Evidence from Observational Studies: What is the Role of the Matching Algorithm?

Master Thesis in Biostatistics (STA495)

by

Priska Heinz
08-912-008

supervised by

Prof. Dr. rer. nat. Ulrike Held

University of Zurich[UZH]

Zurich, March 2021

Evidence from Observational Studies: What is the Role of the Matching Algorithm?

Priska Heinz

Version March 30, 2021

Contents

Acknowledgements

First and foremost, I would like to thank my supervisor Ulrike Held. I am really grateful that she gave me the opportunity to do my master thesis on this interesting topic and that I could analyze real clinical data. She always supported me and gave a lot of her time to discuss and improve my work. Special thanks go to Pedro David Wendel Garcia from the University Hospital Zurich for letting me analyze his data and for answering my questions about it.

I also want to thank Eva Furrer and all the professors involved in the Master Program in Biostatistics. Furthermore, I would like to thank my colleagues from the Master Program. I really liked the time together with them, even though we could only see us in person in the first third of our studies.

Many thanks go to my friends and to my colleagues in the pharmacy who enrich my life besides statistics. Corinne and Marie-Line motivated me in the very first steps for my master studies. Also I am deeply grateful for having Sonja by my side. Finally, I would like to thank my mother for her ongoing support as well as making my studies also financially feasible.

Priska Heinz
March 2021

Chapter 1

Introduction

1.1 Observational studies

The gold standard to investigate the efficacy of a specific treatment in humans are *randomized controlled trials*. The goal of randomization is to obtain groups of patients which differ only by their exposure to the treatment (Silverman, 2009). This maximizes internal validity, a measure for the correctness of the estimand in the study population. In an ideal (but impossible) study we would compare the outcome of the same individual with and without treatment at the same time point. In randomized controlled trials the effect of a treatment on a well defined study population can be examined. The application of the treatment is dictated in detail and there is strong motivation for compliance. But randomized controlled studies have also limitations. Because of highly selected patients, it may be hard to draw conclusions for a general population. There are situations where an experimental study may be unnecessary, inappropriate, impossible, or inadequate (Black, 1996). As an example, clinical trials are inappropriate to examine rare and long-term outcomes, since resources and feasibility limit the number of included patients and the duration of an experiment. Randomized controlled trials can be impossible because of ethical reasons. No one could randomize patients to organ transplantation versus medical management and people could not be forced to expose themselves to toxins like cigarette smoke.

In circumstances where randomized experiments are not suitable, *observational studies* are often a good alternative. Subjects are studied prospectively or retrospectively and patient characteristics and outcomes are reported. Observational studies achieve a smaller internal validity but in exchange a better external validity compared to randomized controlled trials. The external validity is a measure for how good an estimate of a study corresponds to the truth in the whole population of interest. Observational studies can observe a larger and more diverse population in real clinical settings and for a longer time (Silverman, 2009). Sometimes observational studies or data from registries identify questions and outcomes which can be assessed further by a clinical experiment. Additionally, they are used in later stages of drug development to get information about rare side effects and effectiveness in different patient subgroups, after randomized controlled trials have shown a treatment effect in a well defined patient collective that typically excludes patients with many comorbidities or at higher ages.

To be able to estimate the true treatment effect, we try to find groups of treated and control patients that are very similar (Stuart, 2010). This thesis focuses on the case with two treatment groups. In observational studies we do not have control over the treatment allocation (Rosenbaum *et al.*, 2010). At best, the assignment seems random, thus being as close as possible to a randomized study, or the decision for treatment can be explained based on available information.

Confounding variables potentially influence the treatment assignment, may influence the outcome and are measured before treatment (Heinze and Jüni, 2011). The most popular strategies to reduce bias of confounding variables are regression (adjusting for confounders) and matching

(Cochrane and Rubin, 1973). Citing Rosenbaum *et al.* (2010), "adjustments for observed covariates should be simple, transparent and convincing". Methods to achieve this objective will be described on the next pages.

There are some assumptions typically made for observational studies:

- absence of unmeasured confounding, also called "ignorability" or "omitted variable bias" (Ho *et al.*, 2007),
- there is a positive probability of receiving each treatment for all combinations of covariate values, meaning that every subject could get either treatment (Stuart, 2010; Schafer and Kang, 2008),
- the stable unit treatment value assumption (SUTVA) (Rubin, 1980). It says that the outcome of an individual only depends on his or her treatment and is not influenced by the treatment allocation of other individuals.

A common critique of adjustment approaches is that it is never possible to know if there are unmeasured variables which influence the treatment and the outcome strongly. But using field knowledge, experts should be able to judge if there is a high probability for unmeasured confounders to be stronger than the observed variables. Moreover, sensitivity analysis can help to investigate on this.

1.2 Description of the motivating example for the simulation study

Cytosorb is a hemoadsorption device which can be used in clinical conditions with high cytokine levels (Ankawi *et al.*, 2019). There is experience in the use of Cytosorb in sepsis, cardiac surgery, and drug removal. Cytosorb cartridges contain biocompatible polystyrene divinylbenzene copolymer beads that are capable of removing a wide range of molecules from the blood: pro- and anti-inflammatory cytokines, bilirubin, myoglobin, exotoxins, and drugs (Poli *et al.*, 2019).

Ankawi *et al.* (2019) summarized the evidence of Cytosorb in septic patients. The rational of this therapy is to restore a balanced proinflammatory and anti-inflammatory mediators' response. The use of Cytosorb seems to decrease interleukin-6 levels, but there is very little evidence for reduced mortality. Ankawi *et al.* (2019) concluded that the utilization of Cytosorb is safe and has potential benefits. Lacking international validated protocols or guidelines, up to now the decision to employ hemoadsorption remains fully at the discretion of the treating clinician. So further studies are needed to decide who could really benefit from this therapy and which patients have to be monitored carefully during Cytosorb application.

In this thesis we analyze data from patients with therapy refractory septic shock, treated by Cytosorb at the University Hospital Zurich additionally to standard of care or by standard of care only. Moreover, our simulation study is based on this data base.

1.3 Application to a clinical example: lumbar spinal stenosis outcome study

In order to compare the results from the simulation study to real data, we reanalyze data of patients with degenerative lumbar spinal stenosis (Held *et al.*, 2019). This diagnosis describes a substantial narrowing of the spinal canal, which causes buttock or lower extremity pain with or without low back pain (Steurer *et al.*, 2010). The effect of a nonsurgical treatment with drug therapy and physiotherapy was compared to decompression surgery by the influence on quality of life as well as on improvement of symptoms and function at a 12-month follow-up.

Chapter 2

Methods

2.1 Matching

The term *matching* indicates any method used to make treatment groups similar in regard to their covariate distributions (Stuart, 2010). Its goal is to select suitable control patients for the treated individuals. Matching can be used in the beginning of a study to select similar groups, but in this work we will focus on its application after data accumulation facilitating non-biased analysis.

Stuart (2010) described in detail the 4 steps of matching:

1. select covariates to be included and choose a distance measure,
2. implement a matching algorithm,
3. check the quality of the matching by balance measurements,
4. estimate the treatment effect by analyzing the matched data.

In the end, these steps should be complemented by a sensitivity analysis (Caliendo and Kopeinig, 2008). However, in this thesis we confine us to the 4 steps mentioned above.

2.1.1 Distance measures

By the *distance* D_{ij} the similarity of two individuals i and j in respect of their considered covariates $\mathbf{X}$ is determined. Common distance measures are (Stuart, 2010):

- **Exact:** Covariate values correspond exactly. This method is often appropriate for factorial variables with only a few levels.

  $$D_{ij} = \begin{cases} 0, & \text{if } \mathbf{X}_i = \mathbf{X}_j \\ \infty, & \mathbf{X}_i \neq \mathbf{X}_j. \end{cases}$$

- **Absolute difference** in value (only for 1 covariate) (Rubin, 1973).
- **Euclidean distance:** $D_{ij} = ||\mathbf{X}_i - \mathbf{X}_j||$ (O'neill, 2006).
- **Mahalanobis distance:**

  $D_{ij} = (\mathbf{X}_i - \mathbf{X}_j)^\intercal \mathbf{\Sigma}^{-1} (\mathbf{X}_i - \mathbf{X}_j)$,

  where $\mathbf{\Sigma}$ is the variance covariance matrix. The Mahalanobis distance takes the correlations between covariates into account (Gu and Rosenbaum, 1993; Rosenbaum *et al.*, 2010).

- **Propensity score:** the probability of receiving the treatment conditional on patients characteristics (Rosenbaum and Rubin, 1983). It is a balancing score, yielding balanced distributions in the two treatment groups of all considered covariates when matching is done by minimizing the propensity score difference within matched pairs.
- **Linear propensity score**: $\log \frac{\text{propensity score}}{1-\text{propensity score}}$ (Rosenbaum and Rubin, 1985).
- **Prognostic score:** the predicted response when getting the control treatment (Stuart, 2010; Hansen, 2008).

The *propensity score* predicts the probability of receiving the treatment $T = 1$ conditional on the considered covariates, $\text{ps}(\mathbf{X}) = P(T = 1|\mathbf{X})$ (Rosenbaum and Rubin, 1983). It can be calculated by a logistic model or alternatively by a probit model or by non-parametric methods (Stuart, 2010). In contrast to outcome models, the amount of covariates to be included in the propensity score model is not limited by the number of outcome events (Heinze and Jüni, 2011; Glynn *et al.*, 2006). The propensity score is able to better handle not-normally distributed variables than the Mahalanobis distance. If a covariate distribution has long tails, its variance is larger and thus the Mahalanobis distance will tend to ignore that variable in the matching process. In the same way, Mahalanobis distance gives more attention to binary variables with uneven categories than to variables with a 50:50 distribution (Rosenbaum, 2020). If the assumption of *strongly ignorable treatment assignment* holds, analysis using the propensity score can provide unbiased estimates of the treatment effect (Shadish and Steiner, 2010). This assumption means that the treatment assignment and the outcome are independent conditional on the observed covariates. It is given if all confounders (related to the treatment assignment and the outcome) are included in the propensity score model as covariates and if there are no unmeasured confounders (Lee and Little, 2017). As there exists no method to test this assumption empirically, it is important to measure exactly potentially influential values and to include all important variables related to the selection process (Shadish and Steiner, 2010). A sensitivity analysis can help to support the assumption that all relevant variables are considered (Caliendo and Kopeinig, 2008).

Rubin (2001) suggested the use of the *linear propensity score*. The logit of the propensity score tends to have a more symmetric distribution and is more appropriate to assess the efficacy of linear modeling adjustments.

The selection of controls can be restricted by a *caliper*. A caliper is the width of the distance measure which is maximally allowed (Rosenbaum *et al.*, 2010). Individuals having a larger distance between each other than the caliper width can not be matched, resulting in closer pairs. Rosenbaum *et al.* (2010) suggested a caliper width of 20% of the standard deviation of the propensity score. Austin (2011b) recommended to match on the logit of the propensity score and to use a caliper of 20% of the standard deviation thereof.

The propensity score can not only be used for matching, but as well for stratification, or it can be included as a covariate in regression analysis (D'Agostino Jr, 1998). Including the propensity score in a regression assumes a linear relationship of the propensity score with the outcome (Glynn *et al.*, 2006) and does not clearly separate the design part from the analysis (Harder *et al.*, 2010). These are disadvantages compared to the other methods mentioned. Stratification into subclasses was shown to be less precise than matching on the propensity score (Austin *et al.*, 2007). Yet another application of propensity scores is *inverse probability of received treatment weighting* (IPTW) (Heinze and Jüni, 2011; Desai and Franklin, 2019). There, treated subjects get a weight of the inverse of their propensity score and the control patients get a weight of $\frac{1}{1-\text{propensity score}}$. In the case of IPTW the correct specification of the propensity score is important. An advantage of IPTW is that all information is used without discarding any individuals from the analysis. It is similar to full matching which is described later.

There are advancements by combining strategies like coarsened exact matching and methods yielding *fine balance* (see Section 2.3 for an explanation thereof). In the case of *coarsened exact matching* the covariate values are coarsened into strata and then exact matching is performed

respecting these strata (Iacus *et al.*, 2012). Furthermore, combinations of distance measures can be used, as for example Mahalanobis matching within propensity score calipers (Stuart, 2010).

2.1.2 Matching algorithms

After deciding which distance measure to use, the researcher has the choice of different matching algorithms that have been developed over the last decades.

The oldest and probably easiest to understand approach is *greedy 1:1 nearest neighbor matching*. There, one treated individual after the other is considered and the most similar control is selected to each. The result of this process depends on the order in which the subjects are looked at.

Optimal matching solves this problem by looking at the whole data set and guarantees to find the best available 1:1 matching (pair matching). It is an optimization problem that can be solved by network flow theory (Rosenbaum, 1989). Bertsekas (1981, 1990) called his algorithm to perform this task the "auction algorithm".

If matching *with replacement* is used, there is no difference between optimal and greedy matching, because independent of the order for every treated individual the best control is chosen. The drawback is a smaller total sample size and controls which are not independent of each other. The analysis has to consider this dependence (Stuart, 2010).

Instead of 1:1, optimal matching can also be used for *1:k matching* (Rosenbaum *et al.*, 2010). There, for every treated subject k controls are selected. This is especially useful, if many more controls than treated subjects are available. The advantage of a larger total sample size needs to be evaluated against the disadvantage of having not only the best, but the k best matched controls.

To achieve exact balance on some covariates, *exact matching* can be done. Often this is only feasible for binary variables such as gender, if at all. For example exact matching on gender and otherwise matching on propensity score can be done. This is performed by adding an infinite penalty to the propensity score of all pairs that mismatch on gender. If one would like to have exact matches, but it is not possible without discarding several observations, *near-exact matching* is an alternative. Thereby mismatching is only allowed if it is not avoidable. Practically it is done by adding a large penalty to mismatches instead of an infinite one (Rosenbaum, 2020).

Matching all treated to a control and similarly using all controls to match to a treated unit is called *full matching*. It is a stratification that makes the two treatment groups as similar as possible (Rosenbaum *et al.*, 2010). It can achieve a better balance than 1:k matching, because also strata with only 1 control but multiple treated are possible. Depending on the composition of a stratum, matching weights are given to the individuals. As an example, in a stratum consisting of 1 control and 3 treated subjects, the control gets a weight of 1 and the treated ones each get a weight of $\frac{1}{3}$. These weights must be taken into account for the balance diagnostics as well as for the outcome analysis (Rosenbaum, 1991; Hansen, 2004).

Since nowadays computational power is much higher than in the last millennium, more complicated evolutionary algorithms can be used for matching, too (Mebane Jr *et al.*, 2011). Such an approach is *genetic matching*. The R function `Matching::GenMatch` uses a generalized Mahalanobis distance metric including weights for every covariate,
$D_{ij} = (\boldsymbol{X}_i - \boldsymbol{X}_j)^\intercal (\boldsymbol{\Sigma}^{-1/2})^\intercal \boldsymbol{W} \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{X}_i - \boldsymbol{X}_j)$, where $\boldsymbol{W}$ is the diagonal weight matrix. These weights are determined by an automated search algorithm to achieve the best covariate balance (Radice *et al.*, 2012). At every round of the iteration a new group, a so called *generation*, of weight matrices is generated and the resulting balance tested. Then the next generation is produced in a way that the weight matrices tend to yield more balanced matched samples. The group size is called *population size* and should not be too small (Sekhon, 2011a). When after a defined amount of new generations no better result can be found, the algorithm is stopped and the best weight matrix and the corresponding distance metric is used to construct the final matching. In the end it is a nearest neighbor matching, just with a special distance metric. An

advantage of genetic matching is that it is not prone to misspecification of the propensity score (Radice *et al.*, 2012).

2.1.3 Balance diagnostics

After matching the achieved balance, meaning the similarity of the treatment groups, needs to be examined. If satisfactory balance is not achieved, the matching should be improved before going further towards analysis of the outcome. As long as one has not looked at the outcome, it is allowed to try as many matching strategies as required for a good balance.

Let us have a look at the variety of diagnostics to assess balance. Usually, the marginal distribution of each covariate is assessed, because it is not feasible to consider the overall, joint distribution of all considered variables (Stuart, 2010). In addition to look at the single covariates, one can also asses the balance of the distance measure, for example of the propensity score.

There are numerical diagnostics like:

- **Standardized mean difference (SMD):** The difference in means of the two groups is divided by the standard deviation. Stuart (2010) suggested to use the standard deviation in the treatment group, whereas Flury and Riedwyl (1986), Zhang *et al.* (2019) and Austin (2008) used the pooled standard deviation in the two matched groups. Stuart (2008) recommended to use the standard deviation in the full sample to have the same denominator before and after matching, even if treated individuals are discarded by the matching algorithm.

- **Variance ratios:** (Rubin, 2001) It is the ratio of the variance in the treated group to the variance in the control group. The distribution of the estimated ratio follows an F-distribution under the null hypothesis of equal variances in the two groups (Austin, 2009).

- **Five-number summary** of the covariates composed of minimum, first quartile, median, third quartile and maximum (Austin, 2009).

- **C-statistic:** the estimated area under the receiver operating characteristic (ROC) curve from a propensity score model (Franklin *et al.*, 2014). The value corresponds to the estimated probability of a treated subject to have a larger propensity score than a randomly chosen control individual (Heinze and Jüni, 2011). According to Austin (2009) this is not a reliable diagnostic.

- **Overlapping coefficient (OVL):** the overlap of two densities is estimated by kernel densities (Belitser *et al.*, 2011).

- **Kolmogorov-Smirnov distance** is the maximum vertical distance between the cumulative distribution functions of the two groups (Belitser *et al.*, 2011). Smaller values indicate better balance.

- **Lévy distance:** is the side length of the largest square (with sides parallel to the coordinate axes) that can be drawn between the two cumulative distribution curves (Belitser *et al.*, 2011).

The three last diagnostics are only useful for continuous variables. As the variance of a binary variable is only a function of the proportions of the two values that this variable can take, it is not very informative. Thus the consideration of variance ratios and of standardized mean differences does not provide more information than looking at simple mean differences of a binary variable.

It is not recommended to use t-tests and the associated p-values for the evaluation of balance (Imai *et al.*, 2008). The main problem of hypothesis tests is their dependency on sample size (Austin, 2009).

Another possibility is the use of graphic representation. **Density plots** can compare the distribution of a variable in the two groups and show changes after matching. To compare the empirical distributions in the groups, **Q-Q-plots** or plots of the **empirical cumulative distribution functions** are an option (Stuart, 2010; Austin, 2009).

Relating to the numerical five-number summary **boxplots** of the distributions in each group can be drawn for comparison (Austin, 2009).

Ahmed *et al.* (2006a,b) were about the first to show numerical diagnostics in form of a plot. Thomas E. Love became in this way the name giver of the **Love plot**. This plot shows standardized mean differences or other diagnostics for every considered variable. Differences between the original and the matched data set in respect to balance can easily be seen. It is also a good way to recognize covariates with a standardized mean difference above a certain threshold. As an example, it can be decided to use covariates with a standardized mean difference higher than 0.1 for additional adjustment in the subsequent regression (Nguyen *et al.*, 2017). Others used a threshold of 0.25 to judge if balance was achieved, while emphasizing that a smaller value might be better (Harder *et al.*, 2010). However, there is no clear opinion about such a threshold (Austin, 2009).

Instead of calculating the standardized mean difference for the whole matched sample, it can also be done strata-wise (Lee and Little, 2017; Harder *et al.*, 2010).

For all balance diagnostics there exists no general answer to the question what value is good enough. Rubin (2001) listed three conditions which are necessary for regression analyses to be trustworthy: small differences in the means of the propensity score in the two groups, a ratio of the variance of the propensity score close to one, and a ratio of the variances of the residuals of the covariates after adjustment for the propensity score close to one (values of 1/2 or 2 are too far away). These criteria could be used as well to decide if the matching was successful and whether the data is ready for analysis of the outcomes. It is recommended that only effect estimates should be reported which were calculated on balanced matched groups (Harder *et al.*, 2010).

In contrast to the analysis of the outcomes, the matching process does not face a problem of multiplicity if various options of matching are tried. It is just forbidden to look at the outcome before matching is completed (Rubin, 2007). Matching belongs to the design part of a study (Rosenbaum *et al.*, 2010). Afterwards the analysis of the outcome takes place (Stuart, 2010).

2.2 Analysis of the outcome after matching

The quickest approach is to take the matched sample and analyze the outcome by simply calculating a difference in means, odds ratio, or another appropriate measure. But Ho *et al.* (2007) recommended to apply the same analytical method to the matched data as would be done without preprocessing by matching. So regression approaches with adjustment for covariates are better suited (Rubin, 973b; Rubin and Thomas, 2000). Without that, independence of the covariates and the treatment would be assumed, what only holds for the special case of exact matching on all relevant covariates (Ho *et al.*, 2007). Adjustment for covariates in a matched sample can have the same advantages as regression adjustment in a randomized experiment. It can remove residual bias still remaining after matching, enlarge power and decrease variability (Schafer and Kang, 2008; Osborne, 2008). However, there is no general consensus on the need and the extent of covariate adjustment after matching. Successful matching cuts the association between a covariate and the treatment assignment, thus the covariate does not fulfill the criteria to be a confounder anymore (Sjölander and Greenland, 2013). Sjölander and Greenland (2013) confirmed that in the absence of additional covariates (which are not matched for), it is correct to ignore the matching variables in the analysis.

One should not forget that in small samples not too many covariates should be included in

the regression because of the risk of overfitting and of biased effect estimates (Peduzzi *et al.*, 1996; Chen *et al.*, 2016).

Another point of discussion is the question if matched data should be seen as independent or if a correlation structure has to be taken into account, and thus how variances should be calculated for propensity score methods. Austin (2011a) argued that matched subjects resemble one another more than randomly selected ones, because they have a similar propensity score value and so their covariate values originate from the same multivariate distribution. Also Abadie and Spiess (2021) said that matching produces dependence if the outcome depends on the matching variables. They suggested to use clustered standard errors, because standard error estimation without consideration of the matching process is not generally valid in the case of a misspecified regression model (Abadie and Spiess, 2021). On the other hand, Schafer and Kang (2008) did not see any emergence of correlation in matched groups, and Stuart (2010) emphasized that propensity score matching does not yield pairs of individuals sharing the same values of all covariates, but only matched treatment groups with similar covariate distributions. If one believed in correlated pairs, pairs should also be considered for balance diagnostics (Stuart, 2008).

Special methods of analysis are required after full matching, stratification, and matching with replacement. Here, weights should be used in the analysis to account for different numbers of treated and controls in specific strata (Osborne, 2008). In the case of subclassification, estimation should be done in each subclass separately and then aggregated. For example, after full matching Stuart and Green (2008) performed logistic regression with weights, but without taking into account any potential dependence.

Greifer (2020b) emphasized that robust standard errors should be used, when weights are included in the analysis. To consider both matching weights and subclass membership, cluster robust standard errors are the method of choice. In R they can be computed by the function `vcovCL` of the `sandwich` package.

2.2.1 Heteroscedasticity consistent and cluster robust standard errors

Looking at the linear regression model, $\boldsymbol{y} = \mathbf{X}\beta + \epsilon$, the ordinary least squares estimator $\hat{\beta}$ has the following variance:

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{\Omega}\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}, \tag{2.1}$$

where $\mathbf{\Omega}$ is a positive definite matrix.

An important assumption in linear regression is the one of homoscedasticity, saying that the distributions of all errors have the same variance. If one is not sure if this assumption is met, a heteroscedasticity consistent (HC) covariance matrix can be used to calculate standard errors. The HC estimator can be constructed by replacing $\mathbf{\Omega}$ in equation 2.1 by the estimator $\hat{\mathbf{\Omega}} = diag(\omega_1, \ldots, \omega_n)$ for n observations i. For the HC0 estimator, also known as the White, Eicker, Huber or simple "sandwich" estimator, one takes $\omega_i = \hat{r}_i^2$, $\hat{r}_i$ being the residuals. The HC3 estimator should perform better for small samples and is calculated by taking $\omega_i = \frac{\hat{r}_i^2}{(1-h_i)^2}$, where $h_i = \mathbf{H}_{ii}$ are the diagonal elements of the hat matrix (Long and Ervin, 2000; Zeileis, 2004).

In the case of general linear models and hence for logistic regression the sandwich estimator for the variance is:

$$\mathrm{Var}(\hat{\beta}) = \mathbf{BMB}, \tag{2.2}$$

where $\mathbf{B}$ is the "bread" $\mathbf{B} = (-L(\hat{\beta})'')^{-1}$ with $L(\hat{\beta})''$ being the second derivative of the log-likelihood with respect to the parameter β, and $\mathbf{M}$ is the "meat"
$\mathbf{M} = \mathrm{Cov}(L'(\beta_\mathbf{0})) = \sum_{i=1}^{n} g_i(Y_i|\hat{\beta})^\intercal g_i(Y_i|\hat{\beta})$ with $g_i(y|\beta) = \frac{\delta}{\delta\beta}\log f_i(y|\beta)$ (Zeileis, 2006; Freedman, 2006).

If there are clusters, it is assumed that samples within the same cluster are not independent, whereas different clusters are independent of each other. For the calculation of *cluster robust standard errors* the "meat" $\mathbf{M}$ in equation 2.2 changes to $\mathbf{M} = \sum_{j=1}^{m} \sum_{i \in c_j} g_i(Y_i|\hat{\beta})^\intercal g_i(Y_i|\hat{\beta})$, where it is first summed over the m clusters c_j, instead of summing over each individual (Freedman, 2006; Zeileis *et al.*, 2020). When each observation forms its own cluster, the cluster robust standard error simplifies to the (HC0) sandwich standard error.

Other possibilities to deal with clustered correlations are random effects model and generalized estimating equations (GEE). The advantage of robust standard errors, as described above, is that the parameter estimation by regression is not changed but only the covariance matrix is adjusted (Zeileis *et al.*, 2020). This assumes that the score function and thus the parameter estimation is correct. In this work, we decided to use the approach of sandwich standard errors, because in this way it is possible to compute all treatment effect estimates in the same way and then to look at different standard errors. This seems necessary, as not by all matching algorithms clusters are formed, but for others like full matching the resulted subclasses should be considered for inference.

2.2.2 ATE vs. ATT

We need to introduce some notation to explain the different types of treatment effect estimands. According to the Roy-Rubin model (Rubin, 1974), each individual has two potential outcomes, one if it gets the treatment ($T = 1$), and one without treatment ($T = 0$) (Caliendo and Kopeinig, 2008). Of course, only one of these potential outcomes can really be observed. Following the notation of Ho *et al.* (2007), we define the outcome of individual i getting the treatment $T = 1$ as $y_i(T = 1) = y_i(1)$ and equivalently its outcome getting the control procedure as $y_i(T = 0) = y_i(0)$. If the outcomes are considered as random variables, the mean causal effect is obtained, shown in equation (2.3).

$$\mathrm{E}[Y_i(1) - Y_i(0)] = \mu_1 - \mu_0, \tag{2.3}$$

where $\mu_1 = \mathrm{E}[Y_i(1)]$ and $\mu_0 = \mathrm{E}[Y_i(0)]$.

In practice the interest is usually in average treatment effects. The *average treatment effect* (ATE) describes the mean effect on the considered population and is defined in equation (2.4), where $\mathbf{X}_i$ represents the covariate characteristics.

$$\mathrm{ATE} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i] = \frac{1}{n} \sum_{i=1}^{n} [\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)] \tag{2.4}$$

The *average treatment effect on the treated* (ATT) (equation (2.5)) looks only at the estimated effect in the patient group which gets the treatment.

$$\mathrm{ATT} = \frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} T_i \cdot \mathrm{E}[Y_i(1) - Y_i(0)|\mathbf{X}_i] = \frac{1}{\sum_{i=1}^{n} T_i} \sum_{i=1}^{n} T_i \cdot [\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)]. \tag{2.5}$$

The ATE estimates the effect of a treatment if subjects are randomized to it (Caliendo and Kopeinig, 2008). It is the "the difference between the expected outcome if everyone was exposed and the expected outcome if everyone was unexposed" (Woodward, 2014). Applying matching yields estimates of the ATT if a match is found for each treated patient out of the control group. Conversely, stratification and inverse weighting methods estimate the ATE (Desai and Franklin, 2019).

In a randomized controlled trial the ATT and the ATE are identical (Shadish and Steiner, 2010).

If some treated individuals are discarded, the ATT of the whole population is no longer possible to estimate. In this way, the common support region of the propensity score has impact

on the population for which the result can be generalized (Thoemmes and Kim, 2011; Harder *et al.*, 2010).

If the ATE or the ATT better answers the question of a researcher can be decided by asking if it would be possible to give the treatment to all patients included in the study (Desai and Franklin, 2019). The aim of the study influences this choice as well (Lee and Little, 2017).

2.2.3 Non-collapsibility of odds ratios

In contrast to for instance risk differences, odds ratios are not collapsible. This means that even if the conditional estimates in different strata are the same, they do not in general correspond to the marginal effect estimate (Hernán *et al.*, 2011). A conditional odds ratio is always further away of the null effect (OR = 1) than the marginal odds ratio in the absence of effect modification. Whereas marginal odds ratios compare all patients with the treatment and all patients without, conditional odds ratios compare the treatment effect in patients with similar covariates (Samuel *et al.*, 2017).

Pure propensity score methods estimate the marginal effect, the unadjusted effect, without considering covariates and concentrating on the similarity of treatment groups (Martens *et al.*, 2008). In contrast, logistic regression methods with consideration of covariates model the outcome, and thus estimate a conditional treatment effect. This is subject-specific and clinicians can take it into account to make treatment decisions for individual patients (Martens *et al.*, 2008). After matching conditional treatment effects can be calculated by regression considering covariates (Austin, 2008).

2.3 Matching in R

We decided to use the R package `MatchIt` (Ho *et al.*, 2011, 2007) for our analyses. This package provides many matching algorithms and the matched data set can easily be extracted. For full matching and for optimal matching it calls automatically the function `fullmatch` of the package `optmatch` (Hansen and Klopfer, 2006). The `optmatch` package could also be used for optimal 1:1 and 1:k matching as well as for exact matching. To apply genetic matching `MatchIt` calls the package `Matching` (Sekhon, 2011b). Its function `GenMatch` uses a genetic search algorithm executed by the function `Genoud` from the package `rgenoud` (Mebane Jr *et al.*, 2011) to assign each individual a weight leading to matched groups with the best balance (Diamond and Sekhon, 2013). `Matching` could be used as well for various other matching methods, but in contrast to `MatchIt`, it performs matching imputation instead of matching in the sense of a nonparametric preprocessing for subset selection.

Matching with fine balance or with refined covariate balance constraints can be conducted with the package `rcbalance` (Pimentel, 2016). Fine balance in a variable means that the two treatment groups have the same overall distribution of this variable. It can happen that this is not possible with a given data set, for example if there are more treated females than the total number of females in the control group. In such a case near fine balance could be achieved by selecting all control females, and thus achieving the best possible balance of the variable sex. To find refined covariate balance, the covariates are prioritized and then near fine balance is searched for one after the other (Pimentel, 2016).

Another package for propensity score methods is `twang` (Ridgeway *et al.*, 2014). Its default version uses generalized boosted regression (as implemented in the package `gbm` (Greenwell *et al.*, 2020)) to estimate the propensity scores, and assigns weights to the controls. In this way the ATT can be estimated. `twang` can also be used for estimation of the ATE by weighting both treated and control subjects. This method of inverse probability weighting has some similarity to full matching. This package can as well be used for weighting more than two treatment groups (McCaffrey *et al.*, 2015).

To asses the balance in the matched groups, we used the R package `cobalt` (Greifer, 2020a). This package offers convenient functions to calculate and graphically present balance diagnostics. It can be used to evaluate results from different matching packages, like `MatchIt`, `optmatch`, `Matching`, and `twang`.

2.4 Guidelines

2.4.1 Literature about matching and propensity score analysis

In the last years some overview articles were published which compare different methods and give recommendations to researches. Stuart (2010) wrote an exhaustive review article, whereas Thoemmes and Kim (2011) focused on the application in social science and what should be reported when doing a propensity score analysis. Lee and Little (2017) provided a step-by-step guidance including R code. A comparison of different ways to estimate the propensity score as well as instructions for practitioners was made available by Harder *et al.* (2010) with focus on psychological research.

2.4.2 Reporting guidelines in observational studies

The *Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE) guidelines give advice what should be reported about an observational study in the article's title and abstract, the introduction, methods, results, and discussion sections as well as on funding (Vandenbroucke *et al.*, 2007; Von Elm *et al.*, 2014).

Yao *et al.* (2017) provided adjusted guidelines to help researches to correctly report a propensity score analysis. The most important points to mention for propensity score matching are:

- model used to estimate the propensity score,
- variable selection for propensity score model,
- matching algorithm, distance measure, matching ratio, whether sampling with or without replacement, the statistical methods for the analysis of matched data, the package used, and methods for assessing the balance between the matched groups,
- examination of assumption of propensity score analysis,
- handling of missing data,
- sample size before and after matching,
- baseline characterstics in each group before and after matching,
- estimates with confidence interval after matching and unadjusted estimates,
- discussion of remaining imbalance after matching,
- discussion of incomplete matching and potential influence of discarded observations.

2.5 Data used for the simulation study

2.5.1 Data

The data includes patients, who were admitted to the medical intensive care unit of the University Hospital Zurich between 2011 and 2018.

The whole data set consists of 250 patients. After exclusion of patients with missing data for important covariates (namely VPI, Lactate and IL-6), we used 208 patients for the analysis.

The data contains information about therapy, length of in-hospital stay and survival as well as demographic and laboratory values. The ones that were used for our analysis are shown in Table 2.1 including the explanation of special scores (Jones *et al.*, 2009; Le Gall *et al.*, 1993).

Variable	Units / levels	Description	Rationale
PAT_Nr		patient ID	subject
Filter	Cytosorb or control	treatment	treatment
Age	years	age	demographic baseline
BMI	kg/m^2	BMI = weight / height^2, body mass index	demographic baseline
Sex	female or male	sex	demographic baseline
SOFA	0 - 24	Sequential Organ Failure Assessment (SOFA) score: assesses number and severity of organ dysfunction in six organ systems (respiratory, coagulatory, liver, cardiovascular, renal, and neurologic)	validated severity score of the ICU
SAPS	0 - 163	Simplified Acute Physiology Score (SAPS II): includes 12 physiological variables, age, type of admission and three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer, and hematologic malignancy)	validated severity score of the ICU
IL6	pg/ml	interleukin-6 blood level	main target of Cytosorb therapy
VPI		vasopressor index *	value for the cumulative demand for vasopressors on a target MAP of 65 mmHg
PCT	ng/ml	procalcitonin blood level	biomarker in septic shock
Lactate	mmol/l	lactate blood level	marker for tissue perfusion and in general for microcirculation disturbances

Table 2.1: List of variables of the Cytosorb data. MAP: mean arterial pressure.
* VPI = $\frac{\text{Dobutrex}+(\text{Noradrenaline}+\text{Pitressin}\cdot 1000)\cdot 100+\text{Adrenaline}\cdot 100}{\text{Weight}} \frac{10}{\text{MAP}}$.

2.5.2 Outcome

The in-hospital mortality was used as outcome. Thus patients who died during their stay in the intensive care unit or in the normal ward are considered to have encountered the event.

2.5.3 Statistical analysis

We used five different matching algorithms with the objective of obtaining comparable treatment groups. The five matching methods were: nearest, optimal, caliper (nearest matching with a caliper of 20% of the standard deviation of the propensity score), full, and genetic matching. All were implemented with the R package `MatchIt`. The following covariates were used for matching: Age, BMI, IL6, Lactate, PCT, SAPS, Sex and SOFA. Neither interactions nor effects of higher order were included.

Descriptive statistics included absolute standardized mean differences, variance ratios, and Kolmogorov-Smirnov statistics.

The raw as well as all the matched data sets were used for logistic regression to compute the odds ratio for in-hospital mortality. The regression was once done without adjustment and in addition with covariate adjustment for the same variables which were used for matching, thus yielding a marginal and a conditional treatment effect estimate, respectively. Standard errors

were calculated as "simple" standard errors for the unmatched sample, while for the data originating from optimal, nearest, caliper, and genetic matching HC0 standard errors were determined. After full matching cluster robust standard errors were computed.

2.6 Simulation study

We wrote a protocol for the simulation study which is shown in Section A.2 and was uploaded to `https://osf.io/unbka/` on January 27, 2021. Like prior to randomized controlled studies a protocol should be made before conducting a simulation study (Burton *et al.*, 2006). For our protocol we followed the structure proposed by Burton *et al.* (2006).

We used the motivating example of the Cytosorb data as template and simulated data consisting of the treatment Filter, the outcome Death and four covariates: Sex, Age, SOFA and IL6 (interleukin-6).

We analyzed the simulated data using five different matching algorithms offered by the R package `MatchIt`: nearest, optimal, caliper, full and genetic matching. Afterwards, we compared the estimated conditional odds ratios as well as the marginal odds ratios. As a logistic model was used to simulate the treatment and the outcome, the true conditional treatment effect was determined in advance. The corresponding marginal effect was assessed by the simulation of 100'000 data sets consisting of 10'000 individuals each and the formula in equation (2.6) using the β-parameters affecting the outcome (Austin and Stafford, 2008).

$$\text{OR}_{\text{marginal}} = \frac{\overline{p}_1/(1-\overline{p}_1)}{\overline{p}_0/(1-\overline{p}_0)}, \tag{2.6}$$

$$p_{i,1} = \frac{1}{1+exp(-(\beta_0+\beta_{treat}+\beta_1\, x_{i,1}+\cdots+\beta_4\, x_{i,4}))},\ p_{i,0} = \frac{1}{1+exp(-(\beta_0+\beta_1\, x_{i,1}+\cdots+\beta_4\, x_{i,4}))},$$

To compare the different algorithms, the bias of the estimates and coverage were considered. We also had a look at the time needed for computation, the proportion of treated individuals who were discarded during matching as well as false negative and false positive results.

2.6.1 Implementation in R

The code of the programmed functions `dat_simulate` and `match_analyze` can be found in Section A.5.

2.6.2 Deviations from the protocol

We were able to stick to the protocol and no errors did occur during simulation. Only the line-plots were not drawn, because they would be too crowded to show information.

Like written in the protocol, the summary statistics of the weights generated by full matching were stored but they were not analyzed further.

Chapter 3

Results

3.1 Results of the motivating example: Cytosorb

3.1.1 Descriptive statistics

Table 3.1 shows the patient characteristics at baseline for the unmatched data. In the Appendix A.1 tables with the corresponding values after matching can be found.

In Figure 3.1 the distributions of the two treatment groups of the continuous variables are presented. The variables Lactate, IL6, PCT, and VPI seem not to be normally distributed.

3.1.2 Balance

Figure 3.2 and Figure 3.3 show Love plots comparing the different matching algorithms which were applied to the data.

Overall, all matching algorithms yielded a smaller standardized mean difference (SMD). The SMD was calculated by using the standard deviation of the full sample as the denominator. Only for the variables Sex and BMI nearest, optimal, and nearest matching with a caliper enlarged the SMD and for PCT the unmatched data had the smallest SMD of all. From the Figure 3.2 it can be seen that matching algorithms tend in particular to minimize the SMD of variables with relative large SMDs before matching. Genetic matching was able to achieve SMDs smaller than 0.1 for all variables, thus providing the best result in balance measured by SMD.

Looking at the variance ratios, for most variables the matching algorithms could improve the balance. For PCT, VPI, SAPS, and Age the values of the unmatched data was already quite close to 1 and the matching algorithms led to a worsening.

Table 3.1: Patient characteristics of the Cytosorb data. Here, IQR denotes the first and third quartiles, which are given as a range in brackets.

Variable	Level	Overall	Filter	Control	SMD
n		208	160	48	
Age (mean (SD))		61 (16)	63 (15)	57 (16)	0.362
Sex (%)	f	64 (30.8)	47 (29.4)	17 (35.4)	0.129
	m	144 (69.2)	113 (70.6)	31 (64.6)	
BMI (mean (SD))		26 (6)	26 (5)	26 (7)	0.083
SAPS (mean (SD))		63 (19)	62 (19)	68 (18)	0.307
SOFA (mean (SD))		12 (4)	12 (4)	14 (3)	0.769
Lactate (median [IQR])		3 [2, 6]	2 [2, 5]	4 [2, 8]	0.391
IL6 (median [IQR])		1369 [446, 1369]	1037 [302, 1369]	1369 [1369, 1369]	0.988
PCT (median [IQR])		11 [3, 35]	10 [3, 28]	18 [6, 56]	0.062
VPI (median [IQR])		6 [3, 12]	5 [3, 10]	10 [4, 20]	0.507

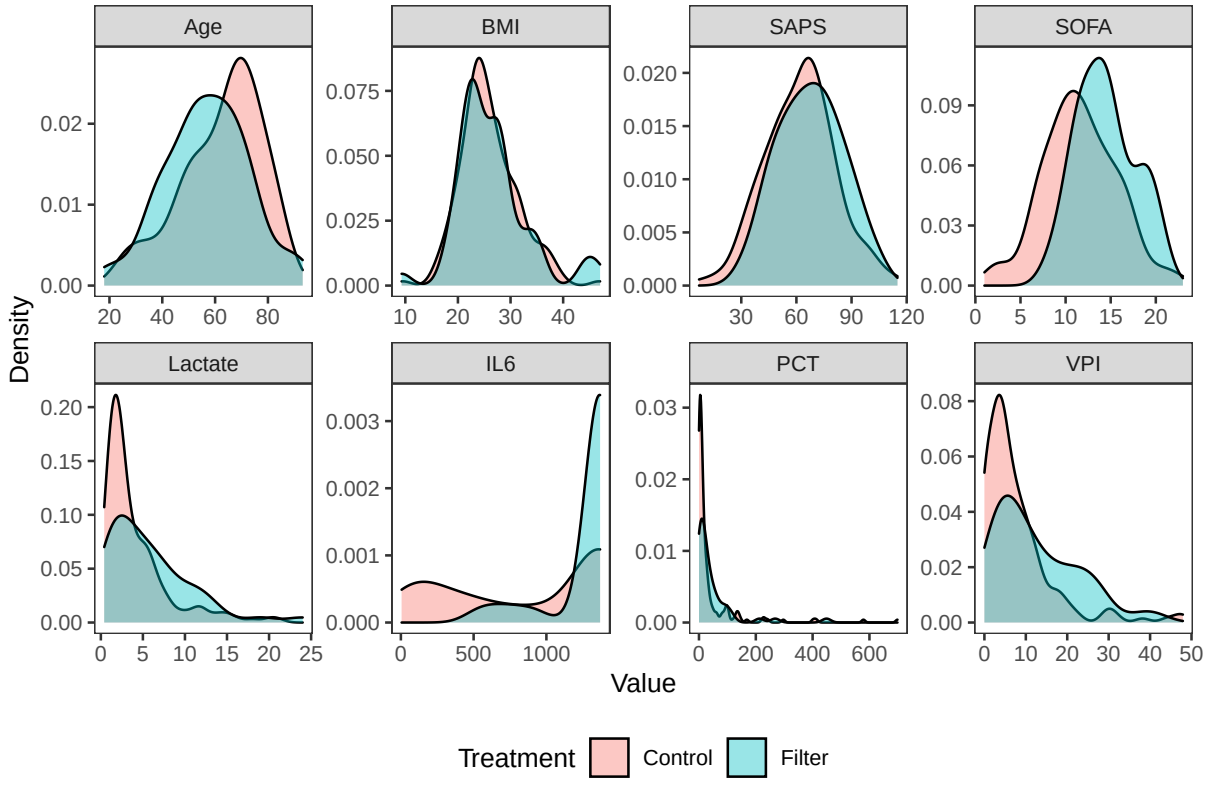

Figure 3.1: Density plots of the continuous variables comparing the control group and the group treated with the Cytosorb filter.

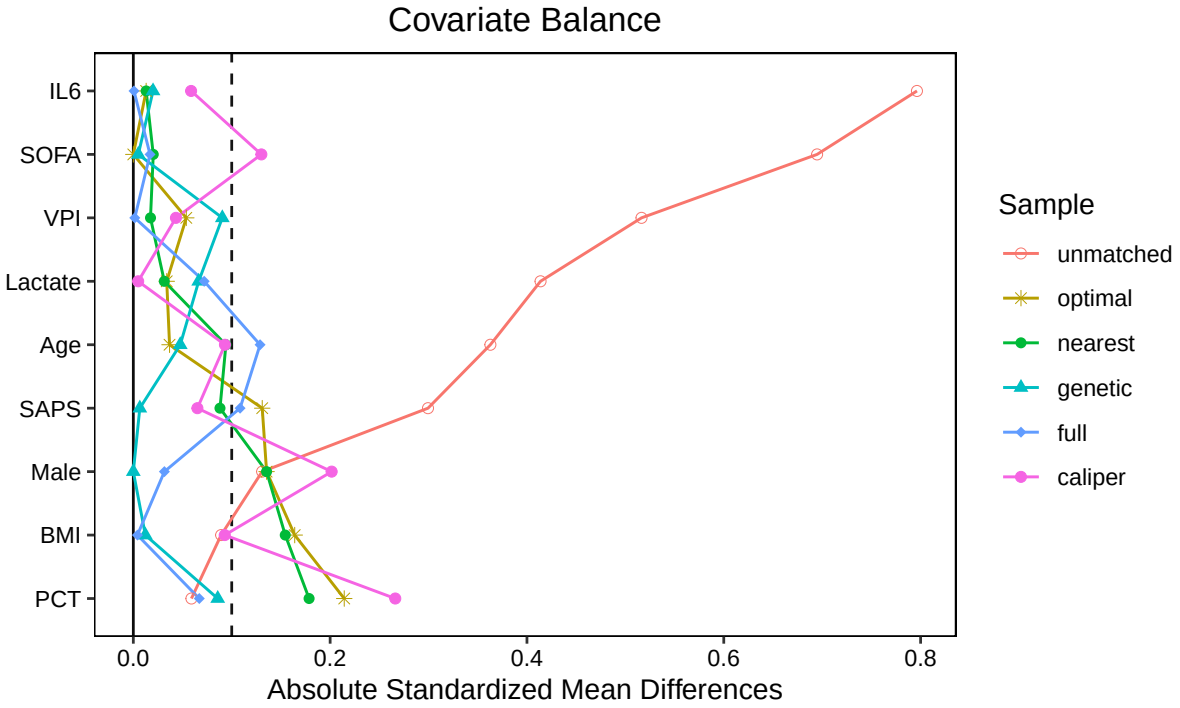

Figure 3.2: Love plot showing the absolute standardized mean differences (SMD) between the treatment groups before (unmatched) and after using 5 matching algorithms. The vertical dashed line marks a SMD of 0.1.

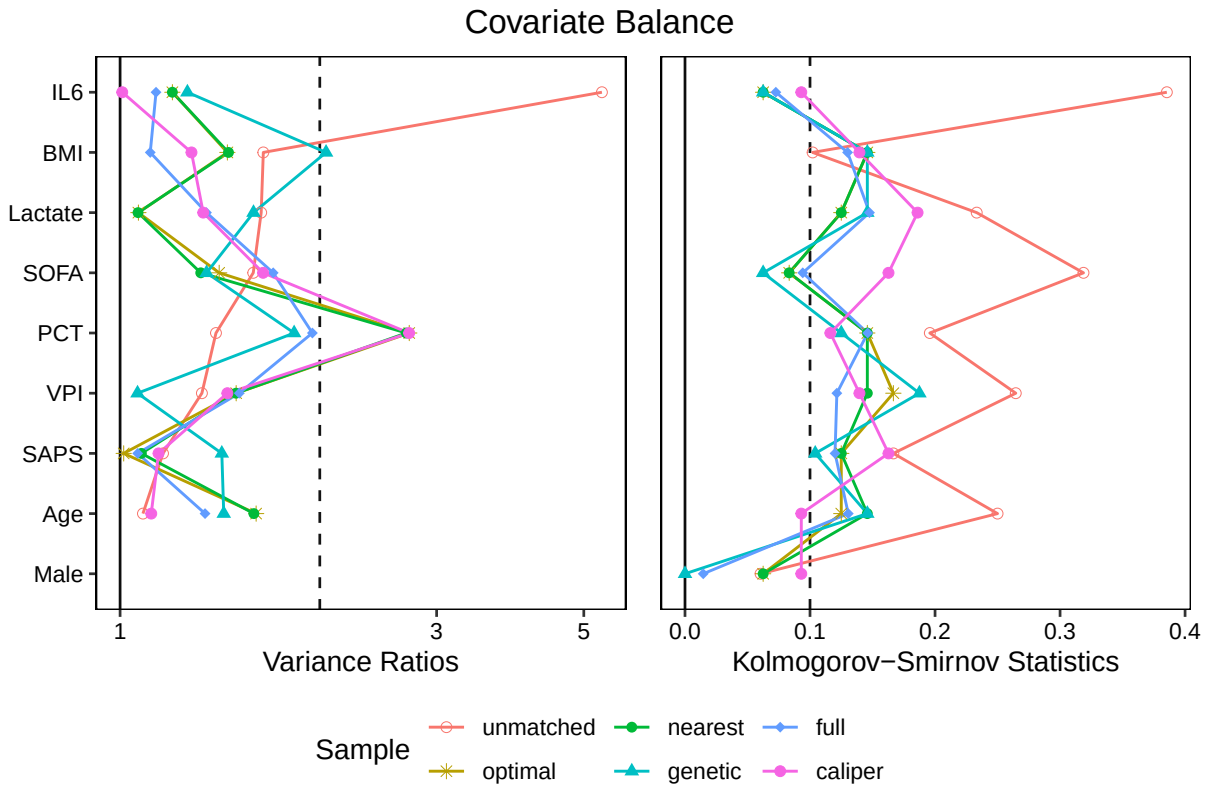

Figure 3.3: Love plot showing the variance ratios and the Kolmogorov-Smirnov statistics between the treatment groups before (unmatched) and after using 5 matching algorithms. The vertical dashed lines mark a variance ratio of 2 and a Kolmogorov-Smirnov statistic of 0.1.

For all variables except BMI, matching decreased the Kolmogorov-Smirnov distance. The result looks quite similar between the different algorithms. Caliper matching performed a bit less well for the variable Sex.

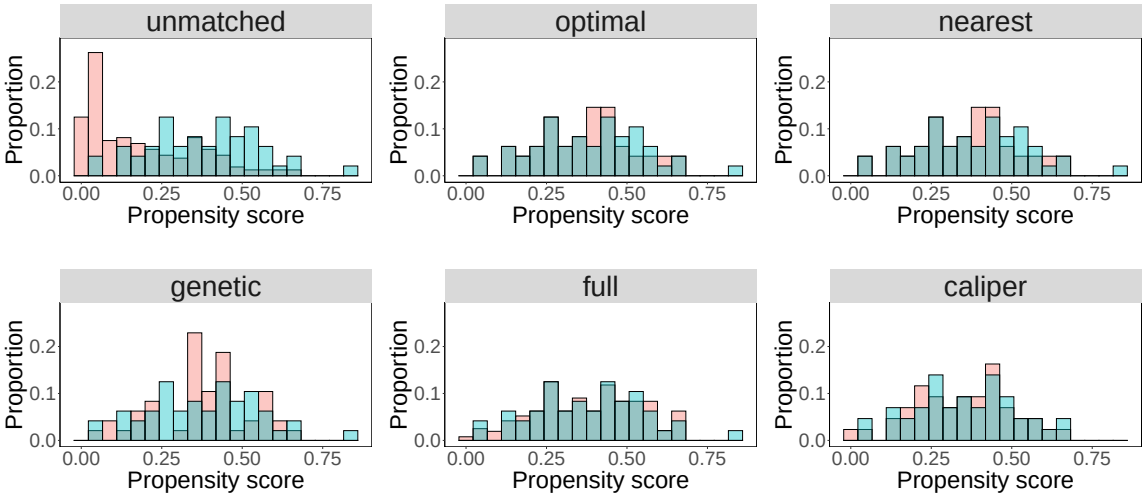

Figure 3.4: Histograms showing the distribution of the propensity score values before (unmatched) and after using 5 matching algorithms. Turquoise: filter group, red: control group.

Another characteristic, that should be looked at, is the common support of the propensity score. Figure 3.4 displays the estimated propensity scores of the matched individuals after applying the different algorithms. In this regard, all matching algorithms seem to perform well.

The number of control individuals contained in the sample differs after the various matching algorithms. Nearest and optimal matching let 48 controls in the sample, which equals the amount of treated individuals. After full matching all 160 and after genetic matching only 35 controls were included in the resulting data set. Caliper matching was the only algorithm also discarding treated patients, in this way including only 43 observations per treatment group.

3.1.3 Treatment effect

Figure 3.5 and Figure 3.6 show the estimated odds ratios (OR) of the Cytosorb filter on in-hospital mortality after unadjusted and multiple adjusted logistic regression, respectively.

All resulting ORs are above 1 stating that treatment by Cytosorb could be rather harmful. Half of the unadjusted estimates are not significant, whereas all odds ratios adjusted for Age, BMI, Sex, SOFA, SAPS, VPI, IL6, PCT, and Lactate are significant on the 5%-level. It is immediately visible, that the marginal treatment effect estimates get smaller after matching, while matching tends to enlarge the conditional estimates.

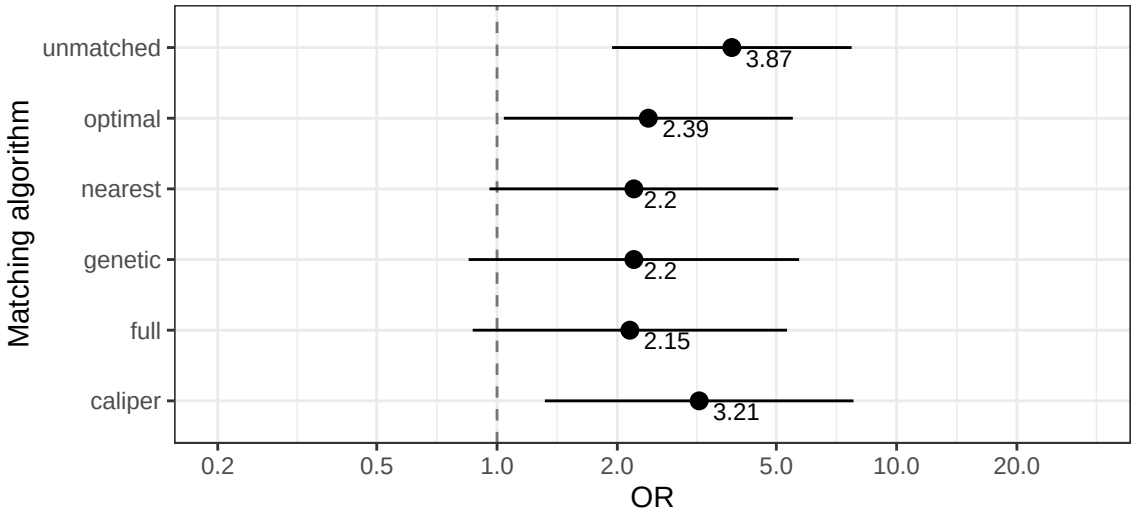

Figure 3.5: Estimated marginal OR without matching and after using 5 matching algorithms, including the 95%-confidence intervals.

3.1.4 Discussion of the motivating example

In the Cytosorb data set there were 42 (17%) observations having missing values for IL6, VPI and Lactate. By exclusion of these patients from the analysis, information and consequently power were lost, even though only 3 of the exclusions were treated patients. We assume that the missing values were missing at random and thus the exclusion should not have included bias.

In Figure 3.1 it can be clearly seen that IL6 and PCT were not normally distributed. So this assumption of the logistic propensity score model and of the logistic outcome model was violated. To solve this problem a transformation of these variables could be a possibility. However, if the balance after matching is good, a misspecification of the propensity score model should be negligible.

The distributions of the propensity score in the two groups were very similar after all matching algorithms. Additionally, the propensity score seems to be more or less normally distributed.

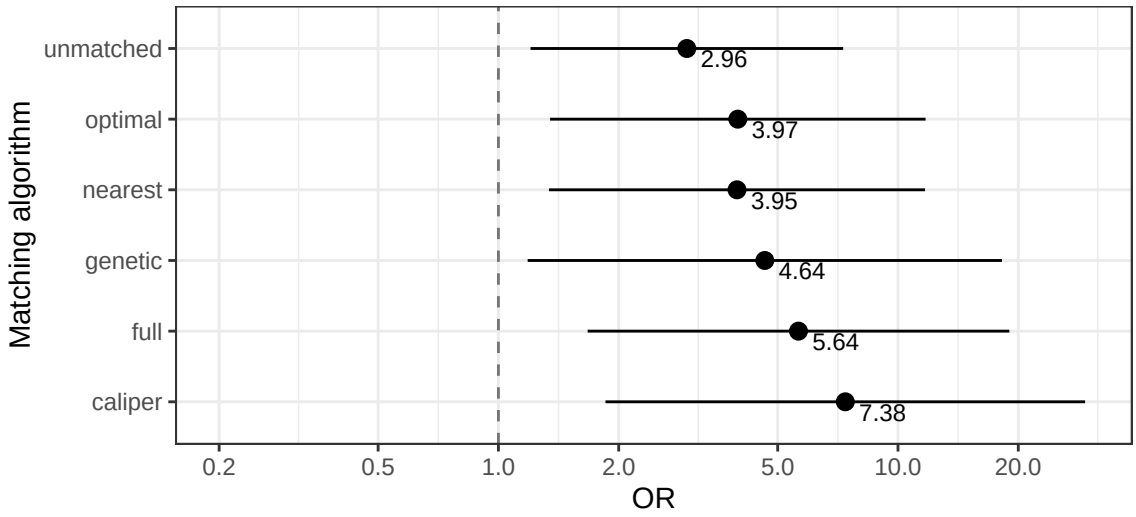

Figure 3.6: Estimated OR without matching and after using 5 matching algorithms, all adjusted for (thus conditional on) Age, BMI, Sex, SOFA, SAPS, VPI, IL6, PCT, and Lactate, including the 95%-confidence intervals.

This confirms that in this case the usage of the linear propensity score would not be more suitable.

Looking at the achieved balance after the different matching algorithms, most variables had SMDs smaller than 0.1. For a better interpretation of balance, clinical knowledge could help. It is desirable that especially variables with a big influence on both treatment decision and outcome are well balanced between the groups (Adelson *et al.*, 2017).

The Kolmogorov-Smirnov statistic of the variable Sex was almost zero after genetic and full matching. It is not surprising that the best value was achieved for a binary variable, because the cumulative distribution of a binary variable only depends on one single value.

Our results suggest that the treatment with the Cytosorb filter has a negative impact on in-hospital survival. Inspecting the unmatched and unadjusted analysis, one could think that the high odds ratio could be caused by selection bias. It is likely that the filter is used more in severely ill patients with a low survival prognosis than in less critically ill patients. The analysis with the matched data sets yielded lower odds ratios and in this way supported this hypothesis. However, double adjustment (matching and covariate adjusted regression) resulted in even higher estimates of the odds ratio, although wider confidence intervals. Here, it must not be forgotten that the unknown true conditional treatment effect is not the same as the true marginal effect (Samuel *et al.*, 2017). For the inclusion of variables in the logistic outcome model the correct specification of the model is more crucial than for the propensity score model. This could be a source of bias in the analysis, because we do not know if all considered variables in reality are confounders and if they all affect the log odds ratio in a linear way.

Our results are consistent with the ones of Schädler *et al.* (2017) who could find a non-significant but mortality increasing effect of Cytosorb in septic patients. Even though there are not many reports of side effects (Poli *et al.*, 2019), it could be imagined that the filter also absorbs molecules with a positive effect on the disease, thus impairing patient conditions.

3.2 Results of the simulation study

3.2.1 Computation time and matching success

Five different matching algorithms were applied and Table 3.2 summarizes how long the computations needed. Nearest matching without and with caliper was the fastest algorithm, followed by optimal and full matching. Genetic matching took the most time.

As the running time of genetic matching depends strongly on `pop.size`, the difference to other algorithms could even be much larger. We used a rather small value in order to ensure that the whole simulation study does not take too long. However, the recommendation is to choose a larger value to get a better matching result (Sekhon, 2011a).

Table 3.2: Average time in seconds needed to run the matching algorithms.

	optimal	nearest	genetic	full	caliper
OR=2	0.14	0.04	7.18	0.11	0.04
OR=5	0.11	0.03	5.58	0.09	0.03
OR=1	0.15	0.04	7.58	0.11	0.04

Caliper matching (with a caliper width of 20% of the standard deviation of the propensity score) is the only algorithm that was not able to match all treated individuals to a control. On average, 18% of the treated group were discarded by caliper matching.

3.2.2 Balance

To give an impression of the balance achieved after matching, Figure 3.7 shows the Love plot of the scenario with the true odds ratio OR = 2. The plots for the other scenarios look the same and are therefore not shown.

It can be seen that optimal and nearest matching were not able to reduce the SMDs as much as the other matching algorithms, especially for variables which were highly unbalanced in the original data set.

3.2.3 Marginal effect estimation

The marginal odds ratios of the treatment on mortality estimated without using any regression adjustment of all three scenarios are shown in Figure 3.8. The logarithmic average of the treatment effect estimates, their standard errors (SE), the bias, the proportion of false negatives or positives (depending on the scenario) as well as the coverage are shown in Table 3.3. The 95%-confidence intervals for the computation of coverage and the decision of significance were calculated by calculating the cluster robust standard errors (which coincide with the simple sandwich standard errors for samples without subclasses) and applying the Wald method.

On average, all estimates are larger than the true value. Comparing the different matching algorithms, the same result can be seen in all three scenarios. The unmatched dataset led to highly overestimated odds ratios and its estimates are much further away from the true than the ones from all matched samples. The estimates after nearest and optimal matching are on average less close to the truth than the estimates after matching with caliper, full matching, or genetic matching. As seen in Figure 3.8, the estimates after full and genetic matching spread more than after the other algorithms.

3.2.4 Conditional effect estimation

Table 3.4 displays the logarithmic average of the conditional treatment effect estimates, their standard errors (SE), the bias, the proportion of false negatives or positives (depending on the

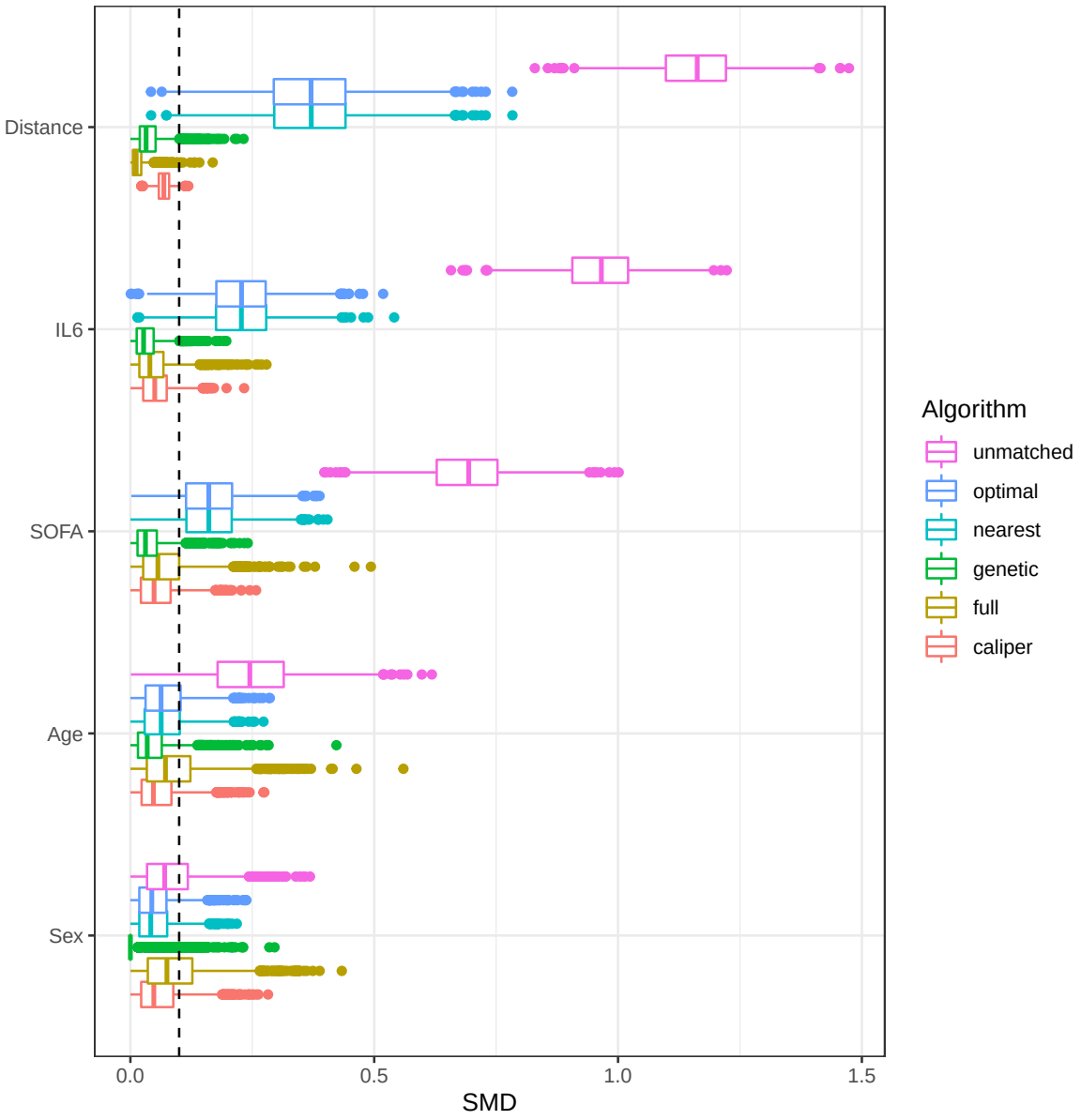

Figure 3.7: Love plot of the scenario OR = 2 showing box plots of the absolute standardized mean differences (SMD) between treated and controls before (unmatched) and after using 5 matching algorithms. The vertical dashed line marks a SMD of 0.1.

scenario) as well as the coverage. In Figure 3.9 the estimated ORs of the treatment on mortality conditional on age, sex, SOFA score and interleukin-6 are presented.

In all scenarios, the averages of the conditional estimates are much closer to the true value than the marginal ones and all matching algorithms as well as the conditional logistic regression without matching yielded similar results. But it is striking that the results of the single simulation runs spread even more than for the marginal estimates.

Table 3.3: Average of the marginal estimates of the log odds ratios $\overline{\hat{\theta}}_j$, their SE as well as bias δ_j, proportion of false negatives or positives and the coverage.
* The proportion of false negatives for the scenarios OR=2 and OR=5 and the proportion of false positives for the scenario OR=1 are based on 95%-confidence intervals.

Scenario	Matching algorithm	$\overline{\hat{\theta}}_j$	SE($\overline{\hat{\theta}}_j$)	Bias	Proportion of false negatives/positives*	Coverage
OR=2	unmatched	1.091	0.216	0.538	0.001	0.290
	optimal	0.696	0.246	0.142	0.215	0.927
	nearest	0.698	0.248	0.144	0.221	0.931
	genetic	0.592	0.321	0.038	0.601	0.964
	full	0.587	0.317	0.033	0.561	0.949
	caliper	0.605	0.262	0.051	0.432	0.970
OR=5	unmatched	1.861	0.242	0.569	0.000	0.340
	optimal	1.466	0.266	0.174	0.000	0.926
	nearest	1.468	0.268	0.176	0.000	0.927
	genetic	1.362	0.336	0.070	0.024	0.968
	full	1.356	0.333	0.064	0.011	0.950
	caliper	1.379	0.285	0.087	0.003	0.960
OR=1	unmatched	0.528	0.214	0.528	0.700	0.300
	optimal	0.133	0.246	0.133	0.080	0.920
	nearest	0.135	0.248	0.135	0.079	0.921
	genetic	0.029	0.325	0.029	0.039	0.961
	full	0.023	0.318	0.023	0.053	0.947
	caliper	0.037	0.263	0.037	0.034	0.966

Table 3.4: Average of the conditional estimates of the log odds ratios $\overline{\hat{\theta}}_j$, their SE as well as bias δ_j, proportion of false negatives or positives and the coverage.
* The proportion of false negatives for the scenarios OR=2 and OR=5 and the proportion of false positives for the scenario OR=1 are based on 95%-confidence intervals.

Scenario	Matching algorithm	$\overline{\hat{\theta}}_j$	SE($\overline{\hat{\theta}}_j$)	Bias	Proportion of false negatives/positives*	Coverage
OR=2	unmatched	0.705	0.263	0.012	0.242	0.950
	optimal	0.718	0.297	0.025	0.318	0.949
	nearest	0.719	0.301	0.026	0.315	0.949
	genetic	0.726	0.390	0.033	0.496	0.937
	full	0.734	0.374	0.041	0.425	0.932
	caliper	0.724	0.326	0.031	0.380	0.952
OR=5	unmatched	1.643	0.290	0.034	0.000	0.955
	optimal	1.670	0.328	0.061	0.000	0.951
	nearest	1.671	0.333	0.061	0.000	0.952
	genetic	1.694	0.415	0.084	0.011	0.947
	full	1.719	0.398	0.109	0.007	0.936
	caliper	1.680	0.363	0.071	0.002	0.954
OR=1	unmatched	0.003	0.268	0.003	0.057	0.943
	optimal	0.008	0.300	0.008	0.056	0.944
	nearest	0.009	0.304	0.009	0.056	0.944
	genetic	0.009	0.402	0.009	0.068	0.932
	full	0.004	0.388	0.004	0.075	0.925
	caliper	0.011	0.322	0.011	0.045	0.955

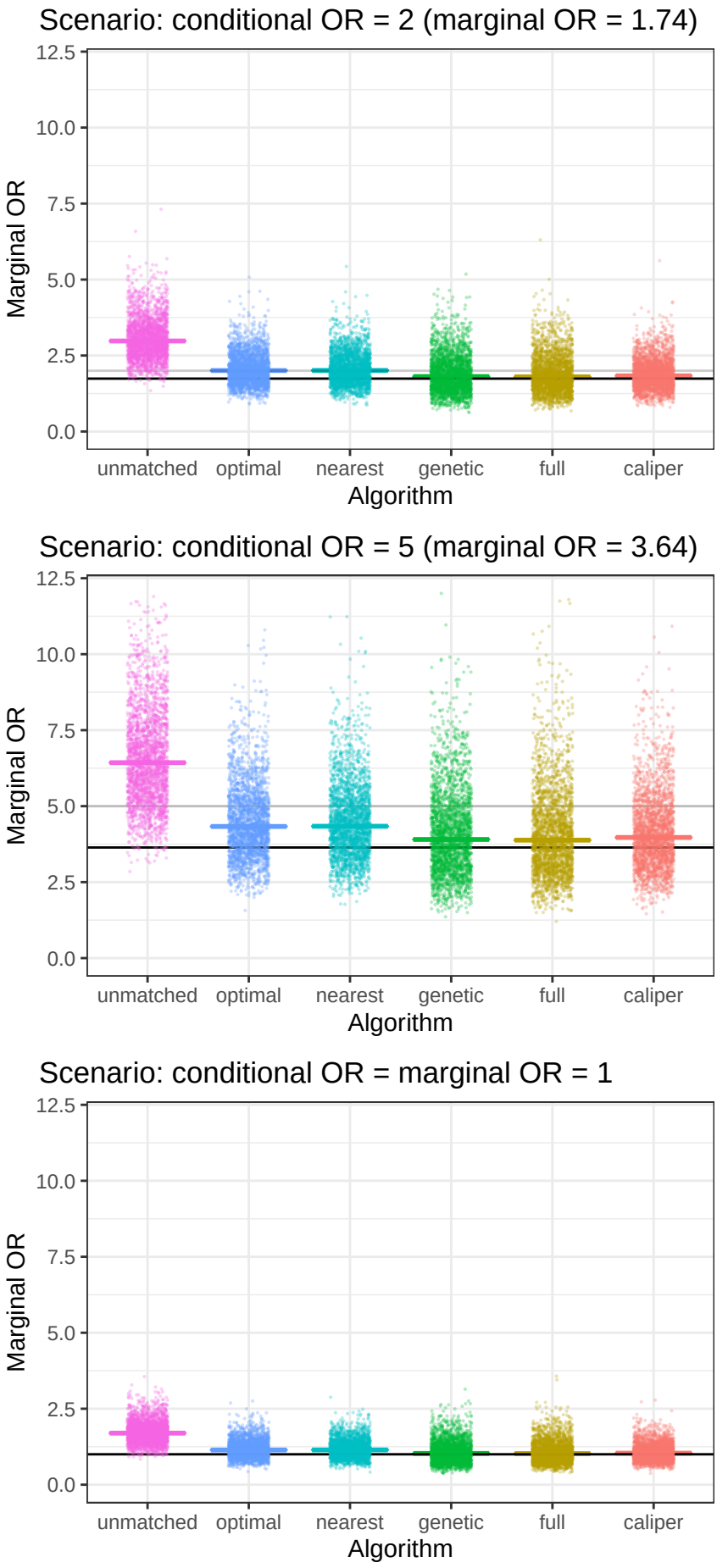

Figure 3.8: Estimated marginal OR of the simulation study.
In the second plot 27 observations are not displayed, because their values are greater than the upper limit of the figure.

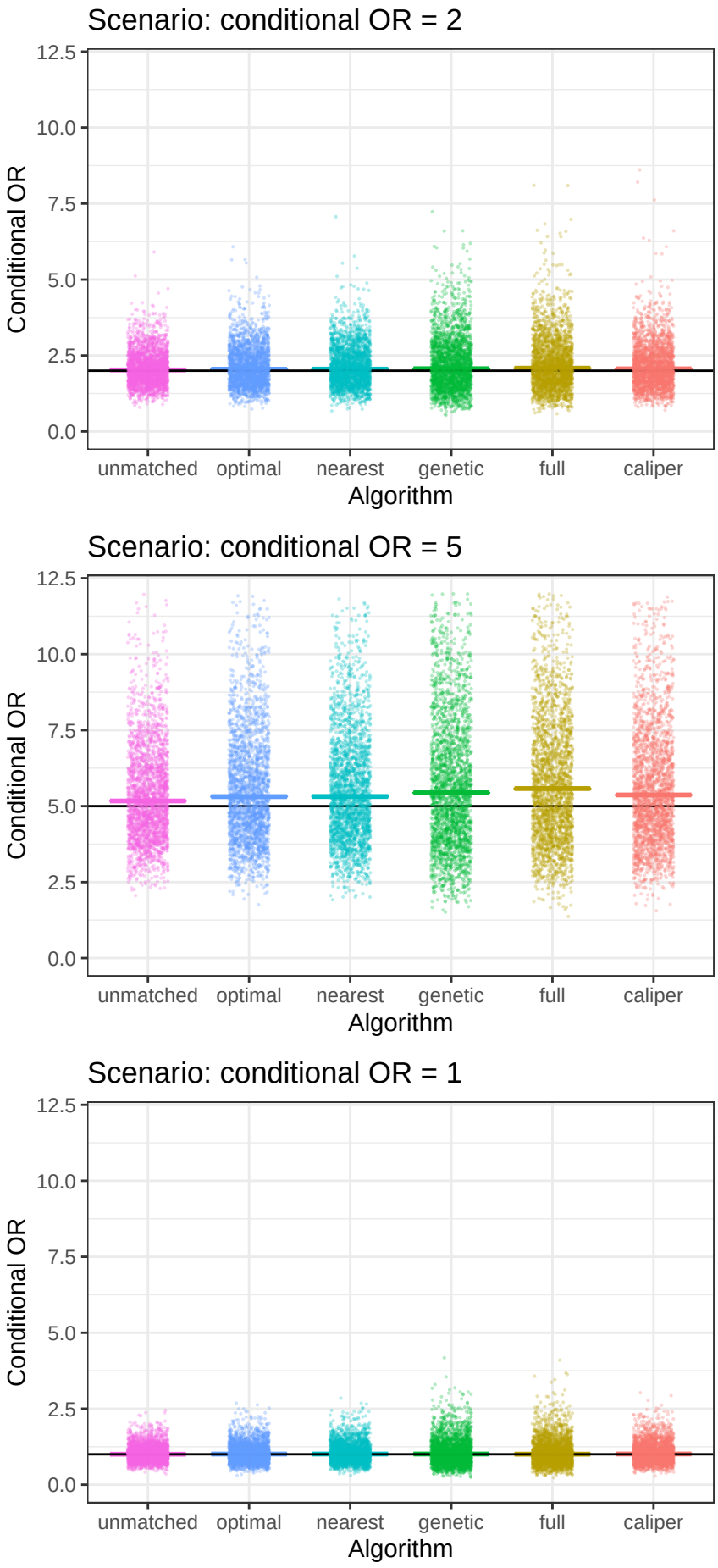

Figure 3.9: Estimated conditional OR of the simulation study.
In the second plot 252 observations are not displayed, because their values are greater than the upper limit of the figure.

3.2.5 Standard errors

In Table 3.5 and Table 3.6 the different standard errors are compared. In the case of full and genetic matching the robust standard errors are substantially larger than the "simple" ones. For the other matching algorithms as well as for the logistic regression of the unmatched sample the differences are only small, leaving the usage of non-robust standard errors acceptable.

The standard errors depicted as "simple" were calculated with the assumption of independent and homoscedastic errors, just applying the `summary` function on the `glm` object. The HC3 standard errors take into account that the errors could be heteroscedastic. Lastly, the cluster robust standard errors incorporate a dependency between the observations in the same subclass. As subclasses are only generated by optimal and full matching, for the other matching algorithms these standard errors are just computed as sandwich standard errors.

The sandwich standard errors (HC0) for the unmatched sample as well as after nearest and caliper matching are exactly the same as the "simple" standard errors, on average. They only differ after genetic matching where not all observations have the same weight. Excluding full and genetic matching, also the HC3 standard errors are just minimally larger than the "simple" ones. This suggests that the assumption of homoscedasticity is met in these samples and thus "simple" standard errors are sufficient.

After full and genetic matching the usage of robust standard errors seems to be necessary since they are considerably larger.

Table 3.5: Comparison of different standard errors (se) of the marginal effect estimation by the simulation study. In parentheses the standard errors of the observed standard errors are shown.
* Cluster robust standard error for optimal and full matching and HC0 standard error for all others.

Scenario	Matching algorithm	Simple se		HC3 se		Cluster robust se*	
OR=2	unmatched	0.215	(0.007)	0.217	(0.007)	0.215	(0.007)
	optimal	0.260	(0.011)	0.262	(0.011)	0.256	(0.016)
	nearest	0.260	(0.011)	0.262	(0.011)	0.260	(0.011)
	genetic	0.295	(0.011)	0.354	(0.030)	0.344	(0.024)
	full	0.213	(0.007)	0.343	(0.044)	0.319	(0.052)
	caliper	0.286	(0.013)	0.289	(0.013)	0.286	(0.013)
OR=5	unmatched	0.243	(0.014)	0.244	(0.014)	0.243	(0.014)
	optimal	0.283	(0.015)	0.285	(0.015)	0.279	(0.019)
	nearest	0.283	(0.015)	0.286	(0.015)	0.283	(0.015)
	genetic	0.316	(0.015)	0.372	(0.030)	0.362	(0.025)
	full	0.241	(0.014)	0.362	(0.044)	0.337	(0.051)
	caliper	0.309	(0.018)	0.312	(0.018)	0.309	(0.018)
OR=1	unmatched	0.211	(0.006)	0.212	(0.006)	0.211	(0.006)
	optimal	0.256	(0.010)	0.259	(0.010)	0.253	(0.015)
	nearest	0.257	(0.010)	0.259	(0.010)	0.257	(0.010)
	genetic	0.292	(0.011)	0.352	(0.030)	0.341	(0.024)
	full	0.209	(0.006)	0.341	(0.044)	0.316	(0.054)
	caliper	0.285	(0.013)	0.288	(0.013)	0.285	(0.013)

In the Appendix A.3 the Monte Carlo standard errors of the coverage are displayed. All errors are smaller than 1%, what was the goal in the sample size calculations (protocol in Section A.2).

Table 3.6: Comparison of different standard errors (se) of the conditional effect estimation by the simulation study. In parentheses the standard errors of the observed standard errors are shown.
* Cluster robust standard error for optimal and full matching and HC0 standard error for all others.

Scenario	Matching algorithm	Simple se		HC3 se		Cluster robust se*	
OR=2	unmatched	0.267	(0.010)	0.271	(0.012)	0.267	(0.012)
	optimal	0.295	(0.015)	0.303	(0.017)	0.295	(0.020)
	nearest	0.295	(0.015)	0.303	(0.017)	0.295	(0.016)
	genetic	0.333	(0.019)	0.402	(0.042)	0.374	(0.029)
	full	0.245	(0.014)	0.385	(0.060)	0.347	(0.047)
	caliper	0.324	(0.020)	0.334	(0.021)	0.324	(0.020)
OR=5	unmatched	0.292	(0.016)	0.297	(0.019)	0.292	(0.018)
	optimal	0.327	(0.023)	0.335	(0.025)	0.327	(0.028)
	nearest	0.327	(0.024)	0.335	(0.025)	0.327	(0.024)
	genetic	0.369	(0.030)	0.430	(0.046)	0.402	(0.035)
	full	0.280	(0.022)	0.412	(0.070)	0.373	(0.054)
	caliper	0.359	(0.029)	0.370	(0.032)	0.359	(0.031)
OR=1	unmatched	0.266	(0.010)	0.270	(0.012)	0.266	(0.012)
	optimal	0.292	(0.014)	0.299	(0.016)	0.292	(0.020)
	nearest	0.292	(0.015)	0.300	(0.016)	0.292	(0.015)
	genetic	0.327	(0.018)	0.401	(0.045)	0.372	(0.031)
	full	0.239	(0.012)	0.390	(0.061)	0.351	(0.048)
	caliper	0.320	(0.019)	0.330	(0.020)	0.320	(0.019)

Chapter 4

Clinical example: nonoperative treatment compared with surgery in patients with lumbar spinal stenosis

In this chapter we reanalyze data which was first published by Held *et al.* (2019).

In the Lumbar Stenosis Outcome Study, a prospective multicenter observational cohort study, 408 patients with degenerative lumbar spinal stenosis were included. The effect of conservative nonsurgical treatment compared to surgery was studied on quality of life, pain, and disability at one year follow-up time.

4.1 Methods of the clinical example

Three outcomes were examined. First, the EQ-5D-3L was considered which is a score for quality of life with higher scores indicating higher quality of life. The value at the 12-month follow-up was assessed. Additionally, the Spinal Stenosis Measure (SSM) assessing both symptoms and function was considered. For the SSM symptoms a minimal clinically important difference (MCID) is defined as an improvement by at least 0.48 points. A MCID in SSM function is defined as an improvement by at least 0.52 points (Stucki *et al.*, 1996). At the 12-month follow-up it was determined if the two SSM, separately, had reached a MCID compared to the individual's baseline value.

Various baseline characteristics were used for matching. The baseline values of the three outcomes were considered, as well as demographic and disease specific patients characteristics. Table 4.1 lists the variables that were used for the analysis.

Matching was done by the following algorithms: optimal, optimal with a 1:2 ratio, nearest, caliper, full and genetic matching. In comparison to the simulation study, the 1:2 ratio matching was included here, because there are more than twice as many controls as treated subjects. Because there were more patients with conservative treatment than who underwent surgery, this nonoperated group was taken as the treatment group. All the matching algorithms were computed by the R function `matchit`. The optimal, nearest, and caliper matching were all performed with a ratio of 1:1, so having as many controls as treated patients in the matched data set. A caliper of 0.1 standard deviations of the propensity score was used, while the matching was done with the optimal algorithm. For the genetic algorithm a population size of 500 was used.

After matching the balance was checked by looking at the standardized mean difference which was calculated using the standard deviation of the treated group (without surgery) in the

denominator. This is well suited for this comparison, because all matching algorithms retained all the treated individuals in the data.

For the sake of completeness we calculated the outcomes of all matched data sets, even though in a practical setting one would probably only use the matched data with the best balance.

The first outcome, EQ-5D-3L, was analyzed by linear regression, whereas the odds ratios for an improvement of the SSM function and SSM symptoms scores were computed by logistic regression. All regressions were once done with the treatment as the only independent variable, and once by including the other covariates which were used for the matching. After full and genetic matching the obtained weights were included in the regression analyses. In the analysis of the full matched data cluster robust standard errors were used in order to account for the subclasses. Robust standard errors (HC3) were also calculated for the results after genetic matching to consider the matching weights. For all other analyses "simple" standard errors were computed. For interested readers other types of standard errors for the EQ5D outcome are shown in Table A.13 in Section A.4.

Table 4.1: List of baseline variables of the lumbar spinal stenosis data.
HADS = Hospital Anxiety and Depression Scale, SSM = Spinal Stenosis Measure.

Variable	Label	Units / levels	Description
age	Age	years	age
female	Female	female or male	sex
diabetes	Diabetes	yes or no	diagnosis of diabetes
bmi	BMI	kg/m^2	BMI = weight / $height^2$
smoking	Smoking	yes or no	smoker
education_risk	Education risk	yes or no	compulsory school only
civil_risk	Civil risk	yes or no	social risk
duration_symptoms_3mo	Duration symptoms	yes or no	duration of symptoms >3 months
cirs	CIRS	score	Chronic Inflammatory Response Syndrome
hads_anx_risk	HADS anxiety	yes or no	HADS anxiety score ≥ 8
hads_dep_risk	HADS depression	yes or no	HADS depression score ≥ 8
ssm_sy	SSM symptoms	score	SSM sympton severity scale
ssm_fu	SSM function	score	SSM function scale
eq5d_ssc	EQ5D	score	EQ-5D-3L sum score
listhese_risk	Listhese risk	yes or no	degenerative spondylolisthesis

4.2 Results of the clinical example

The results of the caliper algorithm are not shown, because optimal matching with a caliper of 0.1 standard deviations of the propensity score did not discard any treated patients and yielded exactly the same matched data set as optimal matching without caliper.

Table 4.2 shows the patient characteristics at baseline. The responding tables after matching can be found in Section A.4.

The achieved balance is displayed in Figure 4.1 and Figure 4.2 by showing the absolute standardized mean differences, and the variance ratios and Kolmogorov-Smirnov statistics, respectively.

Even before matching the balance was not too bad with all variables having a SMD below 0.2 and a variance ratio below 2. The variance ratios could on average not be lowered by matching and genetic matching was the only algorithm which clearly achieved lower SMDs for all except one variable.

In Figure 4.3 to Figure 4.5 the results of the outcome regressions are shown.

As it can be seen in Figure 4.3, a conservative treatment leads on average to a lower EQ-5D-3L sum score after 12 month compared to surgery. This means that lumbar spinal stenosis patients can expect a better quality of life after surgery.

Table 4.2: Baseline characteristics of the patients in the spinal stenosis study.

	Overall	OP	no OP	SMD
No. of patients	408	297	111	
Mean age, years (SD)	72.7 (8.2)	72.3 (8.0)	73.8 (8.4)	0.176
Female, n (%)	210 (51.5)	150 (50.5)	60 (54.1)	0.035
Diabetes, n (%)	46 (11.3)	34 (11.4)	12 (10.8)	0.006
Mean BMI (SD)	27.6 (4.8)	27.4 (4.4)	28.0 (5.6)	0.103
Smoking, n (%)	62 (15.2)	45 (15.2)	17 (15.3)	0.002
Education: compulsory school only, n (%)	94 (23.0)	67 (22.6)	27 (24.3)	0.018
Social risk, n (%)	148 (36.3)	103 (34.7)	45 (40.5)	0.059
Duration of symptoms >3 months, n (%)	368 (90.2)	272 (91.6)	96 (86.5)	0.051
Mean CIRS (SD)	9.3 (3.9)	9.4 (3.9)	9.1 (4.1)	0.087
HADS anxiety score $\geq 8, n(\%)$	77 (18.9)	57 (19.2)	20 (18.0)	0.012
HADS depression score $\geq 8, n(\%)$	72 (17.6)	49 (16.5)	23 (20.7)	0.042
Mean SSM symptoms (SD)	3.1 (0.6)	3.1 (0.6)	3.0 (0.7)	0.137
Mean SSM function (SD)	2.2 (0.7)	2.2 (0.6)	2.2 (0.8)	0.067
Mean EQ-5D-3L sum score (SD)	69.1 (15.8)	68.6 (15.1)	70.6 (17.5)	0.119
Degenerative spondylolisthesis, n (%)	246 (60.3)	186 (62.6)	60 (54.1)	0.086

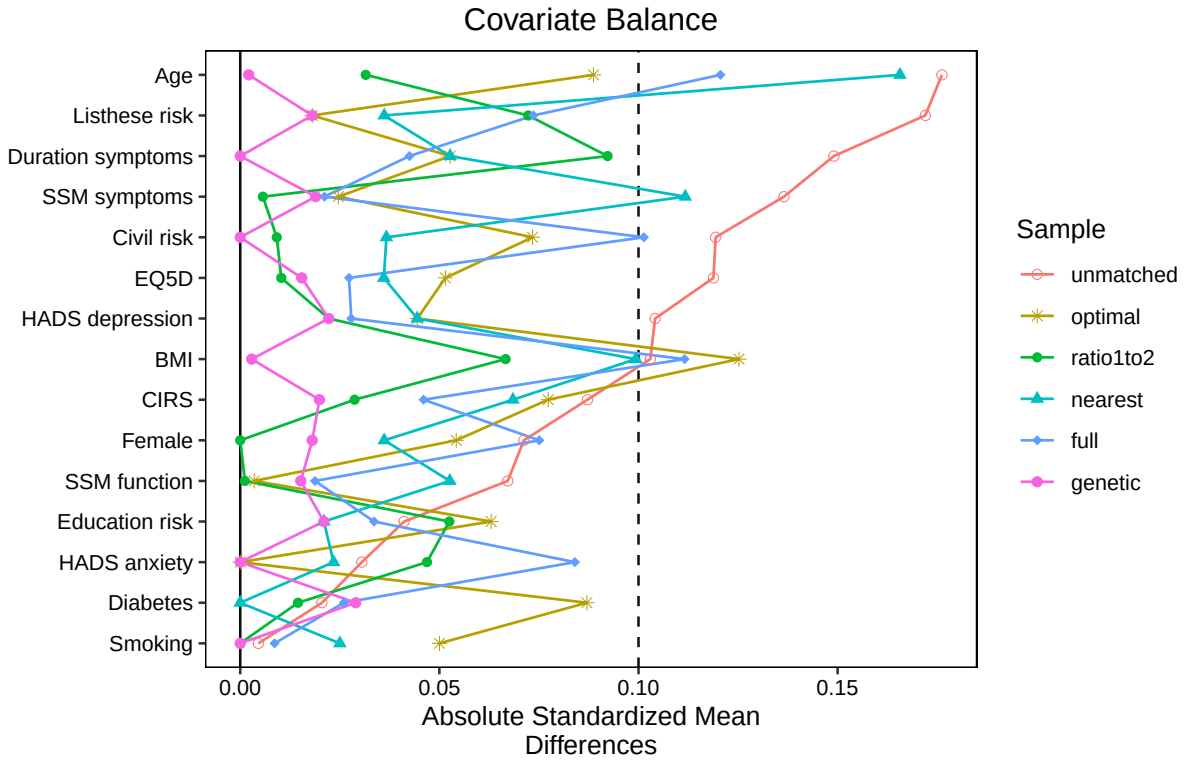

Figure 4.1: Love plot showing the absolute standardized mean differences (SMD) between the treatment groups before (unmatched) and after using 5 matching algorithms.

The odds ratios to gain a clinically significant improvement in both the SSM function and SSM symptoms score after 12 months are below 1 for the nonsurgical treatment compared to surgery, as depicted in Figure 4.4 and Figure 4.5. So there is evidence that patients without an operation have lower chances to reach an improvement.

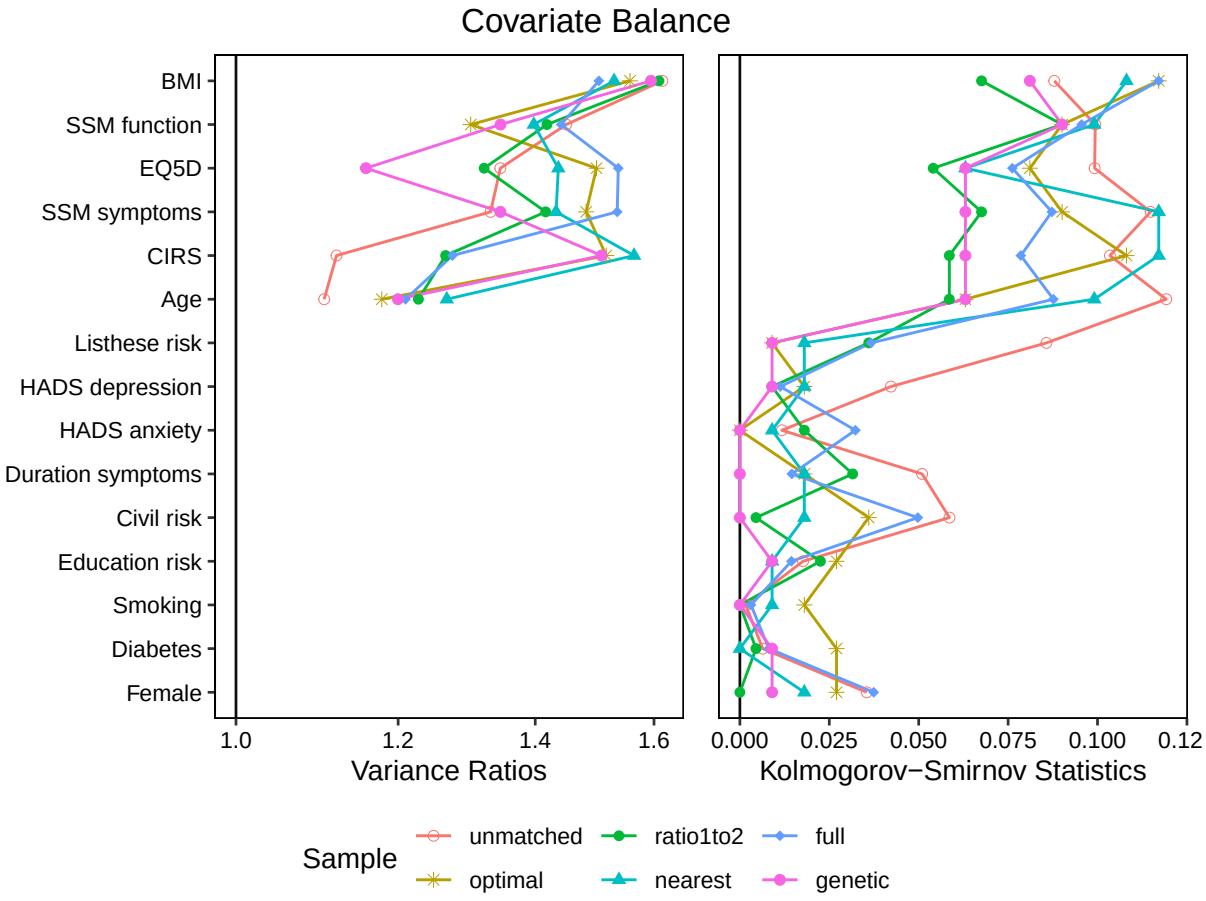

Figure 4.2: Love plot showing the variance ratios and the Kolmogorov-Smirnov statistics between the treatment groups before (unmatched) and after using 5 matching algorithms.

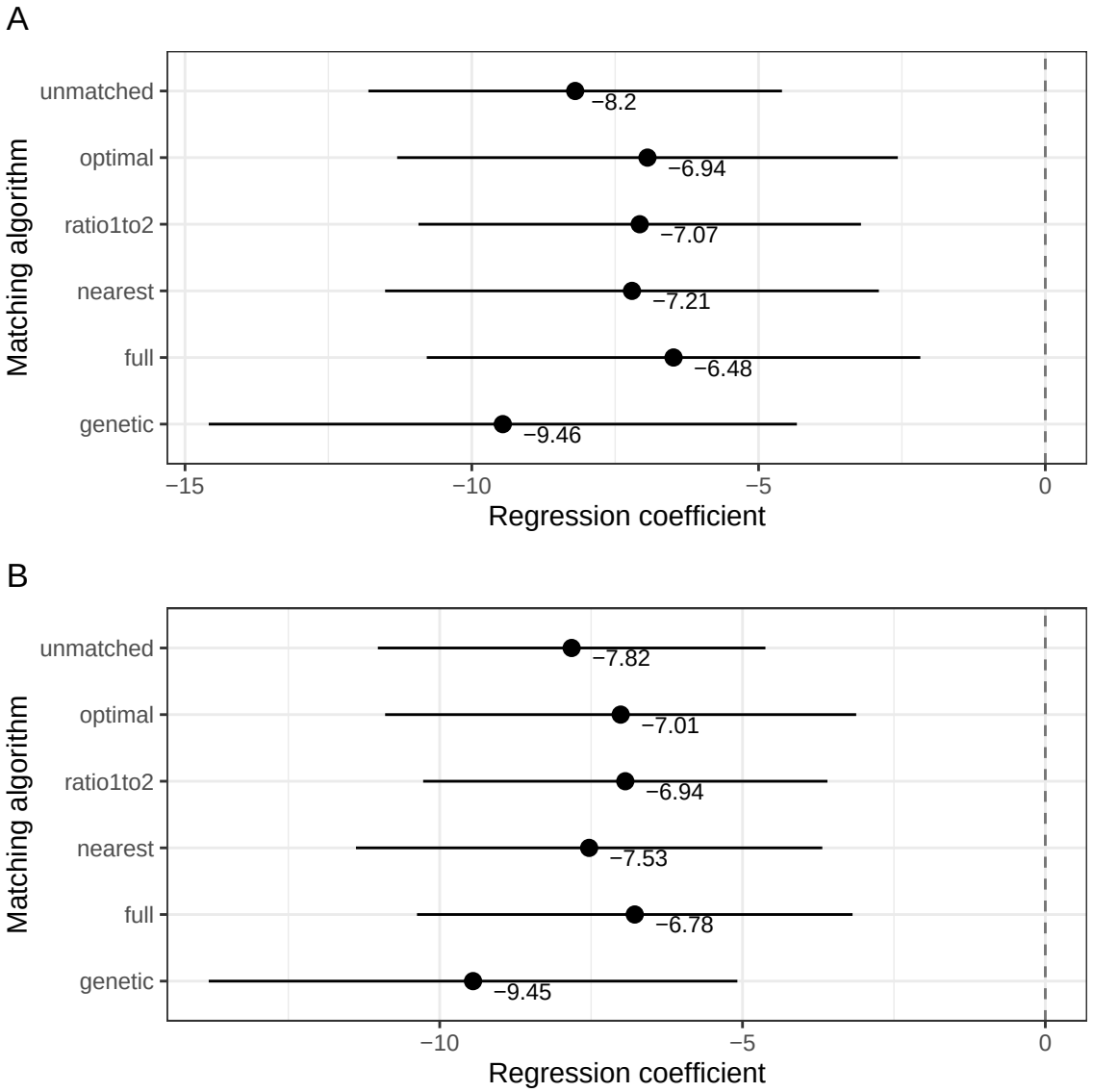

Figure 4.3: The estimated difference at the 12-month follow-up in EQ5D between patients with conservative treatment and patients with surgery, analyzed without matching and after the different matching algorithms, showing the 95%-confidence intervals. Plot A without further covariate adjustment and lower plot B with adjustment for (thus conditional on) age, sex, diabetes, BMI, smoking, education, social risk, duration of symptoms, CIRS, HADS anxiety score, HADS depression score, degenerative spondylolisthesis, baseline SSM symptons, baseline SSM function, and baseline EQ-5D-3L sum score.

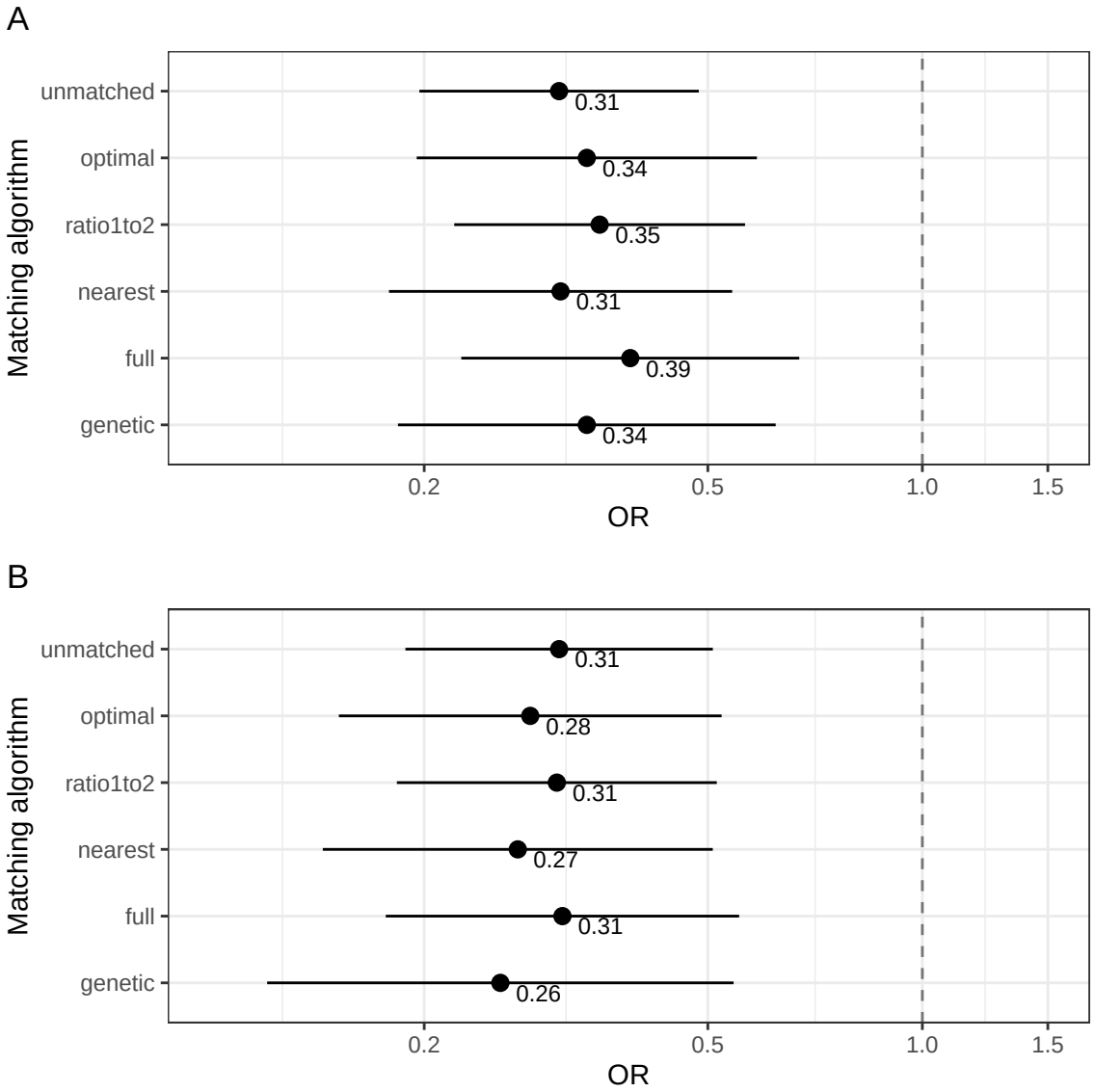

Figure 4.4: The estimated odds ratio (OR) of patients with conservative treatment as compared to patients with surgery for achieving a clinically relevant improvement in SSM symptoms without matching and after the different matching algorithms, showing the 95%-confidence intervals. Plot A without further covariate adjustment (yielding a marginal estimate) and plot B with adjustment for (thus conditional on) age, sex, diabetes, BMI, smoking, education, social risk, duration of symptoms, CIRS, HADS anxiety score, HADS depression score, degenerative spondylolisthesis, baseline SSM symptons, baseline SSM function, and baseline EQ-5D-3L sum score.

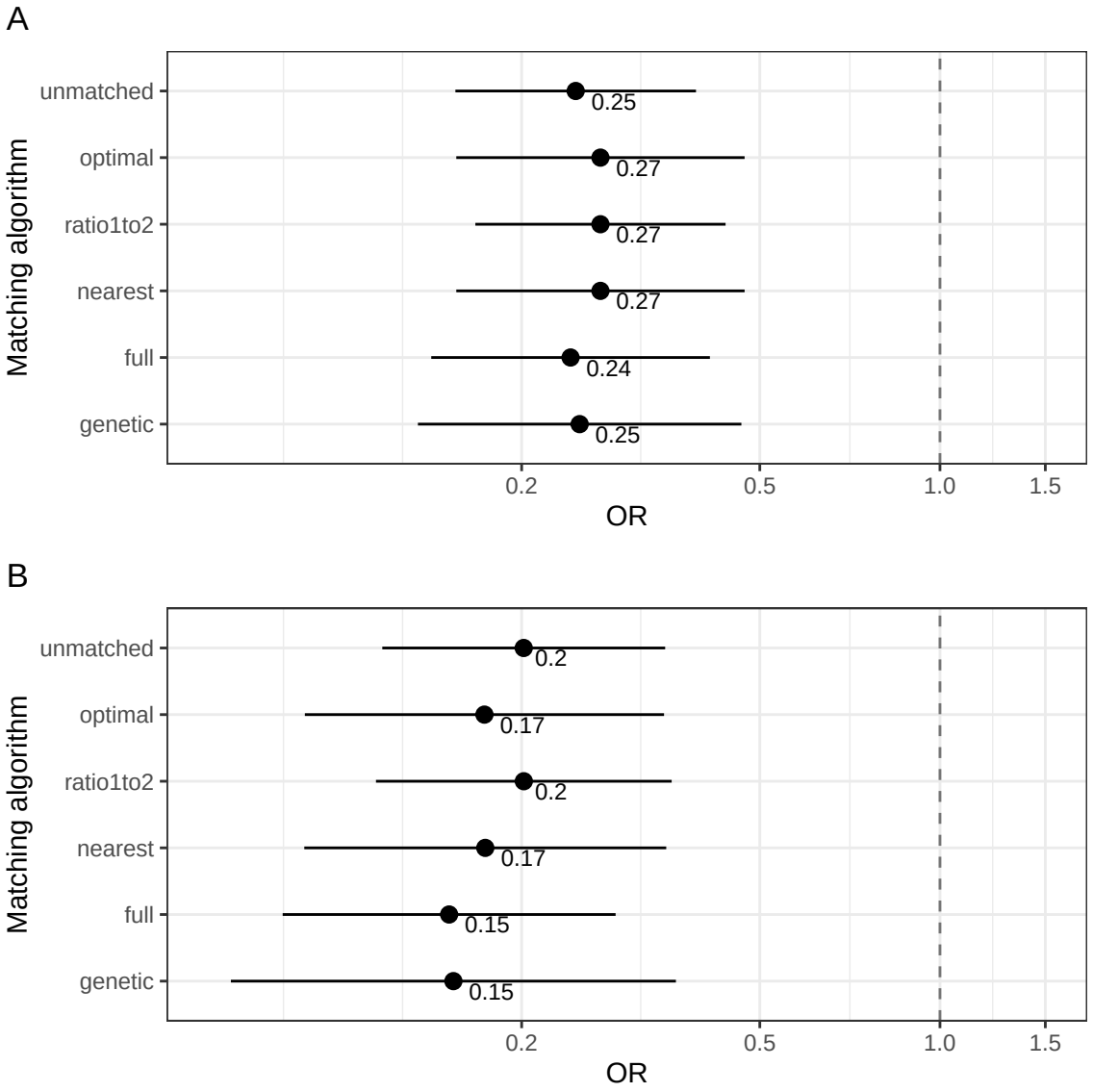

Figure 4.5: The estimated OR of patients with conservative treatment as compared to patients with surgery for achieving a clinically relevant improvement in SSM function without matching and after the different matching algorithms, showing the 95%-confidence intervals. Plot A without further covariate adjustment (yielding a marginal estimate) and plot B with adjustment for (thus conditional on) age, sex, diabetes, BMI, smoking, education, social risk, duration of symptoms, CIRS, HADS anxiety score, HADS depression score, degenerative spondylolisthesis, baseline SSM symptons, baseline SSM function, and baseline EQ-5D-3L sum score.

4.3 Discussion of the clinical example

In this study the patients getting a nonoperative treatment had a lower EQ-5D-3L score 12 months afterwards than patients undergoing surgery, controlling for demographic and clinical covariates. As well, the patients without surgery had smaller chances to gain an improvement in the SSM symptoms and SSM function scores at the 12-month follow-up.

It is standing out that for the EQ-5D-3L analysis all matching algorithms yielded a less extreme result than the regression of the unmatched data with the exception of genetic matching. Genetic matching provided the largest difference, but as well the widest confidence interval. Probably, the reason for that is that the genetic algorithm only chose 93 controls for the 111 treated patients, leading to a smaller sample size and thus to more uncertainty. Of course, this characteristic regarding the uncertainty applies as well to the other outcome analyses.

The estimates considering the SSM function and symptoms scores are all very similar, no matter which matching algorithm was applied. Furthermore, the double adjustment does not seem to change the results much. Most of them got a little bit extremer what is caused by the difference of looking at conditional instead of considering marginal odds ratios. Like already seen in the simulation study, the confidence intervals got larger after double adjustment.

Chapter 5

Discussion

5.1 Summary of findings

The simulation study showed that genetic, full and caliper matching could decrease the standardized mean difference of unbalanced variables much more than optimal and nearest matching. In order to be able to estimate unbiased treatment effects, the SMD should be as small as possible (Harder *et al.*, 2010).

The size of the bias and the coverage of the marginal effect estimates reflected the achieved balance. The unmatched samples yielded the most biased results, followed by the samples of optimal and nearest matching.

On average, the conditional estimates were closer to the truth for all matching algorithms. But attention should be payed to the larger variability of the estimates. By double adjustment the chance to get a result close to the truth is not higher than without. This fact is also depicted by the coverage rates which are smaller for the conditional estimates than for the marginal ones after genetic, full, and caliper matching.

Looking at the conditional treatment effect, the multiple adjusted regression on the unmatched sample yielded the most unbiased result. But it should also be asked, how representative the results are for which population. In the case of matching without discarding treated individuals the average treatment effect on the treated (ATT) is estimated, and thus conclusions can be drawn on patients who typically qualify for the treatment. A drawback of simple regression on data with only a few observations on the edges of a population, for example controls with a very small propensity score, is that the computation relies heavily on extrapolation (Ho *et al.*, 2007). Additionally, the good result of the multiple adjusted regression without matching is caused by the setup of the simulation study. The data was simulated on this model. If real data is looked at, we never now the true underlying structure and it is not easy at all to know which covariates have to be included in what type of model.

All coverage rates were between 92% and 97% with the only exception after the unmatched unadjusted analysis. As an optimal coverage is 95% when using 95%-confidence intervals, this finding encourages the application of matching to analyze observational data.

On average 18% of the treatment group were discarded by caliper matching. When a considerable proportion of the treated is discarded, one has to be careful to whom findings can be generalized to. If all treated patients are kept, the ATT is estimated. Usually this is desirable, because clinicians want to know the effect in patients who qualify for the treatment.

A disadvantage of genetic matching is the time it takes. For the matching of large data sets the difference to other algorithms will probably be much larger than in this simulation study with a sample size of 500.

5.2 Summary of findings in the light of existing literature

5.2.1 Propensity score

To use propensity score matching in the analysis of observational studies is a good possibility to directly target confounding by indication (Glynn *et al.*, 2006). It is able to balance the two treatment groups in a way that an unbiased outcome analysis can be done. The graphical display of the propensity score distribution can detect non-overlapping groups easily (Glynn *et al.*, 2006). This can prevent from wrong conclusions made by comparing apples with pears. Additionally Glynn *et al.* (2006) stated: "Comparison of the distributions of the propensity score between exposed and unexposed subjects can identify those with absolute indications or contra-indications to therapy for whom no comparison may be available.

When deciding which covariates to include in the propensity score model, the ones with high association to both treatment and outcome are most important and they enrich the propensity score model (Adelson *et al.*, 2017). In general it is advisable to include as many covariates as were measured with the exception of variables with a high association to treatment assignment but nearly no association to the outcome because they can bias the effect estimation (Adelson *et al.*, 2017). As the propensity score is a summary score, it is a simple approach to reduce dimensionality of confounders (Glynn *et al.*, 2006).

Harder *et al.* (2010) remembered that even after achieving perfect balance, there could be hidden bias due to unmeasured confounders. This would mean that the ignorability assumption was not fulfilled.

5.2.2 Advantages and limitations of matching

A disadvantage of matching compared to multiple adjusted regression is that no estimation of the effect of covariates is possible (Heinze and Jüni, 2011). Furthermore, the purpose of matching methods is not prediction but the estimation of a minimally biased treatment effect, and thus it is not meaningful to predict the predictive accuracy (Heinze and Jüni, 2011).

King and Nielsen (2019) suggested to prefer coarsened exact matching or Mahalanobis distance matching over propensity score matching. The reason is the difference of what experimental design the different methods imitate. Propensity score matching mimics a completely randomized experiment, whereas coarsened exact matching and Mahalanobis distance matching approximate a fully blocked randomized experimental design. The latter has in general more power and is more efficient. Furthermore, one should keep in mind the counterweights of getting more similar treatment groups and of losing information by discarding extreme observations. King and Nielsen (2019) found that the turning point for propensity score matching is reached faster than for other methods and afterwards imbalance can increase. Thus it should be taken care to really find the best balance between the treatment groups.

5.2.3 Choice of the algorithm

Austin (2014) found that matching with a caliper performs better than simple optimal and nearest neighbor matching. This corresponds to our result. In addition, full and genetic matching are valuable options to get well balanced treatment groups.

I think the developers of R packages, like `MatchIt` and of `cobalt`, did a great job to make available easy-to-use software for researchers in practice. It is really helpful if different methods can be applied by a single package and so only one syntax has to be known. It is worth running more than one matching algorithm and then choosing the one yielding the best balance.

5.3 Limitations and strengths

We only used a linear model for propensity score calculation. One could examine as well different propensity score models, for example including interactions or non-linear terms (Heinze and Jüni, 2011). It might make sense to look both at different propensity score estimation models and various matching algorithms (Harder *et al.*, 2010).

Some matching algorithms have many options which can be changed and which we did not examine. If there are many more controls than treated individuals, another ratio than 1:1 can be used. This option is basically available for all algorithms. A special case is full matching which always builds subclasses with various numbers of treated and control individuals per matched group. But also for full matching it is possible to restrict the treated-to-control ratios (Hansen, 2004). Matching using a caliper can be done by optimal or nearest neighbor matching. In our simulation study only nearest matching was examined this way. Furthermore, different caliper widths could be taken. In the case of genetic matching, the population size can be chosen by finding a trade-off between better results and shorter computation time (Sekhon, 2011a).

A strength of our work is that the simulation study was planned in detail and a protocol was prepared before the simulation was performed, as it was recommended by Burton *et al.* (2006). Even if this preparation takes a lot of time, it pays off in the end when the execution is fast and goes smoothly without requiring bug fixes.

5.4 Implications for further research

A question that should be investigated more, is the correct way of calculating the precision of estimates after matching (Shadish and Steiner, 2010). What standard errors are appropriate? For fixed ratio matching, like optimal or nearest matching, our results reinforce the statement of Stuart (2010) and Schafer and Kang (2008) that "simple" standard errors are sufficient. There is no apparent need to correct for dependence between observations or for heteroscedasticity. On the other hand, it seems important to not only consider the weights in the analysis but also to calculate robust standard errors after full matching. As we found as well a substantial difference between the heteroscedasticity consistent standard errors and the "simple" ones, it should be further investigated if cluster robust standard errors with additional consideration of heteroscedasticity (for example with HC3) could be a even better approach. Greifer (2020b) stressed the importance of robust standard errors when using matching weights in regression, but it remains the question which calculation method suits best.

It would be interesting what the impact of double adjustment is if the outcome model is misspecified. In this simulation study the same model was used for the multiple adjusted outcome regression as was applied for the simulation of the data. Thus it was not expected to insert bias by double adjustment. Conversely, for real data the true underlying structure is not known.

Next to matching, there exist other interesting methods based on the propensity scores. Desai and Franklin (2019) gave a summary of different weighting methods using the propensity score.

In this thesis the focus lies on matching of two treatment groups considering a set of pre-treatment covariates. Additionally, matching and other propensitiy score methods can also be applied to multiple treatment groups as well as to address time-varying confounding (Desai and Franklin, 2019).

Matching can only account for confounding by variables which were observed. In contrast, by randomization all patient characteristics are balanced on average. In observational studies it is recommended to assess the risk of confounding by unmeasured variables by a sensitivity analysis (Caliendo and Kopeinig, 2008). Simulating unobserved variables by ignoring them in the matching algorithm could be another idea for a simulation study to examine their impact on the resulting estimates.

5.5 Implications for practice

Based on the simulation study, the suggestion of Stuart (2010) for the usage of full matching can be supported. Also genetic and caliper matching were able to choose well balanced treatment groups. It is really important that after matching the balance is checked. Only with a satisfying balance the further analysis of outcomes should be started.

Like Nguyen *et al.* (2017) recommended, it seems to be a good idea to adjust also in the regression for variables that still had a standardized mean difference above 0.1 after matching. If the balance is really good for all variables double adjustment is not surely able to improve the results.

Summarizing, I would suggest to try to get balanced data by using full matching. If this is not successful, one can try genetic or caliper matching. Attention has to be payed on potentially discarded observations. If only a small number of variables stay unbalanced, I would recommend to also adjust for these in the regression analysis of the outcome.

5.6 Conclusion

In conclusion, matching is a powerful tool for the analysis of observational studies, nonetheless the matching algorithms, potential loss of power and unmeasurable confounding should be considered when assessing the results for generalizability. Matching begins by forcing the researcher to look in-depth at the differences between treatment groups. After matching, balance and included observations should be examined carefully to allow the right interpretation in the end. By publishing similar results obtained by for example multiple adjusted regression and additionally by usage of one or two matching algorithms the plausibility of a study can be improved.

Appendix

A.1 Additional tables of the motivating example of Cytosorb

Table A.1: Patient characteristics of the Cytosorb data after optimal matching. Here, IQR denotes the first and third quartiles, which are given as a range in brackets.

Variable	Level	Overall	Filter	Control	SMD
n		96	48	48	
Age (mean (SD))		57 (14)	57 (13)	57 (16)	0.040
Sex (%)	f	37 (38.5)	20 (41.7)	17 (35.4)	0.129
	m	59 (61.5)	28 (58.3)	31 (64.6)	
BMI (mean (SD))		26 (6)	25 (6)	26 (7)	0.149
SAPS (mean (SD))		66 (18)	65 (18)	68 (18)	0.140
SOFA (mean (SD))		14 (4)	14 (4)	14 (3)	<0.001
Lactate (median [IQR])		5 [2, 8]	5 [2, 7]	4 [2, 8]	0.029
IL6 (median [IQR])		1369 [1369, 1369]	1369 [1369, 1369]	1369 [1369, 1369]	0.027
PCT (median [IQR])		17 [6, 56]	17 [8, 55]	18 [6, 56]	0.179
VPI (median [IQR])		8 [4, 17]	8 [4, 14]	10 [4, 20]	0.044

Table A.2: Patient characteristics of the Cytosorb data after nearest matching. Here, IQR denotes the first and third quartiles, which are given as a range in brackets.

Variable	Level	Overall	Filter	Control	SMD
n		96	48	48	
Age (mean (SD))		58 (14)	58 (13)	57 (16)	0.102
Sex (%)	f	37 (38.5)	20 (41.7)	17 (35.4)	0.129
	m	59 (61.5)	28 (58.3)	31 (64.6)	
BMI (mean (SD))		26 (6)	26 (6)	26 (7)	0.140
SAPS (mean (SD))		67 (17)	66 (17)	68 (18)	0.096
SOFA (mean (SD))		14 (4)	15 (4)	14 (3)	0.024
Lactate (median [IQR])		4 [2, 8]	4 [2, 7]	4 [2, 8]	0.027
IL6 (median [IQR])		1369 [1369, 1369]	1369 [1369, 1369]	1369 [1369, 1369]	0.027
PCT (median [IQR])		17 [6, 49]	15 [8, 39]	18 [6, 56]	0.150
VPI (median [IQR])		8 [4, 17]	8 [4, 15]	10 [4, 20]	0.014

Table A.3: Patient characteristics of the Cytosorb data after genetic matching. Here, IQR denotes the first and third quartiles, which are given as a range in brackets.

Variable	Level	Overall	Filter	Control	SMD
n		83	35	48	
Age (mean (SD))		57 (15)	58 (13)	57 (16)	0.050
Sex (%)	f	28 (33.7)	11 (31.4)	17 (35.4)	0.085
	m	55 (66.3)	24 (68.6)	31 (64.6)	
BMI (mean (SD))		26 (6)	26 (5)	26 (7)	0.042
SAPS (mean (SD))		67 (17)	67 (16)	68 (18)	0.038
SOFA (mean (SD))		14 (3)	14 (3)	14 (3)	0.113
Lactate (median [IQR])		4 [2, 7]	5 [2, 7]	4 [2, 8]	0.073
IL6 (median [IQR])		1369 [1369, 1369]	1369 [1369, 1369]	1369 [1369, 1369]	0.002
PCT (median [IQR])		18 [7, 51]	16 [9, 47]	18 [6, 56]	0.046
VPI (median [IQR])		9 [5, 16]	9 [5, 14]	10 [4, 20]	0.157

Table A.4: Patient characteristics of the Cytosorb data after full matching. Here, IQR denotes the first and third quartiles, which are given as a range in brackets.

Variable	Level	Overall	Filter	Control	SMD
n		208	160	48	
Age (mean (SD))		61 (16)	63 (15)	57 (16)	0.362
Sex (%)	f	64 (30.8)	47 (29.4)	17 (35.4)	0.129
	m	144 (69.2)	113 (70.6)	31 (64.6)	
BMI (mean (SD))		26 (6)	26 (5)	26 (7)	0.083
SAPS (mean (SD))		63 (19)	62 (19)	68 (18)	0.307
SOFA (mean (SD))		12 (4)	12 (4)	14 (3)	0.769
Lactate (median [IQR])		3 [2, 6]	2 [2, 5]	4 [2, 8]	0.391
IL6 (median [IQR])		1369 [446, 1369]	1037 [302, 1369]	1369 [1369, 1369]	0.988
PCT (median [IQR])		11 [3, 35]	10 [3, 28]	18 [6, 56]	0.062
VPI (median [IQR])		6 [3, 12]	5 [3, 10]	10 [4, 20]	0.507

Table A.5: Patient characteristics of the Cytosorb data after caliper matching. Here, IQR denotes the first and third quartiles, which are given as a range in brackets.

Variable	Level	Overall	Filter	Control	SMD
n		86	43	43	
Age (mean (SD))		59 (14)	60 (14)	58 (15)	0.100
Sex (%)	f	34 (39.5)	19 (44.2)	15 (34.9)	0.191
	m	52 (60.5)	24 (55.8)	28 (65.1)	
BMI (mean (SD))		26 (7)	26 (6)	26 (7)	0.080
SAPS (mean (SD))		67 (17)	68 (16)	66 (17)	0.073
SOFA (mean (SD))		14 (4)	14 (4)	14 (3)	0.148
Lactate (median [IQR])		5 [2, 8]	5 [2, 7]	4 [2, 8]	0.004
IL6 (median [IQR])		1369 [1369, 1369]	1369 [1369, 1369]	1369 [1369, 1369]	0.123
PCT (median [IQR])		17 [7, 64]	17 [7, 80]	18 [6, 56]	0.211
VPI (median [IQR])		9 [5, 17]	9 [4, 15]	10 [5, 20]	0.036

A.2 Protocol of the simulation study

A.2.1 Aims and objectives

The aim of our simulation study is to see how different matching procedures affect the estimated treatment effect from an observational study. The treatment effect of interest is the odds ratio of the Cytosorb filter on in-hospital mortality. The following matching algorithms are compared:

1. nearest neighbor matching,
2. nearest neighbor matching with a caliper,
3. optimal matching,
4. full matching,
5. genetic matching.

All of these are computed by the R package `MatchIt`.

A.2.2 Simulation procedures

Level of dependence between simulated datasets

We will apply each of the aforementioned matching algorithms to all simulated data sets. For different scenarios we will generate different independent data sets. This procedure corresponds to "moderately independent" simulations (Burton *et al.*, 2006).

Allowance for failures

Failure could occur if the logistic model for effect estimation does not converge. When a failure occurs, the concerned sample will be discarded and replaced. We will record the number of failures, the reason and the method for which it occurred.

Software to perform simulations

The simulation study will be performed in R version 4.0.2 using base and tidyverse packages as well as the following packages: `MASS`, `MatchIt`, `rgenoud`, `cobalt`, `sandwich`, `lmtest`.

Random number generator to use

The default random number generator implemented in R is used, which is the "Mersenne- Twister" (Matsumoto and Nishimura, 1998).

Specification of starting seeds

We will use one input seed, namely "202011". Then we will store the state of the random-number generator at the beginning of each repetition, as suggested by Morris *et al.* (2019).

A.2.3 Methods for generating the datasets

We will us the Cytosorb data set as a motivating example. Data are simulated on $n_{obs} = 500$ patients, representing the size of a typical observational study in the field, which the original cytosorb study belongs to. For the sake of some simplification, we will only use the covariates IL-6 (with units ng/ml), SOFA, Age (years) and Sex.

To generate the 3 continuous covariates, we will take the values for the means, standard deviations and correlations similar to the ones of the Cytosorb data. Theses values are shown in

Table A.6: Mean, standard deviation and correlation of covariates.

Variable	Mean	Variance	Covariance to IL-6	Covariance to SOFA	Covariance to Age
IL-6	1	0.3	-	0.6	-0.2
SOFA	12	17	0.6	-	-1
Age	60	200	-0.2	-1	-

table A.6. Assuming a multivariate normal distribution we will generate the data for simulation. The binary covariate Sex will be simulated using a Bernoullli distribution with a probability for female of 0.3.

We will use the logistic link function (see equation (A.1)) to determine the subject-specific probability of treatment. For each subject, an indicator variable for treatment status is generated from a Bernoulli distribution with the appropriate subject-specific probability. Using formula (A.2) the subject-specific probability of the outcome is generated and finally, we will generate an outcome for each subject from a Bernoulli distribution with the subject-specific parameter determined in the prior step.

$$\begin{aligned} logit(p_{\text{treatment}}) = {} & \beta_{(0,\text{treatment})} \\ & + \beta_{(\text{IL6, treatment})} X_{\text{IL6}} + \beta_{(\text{SOFA, treatment})} X_{\text{SOFA}} \\ & + \beta_{(\text{Age, treatment})} X_{\text{Age}} + \beta_{(\text{Sex, treatment})} X_{\text{Sex}}, \end{aligned} \tag{A.1}$$

$$\begin{aligned} logit(p_{\text{outcome}}) = {} & \beta_{(0,\text{ outcome})} + \beta_{(\text{treatment, outcome})} T \\ & + \beta_{(\text{IL6, outcome})} X_{\text{IL6}} + \beta_{(\text{SOFA, outcome})} X_{\text{SOFA}} \\ & + \beta_{(\text{Age, outcome})} X_{\text{Age}} + \beta_{(\text{Sex, outcome})} X_{\text{Sex}}, \end{aligned} \tag{A.2}$$

whereas $T = 1$ for the treated and $T = 0$ for the control group.

This method of data generation is inspired by Austin (2010).

A.2.4 Scenarios to be investigated

1. No treatment effect, conditional OR = 1.
2. Moderate treatment effect as is, conditional OR = 2.
3. Strong treatment effect, conditional OR = 5.

A.2.5 Statistical methods to be evaluated

1. Nearest neighbor matching on propensity score with the order from the largest value of the distance measure to the smallest.
2. Nearest neighbor matching on propensity score with a caliper of 0.2 standard deviations.
3. Optimal matching (loading the add-on package `optmatch`) on propensity score.
4. Full matching (loading the add-on package `optmatch`) on propensity score.
5. Genetic matching (loading the package `Matching`) with a 1:1 ratio of controls to treated units and a population size of 100.

After matching the conditional treatment effect is estimated by logistic regression adjusted for the 4 covariates. After full matching and genetic matching we will incorporate the weights (which in these two cases do not equal 1) in the regression. In addition to the simple standard errors given by the `glm` function, heteroscedasticity-consistent robust standard errors (HC3) and cluster robust standard errors are computed by the R package `sandwich` and the help of the function `coeftest` from the package `lmtest` for all analyses (Zeileis, 2004; Zeileis *et al.*, 2020). To further calculate the 95%-confidence intervals the cluster robust standard errors are used.

The marginal treatment effect is calculated without including the covariates in the regression. To determine the true marginal effect we simulate 100'000 data sets of 10'000 observations and calculate their marginal odds ratio.

A.2.6 Estimates to be stored for each simulation and summary measures to be calculated over all simulations

For each simulation $i = 1, \ldots, n_{\text{sim}}$ and each method $j = 1, \ldots, 5$ we will store:

- the time needed for the matching algorithm,
- number of treated individuals which are discarded,
- absolute standardized differences, a measure for the balance of the covariates, $\text{SMD} = \frac{\overline{X}_1 - \overline{X}_2}{s}$, where s is the standard deviation in the full sample (before matching),
- estimated conditional and marginal treatment effect, θ_{ij}, expressed as log odds ratio (whereof we can calculate the odds ratio afterwards),
- 95% confidence interval of the estimated treatment effect,
- summary of weights obtained by full matching (minimum, maximum, first and third quartile, median and mean),
- failures (specify, when details of potential failures known).

Once all simulations have been performed, we will summarize these estimates for all combinations of scenario and method j:

- the average of the treatment effect estimates, $\overline{\hat{\theta}}_j = \sum_{i=1}^{n_{\text{sim}}} \frac{\hat{\theta}_{ij}}{n_{\text{sim}}}$,
- the SE, calculated as the standard deviation of the estimates, $\hat{\text{SE}} = \sqrt{\frac{1}{n_{\text{sim}}-1} \sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_{ij} - \overline{\hat{\theta}}_j)^2}$,
- the bias, $\delta_j = \overline{\hat{\theta}}_j - \theta$,
- proportion of false positives (in case of a true zero treatment effect) and false negatives (in case of a true treatment effect) using a significance level of $\alpha = 0.05$,
- coverage, meaning the proportion of times the confidence interval includes the true treatment effect, $P(\hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{upp}})$.

For all performance measures the Monte Carlo standard error will be calculated to quantify simulation uncertainty (Morris *et al.*, 2019).

A.2.7 Number of simulations to be performed

The required number of simulations are calculated by taking into account the Monte Carlo standard error (SE) (Morris *et al.*, 2019). We want the Monte Carlo SE of coverage to be lower than 1%. It maximizes for a coverage of 50%, leading to

$$n_{\text{sim}} = \frac{\text{E(coverage)}(1-\text{E(coverage)})}{(\text{Monte Carlo SE}_{\text{max}})^2} = \frac{0.5(1-0.5)}{0.01^2} = 2500.$$

A.2.8 Criteria to evaluate the performance of statistical methods for different scenarios

- Variation of the treatment effect.
- How often is there a treatment effect in the scenario "no treatment effect"?
- How often is there no treatment effect in the scenario "treatment effect"?

A.2.9 Presentation of the simulation results

We will show scatterplots of the estimated treatment effects for the three scenarios, which show the mean estimate and the true value as well. To make comparison of single data sets possible, we will also look at a line plot. A table will show the summarized estimates of all scenarios, as mentioned in Section A.2.6.

A.3 Additional results of the simulation study

Table A.7: Monte Carlo standard errors of coverage.

Scenario	unmatched	optimal	nearest	caliper	full	genetic
OR=2, marginal	0.0091	0.0049	0.0051	0.0034	0.008	0.0052
OR=2, conditional	0.0044	0.0043	0.0043	0.0043	0.0081	0.0057
OR=5, marginal	0.0095	0.005	0.0052	0.0039	0.0072	0.0049
OR=5, conditional	0.0041	0.0042	0.0043	0.0043	0.0074	0.0052
OR=1, marginal	0.0092	0.0051	0.0054	0.0036	0.0079	0.0054
OR=1, conditional	0.0046	0.0046	0.0046	0.0042	0.0083	0.0061

A.4 Additional tables of the clinical example: lumbar spinal stenosis data

Table A.8: Baseline characteristics of the patients after optimal matching.

	Overall	OP	no OP	SMD
No. of patients	222	111	111	
Mean age, years (SD)	74.2 (8.1)	74.5 (7.8)	73.8 (8.4)	0.089
Female, n (%)	123 (55.4)	63 (56.8)	60 (54.1)	0.027
Diabetes, n (%)	27 (12.2)	15 (13.5)	12 (10.8)	0.027
Mean BMI (SD)	27.7 (5.1)	27.3 (4.5)	28.0 (5.6)	0.125
Smoking, n (%)	36 (16.2)	19 (17.1)	17 (15.3)	0.018
Education: compulsory school only, n (%)	57 (25.7)	30 (27.0)	27 (24.3)	0.027
Social risk, n (%)	94 (42.3)	49 (44.1)	45 (40.5)	0.036
Duration of symptoms >3 months, n (%)	194 (87.4)	98 (88.3)	96 (86.5)	0.018
Mean CIRS (SD)	9.2 (3.7)	9.4 (3.3)	9.1 (4.1)	0.077
HADS anxiety score $\geq 8, n(\%)$	40 (18.0)	20 (18.0)	20 (18.0)	0
HADS depression score $\geq 8, n(\%)$	48 (21.6)	25 (22.5)	23 (20.7)	0.018
Mean no. of SSM symptoms (SD)	3.0 (0.6)	3.0 (0.6)	3.0 (0.7)	0.025
Mean SSM function (SD)	2.2 (0.7)	2.2 (0.7)	2.2 (0.8)	0.004
Mean EQ-5D-3L sum score (SD)	71.1 (15.9)	71.5 (14.3)	70.6 (17.5)	0.052
Degenerative spondylolisthesis, n (%)	121 (54.5)	61 (55.0)	60 (54.1)	0.009

Table A.9: Baseline characteristics of the patients after ratio 1:2 matching.

	Overall	OP	no OP	SMD
No. of patients	333	222	111	
Mean age, years (SD)	73.6 (7.9)	73.5 (7.6)	73.8 (8.4)	0.032
Female, n (%)	180 (54.1)	120 (54.1)	60 (54.1)	0
Diabetes, n (%)	37 (11.1)	25 (11.3)	12 (10.8)	0.005
Mean BMI (SD)	27.8 (4.9)	27.6 (4.4)	28.0 (5.6)	0.067
Smoking, n (%)	51 (15.3)	34 (15.3)	17 (15.3)	0
Education: compulsory school only, n (%)	76 (22.8)	49 (22.1)	27 (24.3)	0.023
Social risk, n (%)	134 (40.2)	89 (40.1)	45 (40.5)	0.005
Duration of symptoms >3 months, n (%)	295 (88.6)	199 (89.6)	96 (86.5)	0.032
Mean CIRS (SD)	9.0 (3.8)	9.0 (3.6)	9.1 (4.1)	0.029
HADS anxiety score $\geq 8, n(\%)$	64 (19.2)	44 (19.8)	20 (18.0)	0.018
HADS depression score $\geq 8, n(\%)$	67 (20.1)	44 (19.8)	23 (20.7)	0.009
Mean no. of SSM symptoms (SD)	3.0 (0.6)	3.0 (0.6)	3.0 (0.7)	0.006
Mean SSM function (SD)	2.2 (0.7)	2.2 (0.6)	2.2 (0.8)	0.001
Mean EQ-5D-3L sum score (SD)	70.5 (16.0)	70.5 (15.2)	70.6 (17.5)	0.01
Degenerative spondylolisthesis, n (%)	188 (56.5)	128 (57.7)	60 (54.1)	0.036

Table A.10: Baseline characteristics of the patients after nearest matching.

	Overall	OP	no OP	SMD
No. of patients	222	111	111	
Mean age, years (SD)	74.5 (8.0)	75.2 (7.5)	73.8 (8.4)	0.166
Female, n (%)	122 (55.0)	62 (55.9)	60 (54.1)	0.018
Diabetes, n (%)	24 (10.8)	12 (10.8)	12 (10.8)	0
Mean BMI (SD)	27.7 (5.1)	27.5 (4.6)	28.0 (5.6)	0.099
Smoking, n (%)	35 (15.8)	18 (16.2)	17 (15.3)	0.009
Education: compulsory school only, n (%)	55 (24.8)	28 (25.2)	27 (24.3)	0.009
Social risk, n (%)	92 (41.4)	47 (42.3)	45 (40.5)	0.018
Duration of symptoms >3 months, n (%)	194 (87.4)	98 (88.3)	96 (86.5)	0.018
Mean CIRS (SD)	9.2 (3.7)	9.4 (3.3)	9.1 (4.1)	0.068
HADS anxiety score $\geq 8, n(\%)$	39 (17.6)	19 (17.1)	20 (18.0)	0.009
HADS depression score $\geq 8, n(\%)$	44 (19.8)	21 (18.9)	23 (20.7)	0.018
Mean no. of SSM symptoms (SD)	3.1 (0.6)	3.1 (0.6)	3.0 (0.7)	0.112
Mean SSM function (SD)	2.2 (0.7)	2.2 (0.7)	2.2 (0.8)	0.053
Mean EQ-5D-3L sum score (SD)	70.9 (16.1)	71.3 (14.6)	70.6 (17.5)	0.036
Degenerative spondylolisthesis, n (%)	118 (53.2)	58 (52.3)	60 (54.1)	0.018

Table A.11: Baseline characteristics of the patients after full matching.

	Overall	OP	no OP	SMD
No. of patients	408	297	111	
Mean age, years (SD)	72.7 (8.2)	72.3 (8.0)	73.8 (8.4)	0.176
Female, n (%)	210 (51.5)	150 (50.5)	60 (54.1)	0.035
Diabetes, n (%)	46 (11.3)	34 (11.4)	12 (10.8)	0.006
Mean BMI (SD)	27.6 (4.8)	27.4 (4.4)	28.0 (5.6)	0.103
Smoking, n (%)	62 (15.2)	45 (15.2)	17 (15.3)	0.002
Education: compulsory school only, n (%)	94 (23.0)	67 (22.6)	27 (24.3)	0.018
Social risk, n (%)	148 (36.3)	103 (34.7)	45 (40.5)	0.059
Duration of symptoms >3 months, n (%)	368 (90.2)	272 (91.6)	96 (86.5)	0.051
Mean CIRS (SD)	9.3 (3.9)	9.4 (3.9)	9.1 (4.1)	0.087
HADS anxiety score $\geq 8, n(\%)$	77 (18.9)	57 (19.2)	20 (18.0)	0.012
HADS depression score $\geq 8, n(\%)$	72 (17.6)	49 (16.5)	23 (20.7)	0.042
Mean no. of SSM symptoms (SD)	3.1 (0.6)	3.1 (0.6)	3.0 (0.7)	0.137
Mean SSM function (SD)	2.2 (0.7)	2.2 (0.6)	2.2 (0.8)	0.067
Mean EQ-5D-3L sum score (SD)	69.1 (15.8)	68.6 (15.1)	70.6 (17.5)	0.119
Degenerative spondylolisthesis, n (%)	246 (60.3)	186 (62.6)	60 (54.1)	0.086

Table A.12: Baseline characteristics of the patients after genetic matching.

	Overall	OP	no OP	SMD
No. of patients	204	93	111	
Mean age, years (SD)	73.6 (8.1)	73.4 (7.6)	73.8 (8.4)	0.053
Female, n (%)	112 (54.9)	52 (55.9)	60 (54.1)	0.019
Diabetes, n (%)	22 (10.8)	10 (10.8)	12 (10.8)	0.001
Mean BMI (SD)	28.0 (5.0)	27.9 (4.2)	28.0 (5.6)	0.015
Smoking, n (%)	32 (15.7)	15 (16.1)	17 (15.3)	0.008
Education: compulsory school only, n (%)	49 (24.0)	22 (23.7)	27 (24.3)	0.007
Social risk, n (%)	83 (40.7)	38 (40.9)	45 (40.5)	0.003
Duration of symptoms >3 months, n (%)	177 (86.8)	81 (87.1)	96 (86.5)	0.006
Mean CIRS (SD)	9.1 (3.7)	9.1 (3.3)	9.1 (4.1)	0.006
HADS anxiety score $\geq 8, n(\%)$	36 (17.6)	16 (17.2)	20 (18.0)	0.008
HADS depression score $\geq 8, n(\%)$	40 (19.6)	17 (18.3)	23 (20.7)	0.024
Mean no. of SSM symptoms (SD)	3.0 (0.6)	3.0 (0.6)	3.0 (0.7)	0.031
Mean SSM function (SD)	2.2 (0.7)	2.2 (0.6)	2.2 (0.8)	0.006
Mean EQ-5D-3L sum score (SD)	70.5 (16.7)	70.3 (15.9)	70.6 (17.5)	0.018
Degenerative spondylolisthesis, n (%)	114 (55.9)	54 (58.1)	60 (54.1)	0.04

Table A.13: Comparison of different standard errors computed by the analysis of the continuous outcome EQ-5D-3L of the lumbar spinal stenosis data.

	normal	HC1	HC3	cluster
unmatched	1.63	1.66	1.72	
optimal	1.98	1.97	2.07	1.78
ratio1to2	1.70	1.72	1.78	1.72
nearest	1.96	1.98	2.08	
full	1.64	1.91	2.05	1.83
genetic	2.08	2.09	2.23	

A.5 R code

R functions used in the simulation study.

```
#############################################
####### functions used in simulation study
#############################################

####### simulation of data
############################
simulate_data <- function(n_obs,
                          mean_Age,
                          mean_SOFA,
                          mean_IL6,
                          cov_matr,
                          beta_treat_0,
                          beta_treat_Age,
                          beta_treat_IL6,
                          beta_treat_SOFA,
                          beta_treat_Sex,
                          beta_out_0,
                          beta_out_Age,
                          beta_out_IL6,
                          beta_out_SOFA,
                          beta_out_Sex,
                          beta_out_treat){

### simulation of covariates
multivars <- matrix(mvrnorm(n = n_obs,
                            mu = c(mean_Age, mean_IL6, mean_SOFA),
                            Sigma = cov_matr),
                            ncol = 3)
multivars <- ifelse(multivars < 0, 0, multivars)    # no negative values
sim_dat <- data.frame(multivars)
colnames(sim_dat) <- c("Age", "IL6", "SOFA")
sim_dat$SOFA <- round(sim_dat$SOFA, 0)

# binary variable "Sex" with proportion p(f) = 0.3
sim_dat$Sex <- rbinom(n_obs, 1, 0.3) %>%
               as.factor()
levels(sim_dat$Sex) <- c("m", "f")

### simulation of treatment
sim_dat$Filter_prob <- plogis(beta_treat_0 + beta_treat_IL6 * sim_dat$IL6 +
                      beta_treat_SOFA * sim_dat$SOFA +
                      beta_treat_Age * sim_dat$Age +
                      beta_treat_Sex * (as.numeric(sim_dat$Sex) - 1))
sim_dat$Filter <- rbinom(n_obs, 1, sim_dat$Filter_prob)

### simulation of outcome
sim_dat$Death_prob <- plogis(beta_out_0 + beta_out_treat * sim_dat$Filter +
                             beta_out_IL6 * sim_dat$IL6 +
                             beta_out_SOFA * sim_dat$SOFA +
                             beta_out_Age * sim_dat$Age +
                             beta_out_Sex * (as.numeric(sim_dat$Sex) - 1))
sim_dat$Death <- rbinom(n_obs, 1, sim_dat$Death_prob)

### return simulated dataset
return(sim_dat)
}

####### matching and analysis
#############################
```

```
match_analyze <- function(sim_dat) {

### before matching
n_treated <- nrow(sim_dat[sim_dat$Filter == 1, ])

### matching
start_time <- Sys.time()
matchit_optimal <- matchit(Filter ~ Age + Sex + SOFA + IL6, data = sim_dat,
                       method = "optimal")
end_time <- Sys.time()
time_optimal <- as.numeric(end_time - start_time)
dat_matchit_optimal <- match.data(matchit_optimal)
n_treated_discard_optimal <- summary(matchit_optimal)$nn[3, 2]
n_control_optimal <- summary(matchit_optimal)$nn[2, 1]

start_time <- Sys.time()
matchit_nearest <- matchit(Filter ~ Age + Sex + SOFA + IL6, data = sim_dat,
                           method = "nearest", m.order = "largest")
end_time <- Sys.time()
time_nearest <- as.numeric(end_time - start_time)
dat_matchit_nearest <- match.data(matchit_nearest)
n_treated_discard_nearest <- summary(matchit_nearest)$nn[3, 2]
n_control_nearest <- summary(matchit_nearest)$nn[2, 1]

start_time <- Sys.time()
matchit_caliper <- matchit(Filter ~ Age + Sex + SOFA + IL6, data = sim_dat,
                           method = "nearest", m.order = "largest", caliper = 0.2)
end_time <- Sys.time()
time_caliper <- as.numeric(end_time - start_time)
dat_matchit_caliper <- match.data(matchit_caliper)
n_treated_discard_caliper <- summary(matchit_caliper)$nn[3, 2]
n_control_caliper <- summary(matchit_caliper)$nn[2, 1]

start_time <- Sys.time()
matchit_full <- matchit(Filter ~ Age + Sex + SOFA + IL6, data = sim_dat,
                           method = "full", estimand = "ATT")
end_time <- Sys.time()
time_full <- as.numeric(end_time - start_time)
dat_matchit_full <- match.data(matchit_full)
n_treated_discard_full <- summary(matchit_full)$nn[3, 2]
n_control_full <- summary(matchit_full)$nn[2, 1]
subclasses_full <- length(unique(matchit_full$subclass))
weights_full <- summary(dat_matchit_full[dat_matchit_full$Filter == 0, ]$weights)

start_time <- Sys.time()
matchit_ATE <- matchit(Filter ~ Age + Sex + SOFA + IL6, data = sim_dat,
                       method = "full", estimand = "ATE")   # not further used
end_time <- Sys.time()  # because full matching with "ATE" or "ATT" as estimand
time_ATE <- as.numeric(end_time - start_time)                # -> same weights

dat_matchit_ATE <- match.data(matchit_ATE)
n_treated_discard_ATE <- summary(matchit_ATE)$nn[3, 2]
n_control_ATE <- summary(matchit_ATE)$nn[2, 1]
weights_ATE <- summary(dat_matchit_ATE[dat_matchit_ATE$Filter == 0, ]$weights)

start_time <- Sys.time()
matchit_genetic <- matchit(Filter ~ Age + Sex + SOFA + IL6, data = sim_dat,
                       method = "genetic", ratio = 1, pop.size = 100,
                       print.level = 0)               # default pop.size
end_time <- Sys.time()
time_genetic <- as.numeric(end_time - start_time)
dat_matchit_genetic <- match.data(matchit_genetic)
n_treated_discard_genetic <- summary(matchit_genetic)$nn[3, 2]
```

```r
n_control_genetic <- summary(matchit_genetic)$nn[2, 1]

### balance
# standardized mean differences (for distance, Age, Sex, SOFA and IL-6)
smd_unmatched <- bal.tab(matchit_optimal, abs = TRUE, s.d.denom = "all",
                         binary = "std", un = TRUE)$Balance$Diff.Un
smd_optimal <- bal.tab(matchit_optimal, abs = TRUE, s.d.denom = "all",
                       binary = "std")$Balance$Diff.Adj
smd_nearest <- bal.tab(matchit_nearest, abs = TRUE, s.d.denom = "all",
                       binary = "std")$Balance$Diff.Adj
smd_caliper <- bal.tab(matchit_caliper, abs = TRUE, s.d.denom = "all",
                       binary = "std")$Balance$Diff.Adj
smd_full <- bal.tab(matchit_full, abs = TRUE, s.d.denom = "all",
                    binary = "std")$Balance$Diff.Adj
smd_ATE <- bal.tab(matchit_ATE, abs = TRUE, s.d.denom = "all",
                    binary = "std")$Balance$Diff.Adj
smd_genetic <- bal.tab(matchit_genetic, abs = TRUE, s.d.denom = "all",
                       binary = "std")$Balance$Diff.Adj

### analysis with covariate adjustment -> conditional odds ratio
fit_unmatched <- glm(Death ~ Filter + Age + Sex + SOFA + IL6,
                     data = sim_dat, family = binomial(link = "logit"))
fit_optimal <- glm(Death ~ Filter + Age + Sex + SOFA + IL6,
                   data = dat_matchit_optimal, family = binomial(link = "logit"))
fit_nearest <- glm(Death ~ Filter + Age + Sex + SOFA + IL6,
                   data = dat_matchit_nearest, family = binomial(link = "logit"))
fit_caliper <- glm(Death ~ Filter + Age + Sex + SOFA + IL6,
                   data = dat_matchit_caliper, family = binomial(link = "logit"))
fit_full <- glm(Death ~ Filter +  Age + Sex + SOFA + IL6,
                weights = weights,
                data = dat_matchit_full, family = binomial(link = "logit"))
fit_genetic <- glm(Death ~ Filter +  Age + Sex + SOFA + IL6,
                   weights = weights,
                   data = dat_matchit_genetic, family = binomial(link = "logit"))

logOR_cond_unmatched <- coef(fit_unmatched)[[2]]
logOR_cond_optimal <- coef(fit_optimal)[[2]]
logOR_cond_nearest <- coef(fit_nearest)[[2]]
logOR_cond_caliper <- coef(fit_caliper)[[2]]
logOR_cond_full <- coef(fit_full)[[2]]
logOR_cond_genetic <- coef(fit_genetic)[[2]]

se_cond_unmatched <- summary(fit_unmatched)$coefficient[2,2]
se_cond_optimal <- summary(fit_optimal)$coefficient[2,2]
se_cond_nearest <- summary(fit_nearest)$coefficient[2,2]
se_cond_caliper <- summary(fit_caliper)$coefficient[2,2]
se_cond_full <- summary(fit_full)$coefficient[2,2]
se_cond_genetic <- summary(fit_genetic)$coefficient[2,2]

# robust standard errors
rse_cond_unmatched <- coeftest(fit_unmatched, vcov. = vcovHC)[2,2]
rse_cond_optimal <- coeftest(fit_optimal, vcov. = vcovHC)[2,2]
rse_cond_nearest <- coeftest(fit_nearest, vcov. = vcovHC)[2,2]
rse_cond_caliper <- coeftest(fit_caliper, vcov. = vcovHC)[2,2]
rse_cond_full <- coeftest(fit_full, vcov. = vcovHC)[2,2]
rse_cond_genetic <- coeftest(fit_genetic, vcov. = vcovHC)[2,2]

# cluster robust standard errors
crse_cond_unmatched <- coeftest(fit_unmatched, vcov. = vcovCL, cadjust = FALSE)[2,2]
crse_cond_optimal <- coeftest(fit_optimal, vcov. = vcovCL,
                              cluster = ~subclass, cadjust = FALSE)[2,2]
crse_cond_nearest <- coeftest(fit_nearest, vcov. = vcovCL, cadjust = FALSE)[2,2]
crse_cond_caliper <- coeftest(fit_caliper, vcov. = vcovCL, cadjust = FALSE)[2,2]
```

```
crse_cond_full <- coeftest(fit_full, vcov. = vcovCL,
                           cluster = ~subclass, cadjust = FALSE)[2,2]
crse_cond_genetic <- coeftest(fit_genetic, vcov. = vcovCL, cadjust = FALSE)[2,2]

### analysis without covariate adjustment -> marginal odds ratio
fit_unmatched <- glm(Death ~ Filter,
                     data = sim_dat, family = binomial(link = "logit"))
fit_optimal <- glm(Death ~ Filter,
                   data = dat_matchit_optimal, family = binomial(link = "logit"))
fit_nearest <- glm(Death ~ Filter,
                   data = dat_matchit_nearest, family = binomial(link = "logit"))
fit_caliper <- glm(Death ~ Filter,
                   data = dat_matchit_caliper, family = binomial(link = "logit"))
fit_full <- glm(Death ~ Filter, weights = weights,
                  data = dat_matchit_full, family = binomial(link = "logit"))
fit_genetic<- glm(Death ~ Filter, weights = weights,
                  data = dat_matchit_genetic, family = binomial(link = "logit"))

logOR_marg_unmatched <- coef(fit_unmatched)[[2]]
logOR_marg_optimal <- coef(fit_optimal)[[2]]
logOR_marg_nearest <- coef(fit_nearest)[[2]]
logOR_marg_caliper <- coef(fit_caliper)[[2]]
logOR_marg_full <- coef(fit_full)[[2]]
logOR_marg_genetic <- coef(fit_genetic)[[2]]

se_marg_unmatched <- summary(fit_unmatched)$coefficient[2,2]
se_marg_optimal <- summary(fit_optimal)$coefficient[2,2]
se_marg_nearest <- summary(fit_nearest)$coefficient[2,2]
se_marg_caliper <- summary(fit_caliper)$coefficient[2,2]
se_marg_full <- summary(fit_full)$coefficient[2,2]
se_marg_genetic <- summary(fit_genetic)$coefficient[2,2]

# robust standard errors
rse_marg_unmatched <- coeftest(fit_unmatched, vcov. = vcovHC)[2,2]
rse_marg_optimal <- coeftest(fit_optimal, vcov. = vcovHC)[2,2]
rse_marg_nearest <- coeftest(fit_nearest, vcov. = vcovHC)[2,2]
rse_marg_caliper <- coeftest(fit_caliper, vcov. = vcovHC)[2,2]
rse_marg_full <- coeftest(fit_full, vcov. = vcovHC)[2,2]
rse_marg_genetic <- coeftest(fit_genetic, vcov. = vcovHC)[2,2]

# cluster robust standard errors
crse_marg_unmatched <- coeftest(fit_unmatched, vcov. = vcovCL, cadjust = FALSE)[2,2]
crse_marg_optimal <- coeftest(fit_optimal, vcov. = vcovCL,
                              cluster = ~subclass, cadjust = FALSE)[2,2]
crse_marg_nearest <- coeftest(fit_nearest, vcov. = vcovCL, cadjust = FALSE)[2,2]
crse_marg_caliper <- coeftest(fit_caliper, vcov. = vcovCL, cadjust = FALSE)[2,2]
crse_marg_full <- coeftest(fit_full, vcov. = vcovCL,
                           cluster = ~subclass, cadjust = FALSE)[2,2]
crse_marg_genetic <- coeftest(fit_genetic, vcov. = vcovCL, cadjust = FALSE)[2,2]

### return results
return(c(n_treated, smd_unmatched,
         logOR_cond_unmatched, se_cond_unmatched, rse_cond_unmatched,
         crse_cond_unmatched, logOR_marg_unmatched, se_marg_unmatched,
         rse_marg_unmatched, crse_marg_unmatched,
         time_optimal, n_treated_discard_optimal, n_control_optimal,
         smd_optimal, logOR_cond_optimal, se_cond_optimal, rse_cond_optimal,
         crse_cond_optimal,
         logOR_marg_optimal, se_marg_optimal, rse_marg_optimal, crse_marg_optimal,
         time_nearest, n_treated_discard_nearest, n_control_nearest,
         smd_nearest, logOR_cond_nearest, se_cond_nearest, rse_cond_nearest,
         crse_cond_nearest,
         logOR_marg_nearest, se_marg_nearest, rse_marg_nearest, crse_marg_nearest,
```

```
        time_caliper, n_treated_discard_caliper, n_control_caliper,
        smd_caliper, logOR_cond_caliper, se_cond_caliper, rse_cond_caliper,
        crse_cond_caliper,
        logOR_marg_caliper, se_marg_caliper, rse_marg_caliper, crse_marg_caliper,
        time_full, n_treated_discard_full, n_control_full, subclasses_full,
        smd_full, logOR_cond_full, se_cond_full, rse_cond_full, crse_cond_full,
        logOR_marg_full, se_marg_full, rse_marg_full, crse_marg_full,
        time_genetic, n_treated_discard_genetic, n_control_genetic,
        smd_genetic, logOR_cond_genetic, se_cond_genetic, rse_cond_genetic,
        crse_cond_genetic,
        logOR_marg_genetic, se_marg_genetic, rse_marg_genetic, crse_marg_genetic,
        weights_full))
}
```

A.6 Session info

R version and packages used to generate this report:

R version: R version 4.0.2 (2020-06-22)
Base packages: stats, graphics, grDevices, utils, datasets, methods, base
Other packages: EValue 4.1.1, obsSens 1.3, rbounds 2.1, Matching 4.9-7, MASS 7.3-51.6, sandwich 3.0-0, lmtest 0.9-38, zoo 1.8-8, reshape2 1.4.4, cobalt 4.2.3, rgenoud 5.8-3.0, optmatch 0.9-13, MatchIt 3.0.2, gridExtra 2.3, tidyr 1.1.2, dplyr 1.0.2, plyr 1.8.6, readxl 1.3.1, ggplot2 3.3.2, biostatUZH 1.8.0, survival 3.1-12, xtable 1.8-4, tableone 0.12.0, RColorBrewer 1.1-2, knitr 1.29
This document was generated on March 30, 2021 at 09:32.

Bibliography

Abadie, A. and Spiess, J. (2021). Robust post-matching inference. *Journal of the American Statistical Association*, **0**, 1–13.

Adelson, J. L., McCoach, D., Rogers, H., Adelson, J. A., and Sauer, T. M. (2017). Developing and applying the propensity score to make causal inferences: variable selection and stratification. *Frontiers in Psychology*, **8**, 1413.

Ahmed, A., Husain, A., Love, T. E., Gambassi, G., Dell'Italia, L. J., Francis, G. S., Gheorghiade, M., Allman, R. M., Meleth, S., and Bourge, R. C. (2006a). Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *European Heart Journal*, **27**, 1431–1439.

Ahmed, A., Perry, G. J., Fleg, J. L., Love, T. E., Goff Jr, D. C., and Kitzman, D. W. (2006b). Outcomes in ambulatory chronic systolic and diastolic heart failure: a propensity score analysis. *American Heart Journal*, **152**, 956–966.

Ankawi, G., Xie, Y., Yang, B., Xie, Y., Xie, P., and Ronco, C. (2019). What have we learned about the use of cytosorb adsorption columns? *Blood Purification*, **48**, 196–202.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, **27**, 2037–2049.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, **28**, 3083–3107.

Austin, P. C. (2010). A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics–Simulation and Computation*, **39**, 563–577.

Austin, P. C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, **46**, 399–424.

Austin, P. C. (2011b). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, **10**, 150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, **33**, 1057–1069.

Austin, P. C., Grootendorst, P., and Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in Medicine*, **26**, 734–753.

Austin, P. C. and Stafford, J. (2008). The performance of two data-generation processes for data with specified marginal treatment odds ratios. *Communications in Statistics–Simulation and Computation*, **37**, 1039–1051.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H., De Boer, A., and Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety*, **20**, 1115–1129.

Bertsekas, D. P. (1981). A new algorithm for the assignment problem. *Mathematical Programming*, **21**, 152–171.

Bertsekas, D. P. (1990). The auction algorithm for assignment and other network flow problems: A tutorial. *Interfaces*, **20**, 133–149.

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *BMJ*, **312**, 1215–1218.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**, 4279–4292.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, **22**, 31–72.

Chen, Q., Nian, H., Zhu, Y., Talbot, H. K., Griffin, M. R., and Harrell Jr, F. E. (2016). Too many covariates and too few cases?–a comparative study. *Statistics in Medicine*, **35**, 4546–4558.

Cochrane, W. and Rubin, D. (1973). Controlling bias in observational studies. *Sankhya: The Indian Journal of Statistics*, **35**, 417–446.

D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, **17**, 2265–2281.

Desai, R. J. and Franklin, J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ*, **367**, l5657.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics*, **95**, 932–945.

Flury, B. K. and Riedwyl, H. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician*, **40**, 249–251.

Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B., and Schneeweiss, S. (2014). Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, **33**, 1685–1699.

Freedman, D. A. (2006). On the so-called Huber sandwich estimator and robust standard errors. *The American Statistician*, **60**, 299–302.

Glynn, R. J., Schneeweiss, S., and Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic & Clinical Pharmacology & Toxicology*, **98**, 253–259.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. (2020). *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.

Greifer, N. (2020a). *cobalt: Covariate Balance Tables and Plots*. R package version 4.2.3.

Greifer, N. (2020b). Vignette MatchIt: Estimating effects after matching.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, **2**, 405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, **99**, 609–618.

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

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, **15**, 609–627.

Harder, V. S., Stuart, E. A., and Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, **15**, 234–249.

Heinze, G. and Jüni, P. (2011). An overview of the objectives of and the approaches to propensity score analyses. *European Heart Journal*, **32**, 1704–1708.

Held, U., Steurer, J., Pichierri, G., Wertli, M. M., Farshad, M., Brunner, F., Guggenberger, R., Porchet, F., Fekete, T. F., Schmid, U. D.,*et al.* (2019). What is the treatment effect of surgery compared with nonoperative treatment in patients with lumbar spinal stenosis at 1-year follow-up? *Journal of Neurosurgery: Spine*, **31**, 185–193.

Hernán, M. A., Clayton, D., and Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, **40**, 780–785.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, **15**, 199–236.

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.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, **20**, 1–24.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**, 481–502.

Jones, A. E., Trzeciak, S., and Kline, J. A. (2009). The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical Care Medicine*, **37**, 1649–1654.

King, G. and Nielsen, R. A. (2019). Why propensity scores should not be used for matching. *Political Analysis*, **27**, 435–454.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, **270**, 2957–2963.

Lee, J. and Little, T. D. (2017). A practical guide to propensity score analysis for applied clinical research. *Behaviour Research and Therapy*, **98**, 76–90.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, **54**, 217–224.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., and Klungel, O. H. (2008). Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *International Journal of Epidemiology*, **37**, 1142–1147.

Matsumoto, M. and Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, **8**, 3–30.

McCaffrey, D., Burgette, L., Griffin, B. A., and Martin, C. (2015). Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package.

Mebane Jr, W. R., Sekhon, J. S.,*et al.* (2011). Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software*, **42**, 1–26.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**, 2074–2102.

Nguyen, T.-L., Collins, G. S., Spence, J., Daurès, J.-P., Devereaux, P., Landais, P., and Le Manach, Y. (2017). Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology*, **17**, 1–8.

O'neill, B. (2006). *Elementary differential geometry*. Elsevier, 2 edition.

Osborne, J. W. (2008). *Best practices in quantitative methods*, chapter 11: Best practices in quasi-experimental designs: Matching methods for causal inference, 155–176. Sage Thousand Oaks, CA.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49**, 1373–1379.

Pimentel, S. D. (2016). Large, sparse optimal matching with r package rcbalance. *Observational Studies*, **2**, 4–23.

Poli, E. C., Rimmele, T., and Schneider, A. G. (2019). Hemoadsorption with cytosorb. *Intensive Care Medicine*, **45**, 236–239.

Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z., and Sekhon, J. S. (2012). Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *The international Journal of Biostatistics*, **8**, 25.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. (2014). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, **84**, 1024–1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**, 597–610.

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and its Application*, **7**, 143–176.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**, 33–38.

Rosenbaum, P. R.*et al.* (2010). *Design of observational studies*. Springer.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, **29**, 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185–203.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, **75**, 591–593.

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

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

Rubin, D. B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, **95**, 573–585.

Samuel, M., Schuster, T., Kaufman, J. S., Platt, R. W., and Brophy, J. M. (2017). Differences between conditional and marginal propensity score estimates: a real-world application. *Journal of the American College of Cardiology*, **70**, 117–117.

Schädler, D., Pausch, C., Heise, D., Meier-Hellmann, A., Brederlau, J., Weiler, N., Marx, G., Putensen, C., Spies, C., Jörres, A.,*et al.* (2017). The effect of a novel extracorporeal cytokine hemoadsorption device on il-6 elimination in septic patients: a randomized controlled trial. *PloS one*, **12**, 10.

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, **13**, 279–313.

Sekhon, J. S. (2011a). Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software*, **42**, .

Sekhon, J. S. (2011b). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software*, **42**, 1–52.

Shadish, W. R. and Steiner, P. M. (2010). A primer on propensity score analysis. *Newborn and Infant Nursing Reviews*, **10**, 19–26.

Silverman, S. L. (2009). From randomized controlled trials to observational studies. *The American Journal of Medicine*, **122**, 114–120.

Sjölander, A. and Greenland, S. (2013). Ignoring the matching variables in cohort studies – when is it valid and why? *Statistics in Medicine*, **32**, 4696–4708.

Steurer, J., Nydegger, A., Held, U., Brunner, F., Hodler, J., Porchet, F., Min, K., Mannion, A. F., and Michel, B. (2010). Lumbsten: the lumbar spinal stenosis outcome study. *BMC Musculoskeletal Disorders*, **11**, 254.

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, statistics in medicine. *Statistics in Medicine*, **27**, 2062–2065.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, **25**, 1–21.

Stuart, E. A. and Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, **44**, 395–406.

Stucki, G., Daltroy, L., Liang, M. H., Lipson, S. J., Fossel, A. H., and Katz, J. N. (1996). Measurement properties of a self-administered outcome measure in lumbar spinal stenosis. *Spine*, **21**, 796–803.

Thoemmes, F. J. and Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, **46**, 90–118.

Vandenbroucke, J. P., Von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., and Initiative, S. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLOS Medicine*, **4**, 10.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., Initiative, S., *et al.* (2014). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *International Journal of Surgery*, **12**, 1495–1499.

Woodward, M. (2014). *Epidemiology: study design and data analysis*. CRC press.

Yao, X. I., Wang, X., Speicher, P. J., Hwang, E. S., Cheng, P., Harpole, D. H., Berry, M. F., Schrag, D., and Pang, H. H. (2017). Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. *JNCI: Journal of the National Cancer Institute*, **109**, 8.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, **11**, 10.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, **16**, 1–16.

Zeileis, A., Köll, S., and Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, **95**, 1–36.

Zhang, Z., Kim, H. J., Lonjon, G., Zhu, Y., *et al.* (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, **7**, 1.