, $\widehat{ATT}$, that is close to to the treatment effect of the randomised experiment provided in Lalonde (1986), $\tau = 1794$)).

5.1. Simulation study with Lalonde data

The Lalonde's observational data, which consists of 445 units, serves as the target population in this simulation study from which samples (selected n_t and corresponding n_c) are randomly drawn without replacement. Treated samples of the following sizes are selected: $n_t \in \{8, 10, 15, 20, 25, 30, 50, 100\}$ with R^* of 11, 8, 5, 4, 4, 3, 3, and 2, respectively as presented in Table 7). Accordingly, sample sizes of the control groups are the following $n_c \in \{88, 80, 75, 80, 100, 90, 150, 200\}$. To illustrate this design further: with $n_t = 8$ the assigned R^* is 11 and the corresponding $n_c = 88$. With n_t of 10, the R^* is 8 and the n_c is 80, and so on.

The simulation study with the Lalonde data uses the same simulation design as our theoretical simulation study, but with an important difference: we are not trying to find R^*, but instead, as described in the previous section, we provide R^* that are obtained from the theoretical simulation study for selected n_t.

TABLE 7
Empirical evaluation with Lalonde observational data set

Theoretical simulation study results
$p = 8$ and $IB = 0.19$

n_t	R^*	RB	VR	$99\%CI$	SE
8[a]	11	0.14	1.42	[-0.05,0.08]	0.021
10	8	0.13	1.39	[-0.11,0.02]	0.019
15	5	0.13	1.38	[-0.08,0.02]	0.015
20	4	0.13	1.34	[-0.07,0.02]	0.013
25	4	0.09	1.27	[-0.06,0.02]	0.011
30	3	0.12	1.32	[-0.06,0.01]	0.011
50	3	0.07	1.19	[-0.03,0.02]	0.007
100	2	0.09	1.23	[-0.03,0.01]	0.006
200	2	0.05	1.15	[-0.02,0.01]	0.004
500	2	0.03	1.10	[-0.01,0.01]	0.002

Simulation results of the Lalonde data
$p = 8$ and $IB = 0.19$

n_t	R^*	RB	VR	$\widehat{ATT}$	$99\%CI$	SE	$\widehat{ATT} - \tau$
8[b]	11	0.11	1.74	1380	[577,2183]	230	-414
10	8	0.11	1.69	1908	[1314,2501]	183	114
15	5	0.12	1.56	1760	[1400,2121]	121	-34
20	4	0.11	1.50	1807	[1560,2054]	86	13
25	4	0.08	1.30	1767	[1570,1963]	70	-27
30	3	0.11	1.37	1776	[1606,1945]	62	-18
50	3	0.05	1.15	1876	[1770,1982]	40	82
100	2	0.05	1.12	1789	[1723,1855]	25	-5

[a] the logistic regression used for estimating PSs resulted in extreme values of 0 and 1 for 0.5% of simulation replications.
[b] the logistic regression used for estimating PSs resulted in extreme values of 0 and 1 for 32% of simulation replications.

5.2. Results

The Lalonde data simulation results presented in the right side of the Table 7 are consistent with the results of our theoretical simulations, meaning that for all $n_t > 8$ the RB is below 0.15, and the VR does not go below 0.5 or above 2. The obtained $\widehat{ATT}$ are very close to the $\widehat{ATT}$ of the randomised experiment, $\tau = 1794$. (LaLonde, 1986) and also to the $\widehat{ATT}$ obtained by Dehejia and Wahba (1999), $\tau = 1788$.

The only inconsistency found is with the smallest investigated treated sample, $n_t = 8$. With this treated sample the simulation results in too many simulation replications (32%) with estimated PSs not being effective balancing scores, i.e., $e(X) = 0$ or 1. This was not the case in the theoretical simulation where the percentage was only 0.5% for $n_t = 8$. Such inconsistency likely results from the nature of observed covariates (not all the observed covariates of the Lalonde data are normally distributed as in the theoretical simulation).

6. Conclusion

The results of our simulation studies show that small treated samples require their own guidelines for a successful removal of selection bias, particularly when the number of observed covariates is relatively large in comparison to the the number of treated units. Studies with small treated samples primarily require a control group that consists of significantly more units. The required sample size ratio depends on: (i) the size of a treated group; (ii) the number of observed covariates; and (iii) the level of initial selection bias, i.e., the covariate imbalances between treated and control units. The smaller the treated group, the more observed covariates and the larger the initial selection bias, the bigger the control group must be to remove selection bias.

The results show that the influence of the number of observed covariates ($p \epsilon \{10, 15, 20, 30\}$), on the minimum required group ratio is particularly strong for very small treated samples with $n_t < 25$, whereas negligible for treated samples with $n_t > 100$. The reason that the number of observed covariates plays such an important role for small treated samples has to do with the estimation of PSs. The more observed covariates we have and the smaller the overall sample is, the harder it is to estimate PSs which act as effective balancing scores in the process of removing covariate imbalances.

Additionally, our simulation results show that the choice of a matching algorithm, i.e., *greedy* versus *optimal* matching, matters more with small than with moderately large treated samples. When *optimal* matching is used with very small treated samples, the required minimum group ratio tends to be slightly smaller than when *greedy* matching is used. Also, the treatment effect's standard errors are smaller when *optimal* matching is used, but again only with very small treated samples. However, none of these differences are statistically significant in our study. According to Rosenbaum (2002) the *optimal* algorithm performs better when group ratios are larger due to a bigger competition among the control units. Very small treated samples require larger group ratios, thus, the *optimal* algorithm is a better choice.

The Lalonde simulation results are consistent with our theoretical simulation results, despite the fact that half of the observed covariates of Lalonde's data are discrete and the other half continuous, whereas our theoretical simulation used only continuous covariates. Based

on these results, the type of a covariate (discrete or continuous) does not play a major role. Furthermore, our minimum required group ratios for studies with moderately large treated samples are also consistent with the results obtained by Rubin and Thomas (1996).

Although the focus of our simulation studies were small treated samples, we obtained useful insights also for studies with moderately large treated samples. For example, whenever the number of observed covariates is large relative to the number of units in a treated sample, there will be a demand for a larger group ratio, R. Such demand will be even greater if the initial selection bias is large.

The simulation results in this paper can be taken as general guidelines for designing data collection strategies as also to deepen understanding on how PS matching behaves when attempting to make groups statistically comparable. The insights can be used when thinking about possible combinations of n_t, R and p to remove selection bias. We encourage practitioners to understand the interrelatedness of these factors when preprocessing data to remove selection bias, and always aim for a bigger sample, if possible, because each data set is unique in its own way.

One should never exclude an important observed covariate just because otherwise statistically comparable groups cannot be obtained. This is particularly important when one's aim is to estimate causal effects. It is important to keep in mind that matching methods remove selection bias only with respect to observed covariates that are included in the PS model. If some important covariates are not observed, it is impossible to claim that the entire selection bias was removed and that the effect estimates warrant causal interpretations. Let us emphasize that even when statistically comparable groups on observed covariates are obtained, causal interpretations of treatment effects are warranted only when causal assumptions (as listed in footnote 1) are met. If uncertainty about unobserved confounders remains, sensitivity analyses that probe the effect estimates robustness to unobserved confounding should be carried out. If one believes that the causal assumptions are violated, the estimated effect should be interpreted as a conditional association instead of a causal effect.

Future research should investigate the following: (i) The role of the number of discrete versus continuous covariates included in the estimation of PSs and its impact on removing selection bias. (ii) We used only a continuous outcome variable, thus further research could investigate a discrete outcome variable. (iii) Our scenarios considered only estimation of constant effects, a scenario where effect heterogeneity is present could also be of interest. (iv) We used only a linear model to generate the outcome variable. Future research could investigate also non-linear relations between the covariates and the outcome.

Appendix A: Tables 8-11

TABLE 8

Minimum required group ratios for investigated treated samples, n_t, initial biases, IB, two sets of observed covariates, $p = 10$ and $p = 15$, and the greedy matching algorithm.

$p = 10$						$p = 15$					
n_t	R^*	RB	VR	$99\%CI$	SE	n_t	R^*	RB	VR	$99\%CI$	SE
$IB = 0.5$						**$IB = 0.5$**					
8	13	0.149	1.57	[-0.08,0.03]	0.020	8	22	0.142	1.47	[-0.03,0.07]	0.020
10	10	0.144	1.46	[-0.03,0.06]	0.017	10	15	0.149	1.51	[-0.02,0.06]	0.017
15	7	0.123	1.38	[-0.02,0.05]	0.014	15	9	0.139	1.43	[-0.02,0.05]	0.014
20	5	0.138	1.40	[-0.03,0.04]	0.013	20	7	0.125	1.37	[-0.01,0.05]	0.012
25	5	0.107	1.31	[-0.03,0.02]	0.010	25	6	0.115	1.33	[-0.01,0.04]	0.011
30	4	0.128	1.35	[-0.03,0.02]	0.010	30	5	0.122	1.35	[-0.01,0.04]	0.010
50	3	0.137	1.35	[-0.01,0.03]	0.008	50	4	0.101	1.28	[-0.01,0.02]	0.008
100	3	0.086	1.24	[-0.01,0.02]	0.005	100	3	0.102	1.28	[-0.01,0.02]	0.005
200	3	0.061	1.18	[-0.01,0.01]	0.003	200	3	0.066	1.19	[-0.01,0.01]	0.004
500	2	0.140	1.33	[-0.01,0.01]	0.003	500	2	0.148	1.35	[-0.00,0.01]	0.003
$IB = 1.0$						**$IB = 1.0$**					
8	19	0.149	1.53	[-0.04,0.06]	0.019	8	32	0.149	1.53	[-0.07,0.03]	0.020
10	15	0.137	1.47	[-0.01,0.08]	0.017	10	23	0.144	1.51	[-0.06,0.03]	0.017
15	10	0.137	1.42	[-0.02,0.05]	0.014	15	13	0.142	1.46	[-0.04,0.03]	0.014
20	8	0.131	1.41	[-0.04,0.02]	0.011	20	10	0.137	1.42	[-0.04,0.02]	0.012
25	7	0.130	1.39	[-0.03,0.02]	0.010	25	8	0.143	1.43	[-0.04,0.01]	0.011
30	6	0.136	1.40	[-0.01,0.04]	0.009	30	7	0.139	1.41	[-0.02,0.04]	0.009
50	5	0.131	1.38	[-0.02,0.02]	0.008	50	6	0.114	1.34	[-0.03,0.02]	0.007
100	4	0.138	1.39	[-0.01,0.02]	0.005	100	5	0.101	1.30	[-0.01,0.01]	0.005
200	4	0.112	1.33	[-0.01,0.01]	0.004	200	4	0.119	1.34	[-0.02,0.01]	0.004
500	4	0.101	1.30	[-0.00,0.01]	0.002	500	4	0.104	1.30	[-0.02,0.00]	0.002
$IB = 1.5$						**$IB = 1.5$**					
8	27	0.149	1.54	[-0.01,0.09]	0.020	8	47	0.149	1.52	[-0.08,0.02]	0.020
10	22	0.146	1.50	[-0.04,0.05]	0.017	10	33	0.149	1.52	[-0.06,0.03]	0.017
15	14	0.149	1.51	[-0.03,0.04]	0.014	15	20	0.149	1.50	[-0.03,0.04]	0.014
20	12	0.137	1.45	[-0.04,0.03]	0.012	20	15	0.149	1.49	[-0.03,0.03]	0.012
25	10	0.141	1.46	[-0.01,0.05]	0.011	25	13	0.138	1.45	[-0.03,0.03]	0.010
30	9	0.142	1.44	[-0.02,0.03]	0.010	30	11	0.143	1.46	[-0.04,0.01]	0.010
50	8	0.124	1.39	[-0.03,0.02]	0.007	50	9	0.130	1.41	[-0.01,0.02]	0.007
100	7	0.149	1.45	[-0.03,0.01]	0.005	100	7	0.135	1.41	[-0.01,0.02]	0.005
200	6	0.134	1.41	[-0.01,0.01]	0.004	200	6	0.140	1.42	[-0.00,0.02]	0.004
500	6	0.121	1.38	[-0.00,0.01]	0.002	500	6	0.127	1.39	[-0.00,0.01]	0.002

TABLE 9

Minimum required group ratios for investigated treated samples, n_t, initial biases, IB, two sets of observed covariate, $p = 20$ and $p = 30$, and the greedy matching algorithm method.

$p = 20$						$p = 30$					
n_t	R^*	RB	VR	$99\%CI$	SE	n_t	R^*	RB	VR	$99\%CI$	SE
$IB = 0.5$						**$IB = 0.5$**					
8	33	0.149	1.50	[-0.10,0.01]	0.026	8[a]	58	0.149	1.43	[-0.09,0.04]	0.049
10	23	0.144	1.46	[-0.06,0.04]	0.022	10[a]	45	0.149	1.46	[-0.04,0.07]	0.033
15	13	0.147	1.45	[-0.04,0.03]	0.017	15	22	0.145	1.46	[-0.07,0.00]	0.023
20	9	0.134	1.39	[-0.02,0.04]	0.015	20	14	0.139	1.41	[-0.03,0.03]	0.019
25	7	0.134	1.39	[-0.02,0.03]	0.013	25	10	0.141	1.41	[-0.03,0.03]	0.018
30	6	0.130	1.37	[-0.02,0.04]	0.011	30	8	0.142	1.41	[-0.03,0.02]	0.016
50	4	0.136	1.37	[-0.02,0.02]	0.009	50	5	0.133	1.37	[-0.02,0.01]	0.011
100	3	0.118	1.32	[-0.02,0.01]	0.006	100	4	0.087	1.25	[-0.01,0.02]	0.007
200	3	0.076	1.22	[-0.00,0.01]	0.004	200	3	0.093	1.26	[-0.00,0.01]	0.005
500	3	0.051	1.16	[-0.00,0.00]	0.002	500	3	0.058	1.18	[-0.00,0.01]	0.003
$IB = 1.0$						**$IB = 1.0$**					
8	45	0.149	1.47	[-0.05,0.06]	0.021	8[a]	79	0.157	1.45	[-0.07,0.07]	0.027
10	32	0.143	1.48	[-0.04,0.05]	0.018	10[a]	65	0.140	1.39	[-0.05,0.04]	0.019
15	19	0.137	1.44	[-0.04,0.04]	0.014	15	31	0.148	1.47	[-0.01,0.07]	0.014
20	13	0.143	1.43	[-0.04,0.02]	0.012	20	20	0.145	1.45	[-0.03,0.04]	0.013
25	10	0.146	1.44	[-0.04,0.01]	0.011	25	15	0.140	1.42	[-0.02,0.04]	0.011
30	9	0.136	1.41	[-0.02,0.02]	0.010	30	12	0.141	1.43	[-0.04,0.01]	0.010
50	6	0.139	1.40	[-0.01,0.02]	0.008	50	8	0.129	1.39	[-0.01,0.03]	0.008
100	5	0.115	1.33	[-0.01,0.01]	0.005	100	5	0.145	1.42	[-0.02,0.01]	0.005
200	4	0.131	1.37	[-0.01,0.00]	0.004	200	4	0.149	1.42	[-0.00,0.01]	0.004
500	4	0.111	1.32	[-0.01,0.00]	0.002	500	4	0.113	1.33	[-0.01,0.01]	0.002
$IB = 1.5$						**$IB = 1.5$**					
8	65	0.147	1.50	[-0.06,0.05]	0.021	8[a]	98	0.148	1.42	[-0.11,0.03]	0.027
10	48	0.149	1.52	[-0.04,0.05]	0.018	10[a]	75	0.149	1.45	[-0.08,0.03]	0.020
15	27	0.144	1.46	[-0.05,0.02]	0.014	15	43	0.149	1.45	[-0.05,0.03]	0.014
20	18	0.149	1.48	[-0.02,0.04]	0.011	20	29	0.146	1.45	[-0.04,0.02]	0.012
25	15	0.144	1.45	[-0.01,0.05]	0.010	25	22	0.143	1.44	[-0.03,0.02]	0.011
30	13	0.140	1.43	[-0.02,0.03]	0.010	30	18	0.142	1.44	[-0.02,0.03]	0.010
50	9	0.141	1.43	[-0.03,0.02]	0.007	50	12	0.136	1.42	[-0.02,0.02]	0.008
100	7	0.144	1.44	[-0.02,0.01]	0.005	100	8	0.139	1.42	[-0.01,0.02]	0.005
200	6	0.147	1.44	[-0.00,0.02]	0.004	200	7	0.127	1.39	[-0.01,0.01]	0.004
500	6	0.129	1.40	[-0.00,0.01]	0.002	500	6	0.133	1.40	[-0.01,0.01]	0.002

[a] The logistic regression used for estimating PSs resulted in extreme values of 0 and 1 for 30% of simulation replications. The minimum required group ratio for these cases should be even larger.

TABLE 10

Minimum required group ratios for investigated treated samples, n_t, initial biases, IB, two sets of observed covariate, $p = 10$ and $p = 15$, and the optimal matching algorithm.

$p = 10$						$p = 15$					
n_t	R^*	RB	VR	$99\%CI$	SE	n_t	R^*	RB	VR	$99\%CI$	SE
$IB = 0.5$						**$IB = 0.5$**					
8	13	0.141	1.56	[-0.07,0.03]	0.020	8	21	0.143	1.49	[-0.02,0.08]	0,019
10	10	0.135	1.45	[-0.03,0.06]	0.017	10	15	0.141	1.49	[-0.03,0.05]	0,016
15	6	0.143	1.44	[-0.02,0.05]	0.013	15	9	0.131	1.41	[-0.05,0.04]	0,013
20	5	0.127	1.38	[-0.02,0.04]	0.012	20	7	0.117	1.35	[-0.02,0.04]	0,011
25	5	0.147	1.42	[-0.02,0.04]	0.010	25	5	0.148	1.43	[-0.02,0.03]	0,010
30	4	0.117	1.34	[-0.02,0.03]	0.009	30	5	0.113	1.33	[-0.02,0.03]	0,009
50	3	0.127	1.34	[-0.01,0.03]	0.007	50	4	0.093	1.27	[-0.02,0.02]	0,007
100	3	0.079	1.23	[-0.01,0.03]	0.005	100	3	0.095	1.27	[-0.01,0.01]	0,005
200	3	0.056	1.17	[-0.01,0.00]	0.003	200	3	0.062	1.19	[-0.01,0.01]	0,003
500	2	0.139	1.33	[-0.00,0.01]	0.002	500	2	0.147	1.35	[-0.00,0.01]	0,002
$IB = 1.0$						**$IB = 1.0$**					
8	18	0.148	1.54	[-0.01,0.08]	0.018	8	32	0.146	1.50	[-0.07,0.03]	0.023
10	14	0.146	1.48	[-0.05,0.04]	0.017	10	22	0.146	1.52	[-0.07,0.01]	0.020
15	9	0.148	1.46	[-0.01,0.06]	0.013	15	13	0.136	1.44	[-0.05,0.01]	0.016
20	8	0.124	1.39	[-0.01,0.04]	0.010	20	10	0.131	1.40	[-0.04,0.02]	0.013
25	7	0.122	1.38	[-0.01,0.04]	0.010	25	8	0.136	1.42	[-0.03,0.02]	0.012
30	6	0.129	1.39	[-0.02,0.04]	0.008	30	7	0.132	1.40	[-0.02,0.02]	0.011
50	5	0.125	1.37	[-0.01,0.02]	0.007	50	5	0.149	1.43	[-0.02,0.02]	0.008
100	4	0.135	1.39	[-0.02,0.02]	0.005	100	4	0.148	1.42	[-0.02,0.02]	0.005
200	4	0.110	1.32	[-0.01,0.00]	0.003	200	4	0.118	1.34	[-0.01,0.01]	0.004
500	4	0.101	1.30	[-0.00,0.01]	0.002	500	4	0.103	1.30	[-0.01,0.00]	0.002
$IB = 1.5$						**$IB = 1.5$**					
8	27	0.147	1.52	[-0.02,0.08]	0.019	8	47	0.149	1.48	[-0.06,0.05]	0.020
10	21	0.147	1.52	[-0.07,0.01]	0.016	10	33	0.147	1.49	[-0.05,0.04]	0.017
15	14	0.144	1.49	[-0.03,0.04]	0.013	15	20	0.146	1.48	[-0.03,0.04]	0.013
20	12	0.131	1.44	[-0.02,0.04]	0.011	20	15	0.146	1.47	[-0.03,0.03]	0.011
25	10	0.136	1.45	[-0.01,0.04]	0.010	25	12	0.149	1.48	[-0.03,0.03]	0.010
30	9	0.137	1.43	[-0.02,0.01]	0.009	30	11	0.139	1.45	[-0.02,0.03]	0.009
50	8	0.120	1.38	[-0.02,0.01]	0.007	50	8	0.149	1.47	[-0.02,0.02]	0.007
100	6	0.148	1.44	[-0.02,0.02]	0.005	100	7	0.133	1.41	[-0.01,0.02]	0.005
200	6	0.133	1.41	[-0.02,0.00]	0.004	200	6	0.140	1.42	[-0.01,0.01]	0.003
500	6	0.121	1.38	[-0.00,0.01]	0.002	500	6	0.126	1.39	[-0.01,0.01]	0.002

TABLE 11

Minimum required group ratios for investigated treated samples, n_t, initial biases, IB, two sets of observed covariate, $p = 20$ and $p = 30$, and the optimal matching algorithm method.

$p = 20$						$p = 30$					
n_t	R^*	RB	VR	$99\%CI$	SE	n_t	R^*	RB	VR	$99\%CI$	SE
$IB = 0.5$						**$IB = 0.5$**					
8	33	0.148	1.49	[-0.10,0.01]	0.021	8[a]	58	0.149	1.43	[-0.08,0.05]	0.025
10	22	0.146	1.46	[-0.06,0.03]	0.017	10[a]	45	0.149	1.46	[-0.05,0.05]	0.019
15	13	0.142	1.43	[-0.03,0.04]	0.013	15	22	0.145	1.46	[-0.06,0.01]	0.013
20	9	0.127	1.38	[-0.03,0.03]	0.012	20	14	0.139	1.41	[-0.03,0.03]	0.011
25	7	0.126	1.38	[-0.03,0.03]	0.011	25	10	0.141	1.41	[-0.03,0.03]	0.010
30	6	0.122	1.36	[-0.03,0.02]	0.010	30	8	0.142	1.41	[-0.02,0.03]	0.009
50	4	0.130	1.36	[-0.02,0.02]	0.007	50	5	0.133	1.37	[-0.02,0.01]	0.007
100	3	0.113	1.31	[-0.02,0.01]	0.005	100	4	0.087	1.25	[-0.01,0.01]	0.005
200	3	0.073	1.21	[-0.01,0.01]	0.003	200	3	0.093	1.26	[-0.00,0.01]	0.004
500	3	0.050	1.15	[-0.00,0.01]	0.002	500	3	0.058	1.18	[-0.00,0.01]	0.002
$IB = 1.0$						**$IB = 1.0$**					
8	45	0.149	1.43	[-0.06,0.04]	0.021	8[a]	79	0.157	1.45	[-0.07,0.07]	0.028
10	31	0.148	1.48	[-0.04,0.05]	0.017	10[a]	64	0.140	1.39	[-0.05,0.04]	0.019
15	18	0.145	1.44	[-0.03,0.03]	0.013	15	31	0.148	1.47	[-0.02,0.05]	0.014
20	13	0.138	1.41	[-0.03,0.03]	0.011	20	19	0.145	1.45	[-0.02,0.04]	0.012
25	10	0.140	1.42	[-0.04,0.02]	0.010	25	14	0.140	1.42	[-0.03,0.03]	0.010
30	9	0.131	1.40	[-0.04,0.01]	0.009	30	11	0.141	1.43	[-0.03,0.02]	0.009
50	6	0.134	1.39	[-0.02,0.03]	0.007	50	8	0.129	1.39	[-0.02,0.02]	0.007
100	5	0.112	1.33	[-0.02,0.02]	0.005	100	5	0.145	1.42	[-0.01,0.01]	0.005
200	4	0.129	1.37	[-0.01,0.01]	0.003	200	4	0.149	1.42	[-0.01,0.01]	0.004
500	4	0.110	1.32	[-0.01,0.01]	0.002	500	4	0.113	1.33	[-0.00,0.01]	0.002
$IB = 1.5$						**$IB = 1.5$**					
8	65	0.149	1.45	[-0.06,0.11]	0.021	8[a]	98	0.148	1.41	[-0.10,0.04]	0.019
10	46	0.147	1.47	[-0.08,0.06]	0.017	10[a]	81	0.147	1.38	[-0.10,0.01]	0.013
15	26	0.148	1.46	[-0.04,0.06]	0.013	15	44	0.145	1.42	[-0.04,0.03]	0.012
20	18	0.147	1.47	[-0.03,0.05]	0.011	20	29	0.149	1.44	[-0.05,0.01]	0.011
25	15	0.141	1.44	[-0.04,0.04]	0.010	25	21	0.140	1.42	[-0.02,0.03]	0.009
30	13	0.137	1.42	[-0.03,0.04]	0.009	30	18	0.133	1.41	[-0.02,0.03]	0.007
50	9	0.138	1.42	[-0.01,0.04]	0.007	50	12	0.138	1.42	[-0.02,0.02]	0.005
100	7	0.143	1.43	[-0.02,0.02]	0.005	100	8	0.127	1.39	[-0.01,0.01]	0.003
200	6	0.146	1.44	[-0.01,0.01]	0.003	200	7	0.132	1.40	[-0.01,0.01]	0.002
500	6	0.129	1.40	[-0.01,0.01]	0.002	500	6	0.148	1.31	[-0.01,0.00]	0.019

[a] Approximately 30% of simulation replications where logistic regression for estimating propensity scores results in extreme values of zero and one, thus, the minimum required group ratio in these cases should be even larger.

Appendix B: Derivation of $R^2_{Y,\underline{X}}$

B.1. Derivation of Q - the numerator of $R^2_{Y,\underline{X}}$

The Q denotes the covariance between the outcome variable and the linear combination of observed covariates: $Q = Cov(Y, \beta)$. With normally distributed observed covariates, X_i, $\epsilon \sim N(0, 1)$, and the outcome variable, $Y = \sum \beta X_i + \epsilon$, it follows that $Y = \beta \sum X_i + \epsilon$. Accordingly, the covariance structure between a linear combination of X_i, $\underline{X} = \beta \sum X_i$, and Y is:

$$\begin{aligned} Cov(Y, \beta \sum X_i) &= Cov(\beta \sum X_i + \epsilon, \beta \sum X_i) \\ &= Cov(\beta \sum X_i, \beta \sum X_i) + Cov(\epsilon, \beta \sum X_i) \\ &= \beta^2 Cov(\sum X_i, \sum X_i) + 0 \end{aligned} \tag{1}$$

$Cov(\epsilon, \beta \sum X_i) = 0$ due to the independence of the error term. But because the observed covariates are independently normally distributed, $X_i \sim N(0, 1)$, it follows:

$$\begin{aligned} Cov(\sum X_i, \sum X_i) &= Var(\sum Xi) \\ &= Var(\sum X_i) \\ &= \sum Var(X_i) \\ &= p \end{aligned} \tag{2}$$

As a result:

$$\begin{aligned} Cov(Y, \beta \sum X_i) &= \beta^2 p \\ &= Q \end{aligned} \tag{3}$$

B.2. Derivation of $\sqrt{(Q+1)Q}$ *- the denominator of* $R^2_{Y,\underline{X}}$

The denominator denotes the square root variance of the outcome variable, $Var(Y)$ times the variance of $\beta \sum X$, $Var(\sum X)$. The $1 + Q$ is hence derived from $Var(Y)$ as follows:

$$\begin{aligned} Var(Y) &= Var(\beta \sum X_i + \epsilon) \\ &= \beta^2 Var(\sum X_i) + Var(\epsilon) \\ &= \beta^2 \sum Var(X_i) + 1 \\ &= \beta^2 p + 1 \\ &= Q + 1 \end{aligned} \tag{4}$$

References

Austin, P. C. (2009). Using the standardized differences to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics - Simulation and Computation* **38** 1228-1234.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* **46** 399-424.

COCHRAN, G. W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 295-313.

COCHRAN, G. W. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A* **35** 471-446.

COX, D. R. (1958). *The Planning of Experiments*. Wiley.

DEHEJIA, H. R. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics* **125** 355-364.

DEHEJIA, H. R. and WAHBA, S. (1999). Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of American Statistical Association* **94** 1053-1062.

FLURY, B. K. and RIEDWYL, H. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician* **40** 249-251.

HECKMAN, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* **47** 153-161.

HIRANO, K. and IMBENS, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* **2** 259-278.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42** 1–28.

HOLLAND, P. W. (1986). Statistic and Causal Inference. *Journal of the American Statistical Association* **81** 945-960.

HOLLAND, P. W. and RUBIN, D. B. (1988). Causal Inference in Retrospective Studies. *Evaluation Review* **12** 203-231.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 1st ed. Cambridge University Press.

KANG, J. D. and SCHAFER, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* **22** 523-539.

KELLER, B., KIM, J. and STEINER, P. M. (2015). Neural networks for propensity score estimation: Simulation results and recommendations. In *Quantitative Psychology Research* (L. A. van der Ark et al., ed.) 279-291. Springer International Publishing.

LALONDE, J. R. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review* **76** 604-620.

LUELLEN, J. K. (2007). *A Comparison of Propensity Score Estimation and Adjustment Methods on Simulated Data - unpublished doctoral dissertation*. University of Memphis.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* **9** 403-425.

POHL, S., STEINER, P. M., EISERMANN, J., SOELLNER, R. and COOK, T. D. (2009). Unbiased Causal Inference From an Observational Study: Results of a Within-Study Comparison. *Educational Evaluation and Policy Analysis* **4** 463-479.

ROSENBAUM, P. R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association* **84** 1024-1032.

ROSENBAUM, P. R. (2002). *Observational Studies*. Springer.

ROSENBAUM, P. R. and RUBIN, D. B. (1983a). The Central Role of the Propensity Score

in Observational Studies for Causal Effect. *Biometrika* **70** 41-55.

ROSENBAUM, P. R. and RUBIN, D. B. (1983b). Assessing sensitivity to an unobserved binary covariates in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* **45** 212-218.

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39**.

RUBIN, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics* **29** 184-203.

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

RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* **2** 1-26.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6** 34-58.

RUBIN, D. B. (1979). Using Multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the Royal Statistical Society, Series B* **41** 318-328.

RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. *Journal of the American Statistical Association* **75** 318-328.

RUBIN, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* **25** 279-292.

RUBIN, D. B. (1997). Estimating causal effects from large data sets using the propensity score. *Annals of Internal Medicine* **127** 757-763.

RUBIN, D. B. (2001). Using propensity score to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2** 169-188.

RUBIN, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association* **100** 322-331.

RUBIN, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

RUBIN, D. B. (2008). For Objective Causal Inference. *The Annals of Applied Statistics* **2** 808-840.

RUBIN, D. B. and THOMAS, N. (1992). Characterizing the Effects of Matching Using Linear Propensity Score Methods with Normal Distributions. *Biometrika* **79** 797-809.

RUBIN, D. B. and THOMAS, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **52** 249-264.

SHADISH, W. R., CLARK, M. H. and STEINER, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association* **103** 1334-1343.

SIROKY, D. S. (2009). Navigating Random Forests and related advances in algorithmic modeling. *Statistics Surveys* **3** 147-163.

SMITH, J. A. and TODD, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* **91** 112-118.

SMITH, J. A. and TODD, P. E. (2005). Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators? *Journal of Econometrics* **125** 305-353.

STUART, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* **25** 1-21.

STUART, E. A. and RUBIN, D. B. (2007). Best practices in quasi-experimental design: matching methods for causal inference. In *Best practices in Quantitative Methods* (J. W. Osborne, ed.) 155-176. Sage Publications.

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0.

WAERNBAUM, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in Medicine* **31** 1572 - 1581.

WESTREICH, D., LESSLER, J. and JONSSON-FUNK, M. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and metaclassifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology* **63** 826-833.

ZHAO, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *The Review of Economics and Statistics* **86** 91-107.