Aus dem deutschen Diabetes-Zentrum
Leibniz-Zentrum für Diabetes-Forschung
an der Heinrich-Heine-Universität Düsseldorf
Institut für Biometrie und Epidemiologie
Direktor: Univ.-Prof. Dr. Oliver Kuß

Comparison of different propensity score methods for estimating treatment effects in non-randomized studies with survival data

Dissertation

zur Erlangung des Grades eines Doktors in Public Health
der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Tim Filla

2024

Als Inauguraldissertation gedruckt mit Genehmigung der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:

Dekan: Prof. Dr. med. Nikolaj Klöcker

Erstgutachter: Prof. Dr. Oliver Kuß

Zweitgutachter: Prof. Dr. Holger Schwender

Teile dieser Arbeit wurden veröffentlicht:

Filla, T, Schwender, H, Kuss, O. (2024), Balancing versus modelling in weighted analysis of non-randomised studies with survival outcomes: A simulation study. *Stat Med.* 43(17), 3140-3163.

Zusammenfassung

Vergleich verschiedener Propensity Score-Methoden zur Schätzung der Behandlungseffekte in nicht-randomisierten Studien mit Überlebensdaten.

Beobachtungsstudien werden häufig in der Public Health Forschung eingesetzt, da die Durchführung von randomisierten Studien aus ethischen oder finanziellen Gründen schwierig ist. Die Analyse von Beobachtungsdaten ist jedoch im Vergleich zu randomisierten Studien mit größeren Herausforderungen verbunden, insbesondere im Hinblick auf das Risiko einer Verzerrung durch Störfaktoren. In den letzten Jahrzehnten wurden viele Methoden für die Analyse von Beobachtungsdaten entwickelt. Eine davon ist die Propensity Score (PS) Methode. Zahlreiche Methoden, die den PS verwenden, wurden in den letzten zehn Jahren entwickelt. Dies erschwert für die Forschenden die Wahl der besten Methode. Simulationsstudien können Aufschluss darüber geben, welche Methoden in bestimmten Situationen am besten funktionieren. Es mangelt jedoch an Simulationsstudien zur Verwendung von PS Methoden in Szenarien mit Überlebensdaten, die häufig verwendet werden.
Das übergeordnete Ziel dieser Arbeit ist es, diese Lücke zu schließen und den Anwendenden in der Public Health Forschung eine klare Anleitung für die Wahl der PS Methode zu geben. Zu diesem Zweck wurde eine detaillierte Simulationsstudie mit Simulationseinstellungen durchgeführt, die durch Informationen aus veröffentlichten PS Analysen motiviert wurden, welche durch eine systematische Suche in „PubMed“ gefunden wurden. Darüber hinaus wurde eine neue PS Methode, die verallgemeinerte Methode der Momente-Matching-Gewichte (GMMW), zusammen mit einem korrigierten Varianzschätzer, der für eine angemessene Berechnung des Konfidenzintervalls erforderlich ist, vorgeschlagen.

Die Ergebnisse der Simulationsstudie zeigten, dass alle Methoden (overlap weights (OW), matching weights (MW) und GMMW) zur Schätzung des durchschnittlichen Behandlungseffekts in der Überlappungspopulation (ATO) in fast allen Fällen die besten Ergebnisse und den kleinsten Standardfehler des Schätzers lieferten. Ein Vergleich der Methoden ergab ähnliche Ergebnisse hinsichtlich der Verzerrung und des Standardfehlers des Schätzers. Bei allen Methoden zur Schätzung des durchschnittlichen Behandlungseffekts (ATE) und des durchschnittlichen Behandlungseffekts der Behandelten (ATT) erwies sich die Entropie-ausgleichsmethode (EB) als die beste Methode. Darüber hinaus ist die Leistung der Standardmethode inverse Wahrscheinlichkeitsgewichtung (IPTW) bei der Schätzung des ATEs schlecht, bei der Schätzung des ATTs jedoch recht gut (und ähnlich wie die EB). Was die Leistung des robusten Standardfehlers betrifft, so kam es in allen Simulationseinstellungen zu einer Überschätzung der wahren Varianz und damit zu einer Überdeckung des 95%-Konfidenzintervalls. Der selbst entwickelte neue Varianzschätzer führte zu einer verbesserten Abdeckung des 95%-Konfidenzintervalls für alle Methoden.
Zusammenfassend lässt sich sagen, dass wir den Forschenden im Bereich Public Health vorschlagen, in ihren Auswertungen, EB für den Zielschätzer ATE, EB oder CBPSJ für den Zielschätzer ATT und OW, MW und GMMW für den Zielschätzer ATO zu verwenden. Für die Varianzschätzung des Effektschätzers sollte bei allen Methoden die Verwendung des neu vorgeschlagenen korrigierten robusten Varianzschätzers gegenüber dem robusten Varianz-schätzer bevorzugt werden.

Summary

Comparison of different propensity score methods for estimating treatment effects in non-randomized studies with survival data.

Observational studies are frequently used in public health research, due to ethical or financial complications for conducting randomized trials. However, observational data analysis come with increased challenges in comparison to randomized trials especially regarding the risk of confounding bias. In the last decades, many methods have been developed for the analysis of observational data. One of these are the propensity score (PS) methods. In the last decade many methods using the PS have been developed, which complicates the choice of the best method for the applied researcher. Simulation studies can provide a guidance on which methods perform best in specific settings. However, there is a lack of simulation studies regarding PS usage in scenarios with survival outcome, which is frequently used.
The overarching aim of this study is to fill this gap, and provide clear guidance for the applied public health researcher regarding the best choice of PS method. Therefore, a detailed simulation study was conducted with simulation settings motivated by information extracted from published PS analyses found by systematic research on 'PubMed'. In addition, a new PS method, the generalized method of moment matching weight (GMMW), was also proposed along with a corrected variance estimator, which is necessary for an appropriate calculation of the confidence interval.

The results of the simulation study showed that all methods (overlap weights (OW), matching weights (MW) and GMMW) estimating the average treatment effect in the overlap population (ATO) performed best with unbiased results in almost all settings and the smallest standard error of the estimator. Comparing the methods, similar performances were found regarding bias and standard error of the estimator. Across all methods estimating the average treatment effect (ATE) and average treatment effect of the treated (ATT), entropy balancing (EB) was found to perform best. Further, the performance of the standard method inverse probability weighting (IPTW) is poor for estimating the ATE but reasonably well (and similar to EB) for the ATT. Regarding the performances of the robust standard error, an overestimation of the true variance and thus, an over-coverage of the 95% confidence interval occurred in all simulation settings. The self-developed new variance estimator resulted in improved coverage of the 95% confidence interval for all methods.
In conclusion, for the applied public health researcher, we propose to use EB for target estimand ATE, EB or CBPSJ for target estimand ATT and OW, MW and GMMW for target estimand ATO. For the variance estimation of the effect estimator, the usage of the corrected robust variance estimator should be preferred over the robust variance estimator for all methods.

Abbreviations

PS = propensity score

IPTW = inverse probability weighting

IPTWST = stabilized inverse probability weighting

MW = matching weights

OW = overlap weights

EBAL = entropy balancing

VBAL = variance balancing

CBPSJ = just identified covariate balancing propensity score

CBPSO = overidentified covariate balancing propensity score

GMMW = generalized method of moment matching weights

ATE = average treatment effect

ATT = average treatment effect of the treated

ATO = average treatment effect of the overlap population

Contents

1 Introduction

In the field of public health research, the effort to establish causal relationships among interventions, exposures, and outcomes has long been a cornerstone of improving our understanding and enhancing population well-being [1],[2]. In complex real-world scenarios, researchers often encounter challenges that make direct assessment of causal effects difficult [3],[4]. The possibility of confounding variables, selection bias, and uncontrolled factors can impede the direct interpretation of results and lead to inaccurate conclusions that may ultimately impact health policy decisions and resource allocation [5]. Therefore, the development of robust statistical methods is necessary, and the choice of the best method is essential.

Propensity score methods are a sophisticated approach to mitigating confounding bias and improving causal inference in public health research [6]. These methods provide a way to overcome the inherent limitations of observational data, where randomization is often impractical or impossible due to ethical reasons [7]. By systematically accounting for imbalances in covariates and creating a quasi-experimental setting, propensity score methods promise to isolate treatment effects and illuminate the true effects of interventions [6].

The concept of PS methods is to use a two steps procedure for treatment effect estimation. First, a pseudo-population is created in which the values of pre-treatment variables do not differ between treatment groups. Second, treatment effect estimation is conducted on the unconfounded pseudo-population instead of the original population. This methodology enables the estimation of causal effects from observational data and makes PS methods an important tool in the arsenal of public health research.

In the following, the main differences between RCTs and observational studies are outlined, in particular the challenges in estimating the causal effect of an intervention in both designs. Subsequently, the PS method and especially PS weighting will be presented. Hereby, the focus is on underlining how PS methods can be used for causal effect estimation. Finally, the different approaches in PS weighting are considered and embedded in the current state of research.

1.1 Randomized trials vs. observational studies

For answering a specific research question in public health, two main classes of study types exist: randomized trials (RCT) and observational studies [7]. As the major difference between these two groups, the intervention in an RCT is actively given to a random set of patients in the cohort, while, in observational studies, it is only observed which patient received which intervention [8]. Although the difference may sound minor at first glance, it has critical consequences on estimating the causal treatment effect.

The design of RCT implies that balanced distributions of all pre-treatment variables are given, including those variables that are not observed or recorded [9]. Thus, for a sufficient sample size, the in-treatment group differences of all pre-treatment covariate values should be small. In contrast, observational studies run the risk that the pre-treatment variables value sometimes show considerable differences between the treatment groups. Thus, the direct calculation of the treatment effect is biased [7].

For a better understanding of the utility of an RCT, the distribution of pre-treatment variables on the one hand, and the actual values in the dataset on the other hand, will be examined. For simplicity, it is assumed that the dataset contains only one pre-treatment variable, namely gender. Then, the term balanced distribution of gender is used if the probability of being a woman in the treatment group is the same as in the control group. Instead, the term the values of gender are balanced is used when the actual proportion of women in the treatment group is equal to the one in the control group. It is important to recognize that a balanced distribution does not necessarily imply balance of the actual values in the dataset. All that is true, is that for equal covariate distribution between intervention groups, the values of the covariate should become more similar as the sample size increases. In the end, however, only the balance of the covariate values counts, while the balance of its distribution is only used as a tool to achieve this.

A drawback of observational data in comparison to RCTs is, that the treatment effect estimator can only be adjusted for differences in observed pre-treatment covariates, but this is impossible for any unobserved pre-treatment covariate [10]. Thus, to reduce the risk of confounding bias, an important task in designing an observational study is to identify potential confounding factors, e.g., through literature research [11] optimally before data collection. However, there is always the risk of missing confounding variables, because either a variable was not accessible or not known as a confounder at the time of the study. To conclude, RCT

are the gold standard for assessing intervention effects as they offer the highest level of evidence for causal effects [12].

Although RCTs are the gold standard for causal inference, there are still good reasons for researcher to use observational studies. First, due to the high cost in terms of time and money, the sample size in RCT will generally be small in comparison to observational studies [13]. A small sample size in an RCT increases the risk of larger differences between treatment groups in observed and unobserved pre-treatment variables values and within this the risk of confounding bias. Second, the duration of a randomized trial is often limited in time and statements regarding long-term effects may therefore not be possible. Third, the effect in an RCT tends to be estimated under ideal conditions among highly selected populations, whereby the populations used in observational studies tend to be closer to the real-world setting [14]. Fourth, there are research questions that RCTs are not feasible to answer for ethical reasons, and in these cases observational studies are the only choice. For example, the impact of aircraft noise on the quality of people's life should be examined. In this case, an RCT is obviously not possible for ethical reasons.

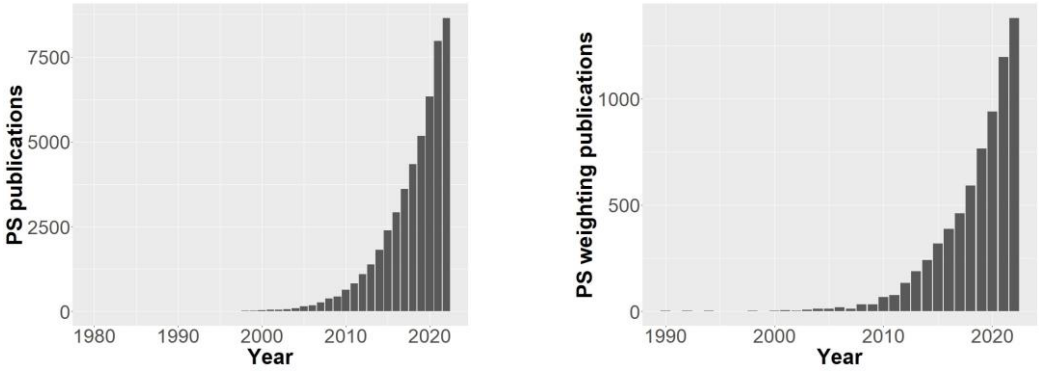
Many studies have examined how treatment effect estimator of RCT differ to those of observational studies based on the same research question. First studies were conducted in the 1970s and 1980s and found observational studies inflate positive treatment effect in comparison with RCTs [15],[16],[17],[18]. However, later studies conducted in the 2000s came to the conclusion that there is little evidence of systematic differences between both methods [19],[20]. They argue that observational studies used in the earlier comparisons were methodologically weaker and that less statistical methods had been available for the analyses at that time. For example, the conceptual PS methods paper was published in 1983 [21]. An actual comparison in 2014 found less evidence for differences in treatment effect estimates between RCT and observational studies. The authors stated that publications in which RCT and observational studies draw different conclusions regarding the treatment effect are easily remembered, while those revealing effect estimators to match are easily forgotten [22].

1.2 Propensity score methods

Propensity score methods are a relative new method to account for confounding bias in the treatment effect estimator and thus are especially helpful for the analysis of observational data. The conceptual idea of the PS was developed by Rubin and Rosenbaum in 1983 [21].

Based on their results, several new methods for using the PS to adjust for differences in pre-intervention variables were developed, which resulted in an increased usage in applied research since the early 2000s (see Figures 1a, 1b). The idea of PS methods is to remove confounding bias by creating a pseudo-population in which the values of all observed pre-treatment covariates are balanced between treatment groups. This pseudo-population is then used in a second step to estimate the intervention effect.

Overall, PS methods pursue two competing goals. On the one hand, the pseudo-population is intended to be as similar as possible as the pre-treatment covariates values to avoid confounding bias in treatment effect estimation. On the other hand, the size of the pseudo-population should be as close as possible to the one of the original population, because higher sample sizes lead to treatment effect estimators with lower variances and, thus, smaller confidence intervals [23].



(a) (b)

Figure 1: Number of publications found on PubMed for the phrases ‘propensity score’(PS) (a) and ‘propensity score weighting’ (b).

In total, four main PS approaches to remove confounding bias exist: PS matching, PS weighting, PS stratification, and PS adjustment. In this work, the focus will be set on PS weighting including also a brief introduction to PS matching for better interpretation of the conceptual ideas and challenges of PS methods. Thus, I will refer to PS methods for the group of PS weighting methods or PS matching methods in in the following.

1.2.1 Subpopulation

An important task for researchers using PS methods is to define a target population the treatment effect should be estimated for. PS methods will then generate a pseudo population in which the values of the pre-treatment covariates are similar to those of the preselected target population. There are three broad target populations with each having its own treatment effect that does not necessarily need to be identical to the others. The first one is the total population represented by the complete cohort with the resulting treatment effect called the average treatment effect (ATE). In this case, the values of the pseudo-population in each treatment group need to mimic the values of all samples in the cohort. The second population is the population of treated patients for which the term average treatment effect of the treated (ATT) is used for the treatment effect. Hereby, the values in the pseudo-population should be as similar as possible to the values of pre-treatment covariates in the treatment group. The third population is the one of best overlap between pre-treatment covariates, which effect is named the average treatment effect of the overlap (ATO) in which the values of the covariates are moved into an area of greatest overlap for each treatment group. For this target population, it is impossible to exactly define how the values in the pseudo-population should look like, as the different methods choose slightly different regions of best covariate overlap. Li et al. [15] noted that the overlap pseudo-population is the one being closest to an RCT. However, the right choice of the target population always depends on the specific research question. As a hypothetical example, the effectiveness of a smoking cessation program on reducing smoking rates and improving health outcome should be assessed. This program includes counseling, nicotine replacement, and support groups. The study population consists of smokers either having participated in the program (intervention group) or not (control group). In this case, the researcher would want to estimate the ATT if it is assumed that only individuals with covariate values similar to those in the intervention group might participate in such a smoking cessation program. The ATE would be the desired estimate if the interest is in the intervention effect in the complete study population (for example if it is mandatory to participate in such a program). Lastly, the ATO should be chosen if the interest is to mimic a randomized trial, thus focusing on individuals with similar probability of being in the program or not.

In the next sections, the following notation will be used $T \in \{0, 1\}$ for the binary treatment variable, $X \in R^m$ is the covariate vector of the m pre-treatment covariates and Y will be used as the outcome variable.

1.2.2 Propensity score matching

The idea to match each sample of one intervention group to a sample in the other intervention group that is similar in its pre-treatment covariates was developed early by researchers [24],[25],[26]. All samples without matching partners are then excluded from treatment effect estimation. Thus, a pseudo-population is derived – the matched population – with similar pre-treatment covariate values between treatment groups. However, in this approach, the critical point is the definition of similarity between patients for the m dimensional pre-treatment covariate vector. Therefore, Rosenbaum and Rubin presented a highly noted concept in which they showed that, instead of balancing all covariates, it is sufficient to balance the PS (π). The latter is defined as the probability of being in the intervention group given a set of pre-intervention covariate values

$$\pi := P(T = 1 | \boldsymbol{X}=\boldsymbol{x}) \qquad (1.1)$$

[21]. This reduces the number of variable values to balance from m to 1, but it brings along another problem, namely the requirement to estimate PS. Rosenbaum and Rubin proposed to use a logistic regression approach for PS estimation [21], which is still the standard approach nowadays. Recent studies compared the performance of PS matching with PS being estimated by either a logistic regression model or a machine learning approach and found they were very similar [27].

The competing goals pursued by PS matching and shared by PS weighting are highlighted by the idea of PS matching. When matches are restricted to be extremely similar, a small pseudo-population with high similarity is obtained. Conversely, the consequence of including patients with low similarity results is a larger pseudo-population (excluding less samples) that is more different in regard of the pre-treatment covariate values.

1.2.3 Propensity score weighting

Another approach to create a pseudo-population with balanced pre-treatment covariate values is the PS weighting method. PS weighting was first developed in the early 2000s and

has fast gained attention in science (see Figure 1(b)). The main idea of PS weighting for creating a pseudo-population with balanced pre-treatment covariate values is that a weight $w_i > 0$, $i = 1, ..., n$ is calculated for each patient, which reflects the importance of each sample. Hereby, larger weights indicate higher importance. For example, a weight w=2 means that the corresponding participant is counted twice.

Moreover, PS weighting can be seen as a generalization of PS matching in which the weights are restricted to $w_i = 1$ if a matching partner was found and the sample is alternatively weighted as $w_i = 0$. Beyond that, PS matching can be only used to estimate the ATE or the ATT, but ATO estimation is not possible [28]. In contrast, PS weighting can be used to create a pseudo-population with any pre-treatment covariate distribution [13].

There are three main approaches to calculate the weight vector **w**, which reflect the difference in balance definition (balance of covariate values vs. balance of covariate distribution). The first approach, the modelling approach, tries to balance the distribution of all observed pre-treatment covariates. The second approach, the balancing approach, tries to balance some aspects of the values of the observed pre-treatment covariates. Finally, the hybrid approach tries to balance the distribution of all observed pre-treatment covariates as well as some aspects (predefined set of moments) of the observed pre-treatment covariate values.

There are some main differences between the approaches and their methods in term of achieved balance and the target estimand (see Table 1). First, the modelling and the hybrid approach balance the pre-treatment covariate distribution between treatment groups in case the assumed PS model is correct. This results in the asymptotic balance of all pre-treatment covariate values. Second, the balancing and some hybrid approach methods (CBPSJ, GMMW) result in exact balance of pre-treatment covariate moments. Third, only hybrid (GMMW) and modelling approach methods (OW, MW) can assess the highly relevant overlap population, whereas balancing approach methods cannot.

To assess the performance of a PS weighting method, it is necessary to measure how well the objectives of the PS method, namely the similarity between the values of the pretreatment covariates in the weighted pseudo-population and the sample size in the pseudo-population, have been achieved. The discussion has so far focused on balancing the pre-treatment covariates values. However, the specific definition of balance of the pre-treatment covariates

values has not been stated. For the balance task, several measurements exist of which some compare the average value of each covariate between treatment groups, such as the standardized difference [29] or the z-difference [30], while others

Approach	Method	Exact balance	Balance distribution	Target estimand
Modelling	IPTW	no	yes†	ATE, ATT
	IPTWST	no	yes†	ATE, ATT
	MW	no	yes†	ATO
	OW	yes*	yes†	ATO
Balancing	EB	yes	no	ATE, ATT
	VB	yes	no	ATE, ATT
Hybrid	CBPSJ	yes	yes†	ATE, ATT
	CBPSO	no	yes†	ATE, ATT
	GMMW	yes*	yes†	ATO

Table 1: Difference between the pseudo-populations created by the different propensity score weighting methods. *Only if the propensity score was estimated via a logistic regression model. †Only if the propensity score model is correctly specified. IPTW= inverse probability weighting, IPTWST = stabilized inverse probability weighting, MW= matching weights, OW = overlap weights, EB = entropy balancing, VB = variance balancing, CBPSJ = just identified covariate balancing propensity score, CBPSO = overidentified covariate balancing propensity score, GMMW = generalized method of moment matching weights, ATE = average treatment effect, ATT = average treatment effect of the treated, ATO= average treatment effect of the overlap population.

compare the whole distribution of each covariate, e.g., the Kolmogorov-Smirnov test [31] or graphical comparisons like side-by-side boxplots [32]. To analyze the sample size in the weighted pseudo-population, the effective sample size is frequently used [23]. The ESS represents the size of an unweighted sample that results in the same precision as the weighted sample [33]. For a weight vector $\boldsymbol{w} \in R^n$ calculated by PS weighting, the ESS is defined via

$ESS(\boldsymbol{w})$: $\frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2}$. The ESS is equivalent to the raw sample size (n) in the unweighted case (w_i = 1, i = 1,..., n).

1.2.4 Propensity score weighting vs. classic regression

For assessing the treatment effect, a regression approach is frequently chosen. In this approach, the outcome of interest is taken as the dependent variable and the intervention variable and all potential confounder variables are included as independent covariates. Hereby, the choice of the regression model depends on the class of the outcome variable, e.g., a Cox model for survival outcome or a logistic regression model in case of a binary outcome.
The usage of PS weighting methods has several advantages in comparison to the regression approach. First, in PS weighting, balancing the pre-intervention covariates is independent of the outcome as weight calculation is independent of the outcome [34]. Thus, PS weighting can be seen as part of the study design and independent of treatment effect estimation [35]. In contrast, adjusting for confounding and treatment effect estimation is conducted in the same step in the regression approach [36]. This is problematic as it might lead scientists to fit several models with different sets of confounders until they reach the desired or expected answer [37].

Second, a general problem of the regression approach is that it always estimates a treatment effect, even when treatment groups are completely different [37]. In contrast, PS weighting identifies that it is impossible to balance the pre-treatment covariates adequately and, thus, treatment effect estimation is impeded.

Third, PS weighting is conceptual better suited for modelling the effect in cases of rare outcome events. As an example, 500 observations of a binary outcome variable may include only 20 events. If a logistic regression approach for treatment effect estimation is used, it is recommended to have at least 10 [38],[39] or even 20 [40] events per variable included in the model. Thus, even with 10 events per variable only one pre-treatment variable can be included in the logistic regression model as it is necessary to include the treatment variable. In such a situation, the researcher has the choice of either confounding bias due to pre-treatment covariates not included in the model or of having an unstable treatment effect estimator as the model contains too many covariates. In PS weighting, that problem is avoided as the treatment variable is the only independent variable in the outcome regression model.

Moreover, there is a fundamental difference between PS weighting and the regression approach, namely that they estimate different effects. The regression approach adjusted for pre-treatment covariates estimates a conditional effect, that is the average effect of treatment on the individual. The PS weighting approach instead estimates a marginal effect, which is the average effect of treatment on the population [41]. For collapsible measures, the two effects are the same, but for non-collapsible measures, such as the odds ratio used in logistic regression model or the hazard ratio as used in the Cox model, they are different. Hereby, the marginal effect is generally closer to the null effect than the conditional effect. The choice of a conditional or a marginal effect estimator depends on the research question, but it should be noted that randomized controlled trials estimate the marginal effect. Thus, whenever researchers want to answer the same research question as in RCT, the regression approach should not be used if the effect measure is not collapsible.

Finally, another option is to combine PS weighting and the regression model [42]. In this approach, PS weighting is first used to calculate weights that balance the pre-treatment covariates between treatment groups. In a second step, a weighted regression model is used, in which all pre-treatment covariates are included as independent covariates. As long as either the PS model or the outcome regression model is correctly specified, the resulting treatment effect estimator is unbiased. However, in this approach equivalent to the standard regression approach, treatment effect estimation and adjustment for confounding variables are done in the same step and all the drawbacks previously mentioned still hold. Likewise, this approach also estimates a conditional treatment effect rather than a marginal.

1.2.5 Examples of propensity score methods in public health

In this chapter, I present three examples of studies using PS methods in public health research. Results of this study are currently prepared for publication.

The first example is a Chinese study in which the relationship between retirement and health was analyzed. Observational data were taken from the 'China Health and Retirement Longitudinal Study' (CHARLS) cohort [43]. Two different PS matching methods were used to adjust for potential differences at baseline. The set of variables, which was adjusted for, contained standard patient information like age and sex as well as emotional support

(frequency at which respondents saw their children), health behavior (e.g., smoking (yes/no), chronic disease (yes/no)), and social characteristics (medical insurance (yes/no)).
The authors found an odds ratio of 0.78 [95%CI: 0.65-0.94, P = 0.026] for self-reported physical pain and 0.76 (95% CI: 0.62-0.93, P = 0.023) for depression comparing retired to working people [44]. These results suggest that retirement benefited health, which is an important aspect for political decision making. On the one hand, early retirement causes costs that need to be paid by the public. On the other hand, deteriorating health also increases public costs. Thus, both aspects need to be taken into account by decision makers.

The second example is a US study in which the impact of an enriched medical home intervention using community health workers on immunization adherence among young children was analyzed [45]. Vaccination of children is very important for several reasons. First, the vaccination protects children from potentially life-threatening diseases [45]. Second, a high vaccination rate in a community also protects non-vaccinated children [45]. For their study, the scientists analyzed the data of 311 children with 110 individuals belonging to the intervention group and 201 children being in the usual care group. The authors adjusted for potential in-group-differences by using inverse probability of treatment weights in which they adjusted for the variables: gender, race, maternal education, living situation, insurance, mother's health rating, prenatal care, maternal depression screening, mother's report about whether adults in neighborhood care about her child, frequency of reading, number of books read, child care use, and car ownership.
The study found an increase of 20.9% [4.6%, 37.2%] for newborn and 16.8% [4.1%, 29.5%] for infants in up-to-date immunization proportions when comparing the intervention group to the usual care group [45].

The third one is the currently ongoing '10,000 Steps' Duesseldorf project [46]. This study examines how complex interventions to promote physical activity influence physical activity. Physical inactivity is the fourth largest risk factor for mortality worldwide and is one of the main drivers contributing to the etiology of noncommunicable diseases, such as type 2 diabetes, cardiovascular diseases, and certain cancers [47]. Currently, only 43% of woman and 48% of men in Germany meet the WHO and American College of Sports Medicine

recommendations for physical activity [46]. Therefore, an increase in physical activity is important for public health.

To examine the intervention effect, 400 people in the intervention city Duesseldorf and 400 people in the control city Wuppertal were randomly selected [46].

PS weighting is intended to be used to adjust for potential baseline differences in the control group and the intervention group. More concrete matching weights are planned to be used to adjust for the number of steps taken before study entry, age, gender, level of education, and household income.

In all three examples described above, it is not possible to perform randomized studies to answer the research questions due to ethical reasons. However, PS methods can be used to adjust for potential in-group differences at baseline and result in a more efficient estimation of the intervention effect.

1.3 Aim of this thesis

In the previous chapters, I showed that PS methods play an important role in public health research. The advantages and disadvantages of observational studies compared to controlled (randomized) trials were presented. Furthermore, it was demonstrated to what extent PS methods, especially PS weighting methods, can be valuable tools in the analysis of observational studies.

Different approaches to calculate these weights can be grouped into three main categories: modeling, balancing, and hybrid. Researchers need to choose the most suitable methods for their analyses. To investigate the weighting methods with highest performance, simulation studies have been conducted, but most of these focused either on the modelling or the balancing approach. The studies that consider both approaches usually limit themselves to using IPTW as the only modeling approach method or EBAL as the only balancing approach method. To the best of our knowledge, there is no publication that examines a larger set of methods from all approaches with a survival time outcome. Further, the choice of the parameters used in these simulation studies did not necessarily mimic real world data.

Therefore, we conducted an extensive simulation study with a larger set of methods for all three approaches. Aiming for maximal practical relevance, settings in the simulation were informed by published studies that applied PS methods to real data. The detailed findings and

methodology of the simulation results are presented in the published paper ‘Balancing versus modelling in weighted analysis of non-randomized studies with survival outcomes: A simulation study’.

2 Balancing versus modelling in weighted analysis of non-randomized studies with survival outcomes: A simulation study.

Filla, T., Schwender, H., Kuss, O. (2024), Statistics in Medicine. 43(17), 3140-3163.

3 Discussion

3.1 Main findings

This primary goal of this thesis was to provide clear guidance for the applied researcher regarding appropriate methods when applying PS methods to survival data analyzed by a Cox regression model. For maximal practical relevance, a large simulation study was performed which settings mirrored published PS analysis sets with respect to outcomes as well as covariates as close as possible. All major analysis methods were included. The simulation study drew four main conclusions. First, none of the three approaches (modelling, balancing, hybrid) outperformed the others in all settings. More precise, EB had the best performance across all methods estimating the ATE or ATT for most simulation settings. In contrast, VB as the other balancing approach method resulted in large bias for ATE and ATT in some settings. Second, the methods estimating the ATO performed best across all settings regarding the bias and standard error of the estimator. This was expected as these methods use the overlap population as the target population and weight the distribution of the covariates into an area of best overlap. Third, the performance of IPTW was poor for the target estimand ATE, but reasonably well and similar to EB for target estimand ATT. Fourth, the standard robust variance estimator overestimated the variance for all methods and target estimands, which resulted in an over-coverage of the 95% confidence interval for all methods. In comparison, the corrected robust variance estimator resulted in improved variance estimation for all methods.

3.2 Detailed findings

Both methods of the balancing approach had difficulties in numerical stability in settings with low sample sizes. EB failed to converge in 1.8% of the simulated data sets for target estimand ATT and in 0.3% of the simulated data sets for ATE. VB had even stronger convergence issues in 2.8% of the simulated data sets for ATT and in 4.9% of the simulated data sets for ATE. For both methods, non-convergence occurred mainly in simulation settings with bad covariate overlap and low sample size, which is not surprising as these are the most challenging settings for all methods. The convergence problem did not only occur in our simulated data sets but

also in the example data set. Potential explanations for the non-convergence are the lack of a solution to the balancing equations and the failure in finding the solution. At least for some settings the latter case was true as EB or VB found a solution, whereas the other method failed. According to Wang et al. [48], these convergence issues are a disadvantage of all methods that aim for exact balance in comparison to those methods with approximate balance as only the latter guarantee convergence.

Comparing the results of the simulation study for treatment effect estimation to the results of previous published simulation studies, the results match in most of the cases.
This work suggests good performances of IPTW and EB for the target estimand ATT. This is in line with Amusa et al. [49], who conducted a simulation study with a similar design as the one in our analysis. The authors compared the performances of IPTW and EB in a setting with ten covariates, a logit model with parameters very similar to the ones used in this work in the bad overlap settings, and a survival outcome analyzed via a Cox regression model. They also used the robust variance estimator for coverage estimation. Equivalent to our study, they found a similar bias among both methods, but the EB had superior performances for the mean squared error. In our work, the relative efficiency comparing EB and IPTW with target estimand ATT was similar, and in settings with a correctly specified PS model even slightly lower for the IPTW method. The reason for this could be found in the higher effective sample size for IPTW.

We found EB to perform excellent across both target estimands (ATE and ATT) and all settings. In our study, EB outperformed the other methods for target estimand ATE and was one of the top performing methods for ATT. This fits to the results of previous simulation studies in which EB was found to perform best across all compared methods [50],[51],[52].

Our results suggest a mixed performance of VB. On the one hand, VB showed the worst performance across all methods with biased estimator for ATE and ATT in some settings even when the PS model was correctly specified. On the other hand, promising results were found regarding the effective sample size and empirical standard error with similar or even better results than EB. In a previous simulation study with a continuous outcome and target estimand ATE, Chattopadhyay et al. [23] found VB outperforming IPTW in all scenarios. However, in comparison to linear regression for estimating the hazard ratio, which was used in our study,

it is not sufficient to balance only the first moment of all covariates, but the whole covariate distribution. Therefore, bias might be introduced by remaining imbalances in higher covariate moments or interactions. In general, this could be solved by also balancing these. However, in real world settings it remains unclear to which degree it is necessary. Increasing the number of moments to balance also increases the variability of the weights and thus results in larger standard error for the estimator. This was shown in a simulation study comparing IPTW, CBPSJ, and EB to methods using machine learning techniques, which balanced the complete distribution of covariates including interactions and higher moments [52]. The methods balancing the complete distribution showed good performances in general, but for settings with many covariates or a low number of samples, large variances of the estimators were found [52].

For estimating the ATE, the methods CBPSJ and CBPSO created pseudo-populations with covariate distributions that did not necessarily match those of the whole cohort. The average values of the covariates in the pseudo-populations were found to be different to those of the original unweighted cohort. As a possible explanation, per construction CBPSJ and CBPSO do not necessarily weight the covariates of the pseudo-population to the whole cohort. In fact, they rather weight the covariates to an area of good covariate overlap. For target estimand ATT, this problem does not exist as the weights for the treatment group are fixed and the control group is weighted towards the treatment group.

The performance of all methods estimating the ATO was very good in all settings and results were similar across the applied methods. This is in line with previous simulation studies that also showed good performance in balance and higher preserved sample sizes compared to IPTW [13]. Also, another study comparing the performances of OW and MW found both methods to perform well with small differences across methods [53]

The self-developed method GMMW outperformed MW in terms of achieved balance measured via the absolute weighted standardized difference. This was expected as GMMW was designed to remove any imbalance of the covariates moment, which is exactly what is measured by the weighted standardized difference. When comparing the performance of GMMW in terms of bias or the standard error of the estimator, the results were very similar

to MW. However, as the bias is related to the achieved balance and the strength of the covariable outcome effect, probably an improvement in bias for stronger confounder outcome effects (effect was fixed $\beta_C = 0.5$) would be found. This is considered to be important for two reasons. First, if a researcher prefers MW to OW method due to the asymptotic equality to PS-matching (and within this more intuitive interpretation), GMMW offers better balance of the pre-treatment covariates. Second, in case of more than two treatment groups, also the OW method does not provide an exact but only an asymptotically balance of the covariates. Thus, in this case both GMMW and the new method generalized method overlap weights (GMOW) could be calculated. I expect that GMMW and GMOW method would then increase efficiency for treatment effect estimation with more than two treatment groups. Indeed, first results confirmed this assumption and we found an improved performance for many settings. The manuscript is currently being prepared for submission.

As another important finding of the present study, the coverage of the 95% confidence interval using robust variance estimator was too high independent of the method used or the target estimands. This matches to the result of Austin [54] who also found an over-coverage in a simulation study with survival outcomes and IPTW. The standard robust variance estimator has the disadvantage of assuming the weights values to be given instead of being estimated. Therefore, the variability that is introduced by estimating the weights is added to the standard error of the treatment effect estimator of the final Cox-model.

The corrected robust variance estimator takes into account the uncertainty of weight estimation and was originally proposed by Shu et al. [55] for IPTW with target estimand ATE and was expanded to IPTW with target estimand ATT, OW, MW, and GMMW in this work. The corrected robust variance estimator performed better in all settings across all methods than the robust variance estimator.

The corrected robust variance estimator was found to give adequate coverage for IPTW with target estimand ATE for large sample sizes or good covariate overlap. However, it underestimates the true variance in settings in which sample size and treatment prevalence are low. This is in line with the results of Shu et al [55] who found similar results for low sample size, high censoring proportion, and low treatment prevalence. Amusa et al. [49] conducted a simulation study using a survival outcome and IPTW weights with target estimand ATE. They

found an over-coverage of about 97.5% in settings with low treatment prevalence, while the coverage was reduced to about 92.5% in settings with high treatment prevalence. However, this under-coverage was most likely related to a biased estimator in these settings and, thus, their results match the ones presented in this work.

For the performance of the coverage using the corrected robust variance estimator with IPTW and target estimand ATT, adequate coverage in all settings was found, including those with low sample sizes. When using the corrected robust variance estimator for OW, MW, and GMMW, the coverage appeared to be adequate for all settings with a null treatment. However, in those settings with a non-null treatment effect, the true variance was slightly underestimated. The coverage results for the balancing approach methods using the robust variance estimator were similar to these of the other approaches with an increased coverage in all settings, although both methods calculated the weights directly and avoided estimation of the PS.

3.3 Limitations

This work has some limitations that might have affected our results and the conclusions we draw. First, more weighting methods exist that would have been worth to discuss here. For example, the *energy balancing* method is another promising approach that fits to the balancing approach methods and minimizes the energy distance between treatment groups [56]. Nevertheless, the set of methods to use is limited in every simulation study and best was done to include different methods for each of the three approaches.

Further, some authors propose to combine the regression model and the use of the PS [57], which we did not do in our study. Hereby and equivalent to the modelling approach methods, the PS is estimated in a first step. In contrast to the modelling approach, in the second step of the combined approach, the pre-intervention covariates are added as independent variable in the weighted outcome regression model, while in the traditional modelling approach only the treatment variable is used. In this approach, the researcher must estimate the PS correctly or define the covariate outcome association correctly to obtain valid results. The approach has been criticized by other authors, because it cancels out the advantage of the PS being part of the study design and independent of the outcome as the treatment effect depends on how the covariate outcome association is modelled. Thus, it might temp to work towards the

desired or anticipated result. Therefore, we decided not to include this approach in our simulation.

In addition, all methods were run with their default settings, and it is possible that some methods might show improved performances by carefully tuning their settings. For example, trimming the weights can be conducted or samples with extreme PSs could be deleted in IPTW methods. However, the researcher should be aware that both actions change the covariate distribution and as a consequence the final treatment effect estimand. Hereby, trimming weights is equivalent to making the covariate distribution more similar and within this weighting the covariate distribution to an area of best overlap. However, the choice of the weighting method should depend on the specific research question. Thus, trimming is not adequate if, for example, the ATE should be estimated. If the focus is on estimating the ATO, a previous simulation study showed that overlap weight consistently outperforms trimmed IPTW [58]. Therefore, we decided not to include trimming for IPTW weights. Beyond that, balancing approach methods could be tuned by not exactly fulfilling the balancing constraints, but allowing some level of imbalance. As a disadvantage of that approach, the introduced bias by allowing some imbalance for a specific covariate depends not only on the level of imbalance but also on the strength of the covariate-outcome effect. Thus, it is complicated to set a good cutoff value for the allowed level of imbalance.

The PS was always estimated via a logistic regression model, although other methods, such as machine learning methods, might be suitable candidates. We decided on our procedure for two reasons. First, in the 50 published PS publications analyzed for the design of the simulation study, we found that most of the studies used a logistic regression approach for PS estimation. Second, previous studies did not result in systematically worse performance of logistic regression models in comparison to machine learning algorithms. To be more concrete, one simulation study found machine learning algorithms to slightly improve the performance of IPTW in scenarios with both moderate non-additivity and moderate non-linearity of the PS model [59]. Another study found the results of logistic regression model to be similar to those of neural networks and only slightly worse than those of a random forest model [60]. In a third study, Goller et al. [61] found random forest model to perform worse than logistic regression model.

As another potential limitation, we only used the robust variance estimator and corrected robust variance estimation to estimate the variance. Here, the bootstrap estimator might be another promising approach [60]. Equivalent to the corrected robust variance estimator, the bootstrap estimator accounts for the weights being estimated rather than fixed by incorporating the weight estimation step into the bootstrap procedure. In comparison to the corrected robust variance estimator, the bootstrap estimator has the advantage of being easily applicable to all weighting procedures, while the corrected robust variance estimator needs to be developed for all weights independently. Austin [54] found the bootstrap algorithm to perform well in a variety of settings. Thus, especially for balancing approach methods for which no corrected robust variance estimator exists so far, it might be a promising method. However, the bootstrap algorithm also has some disadvantages. First, it can be very time-consuming, especially for the methods in which the weight estimation step is computational challenging (e.g., balancing approach methods with large sample size). Second, for datasets with low number of events, bootstrapping can result in bootstrap samples with very little number of events and the outcome model might result in extreme effect estimates making the variance estimator unstable. Such a behavior was observed by Shu et al. [55], who found the bootstrap variance estimator to overestimate the true variance in settings with high censoring rate and low sample size. In summary, it would have been computationally extremely heavy to incorporate the bootstrap algorithm as not only the treatment effect estimation step but also the weight estimations step needs to be repeated. Therefore, we did not incorporate this algorithm into our simulation. Nevertheless, it would be very interesting and worth further investigation to compare the performance of the corrected robust variance estimator and the bootstrap algorithm especially in settings with lower sample size and/or higher censoring rate.

4 Conclusions

In summary, this work clearly demonstrated advantages and disadvantages of different weighting approaches and compared the performances of different variance estimators, which are important for assessing the strength of intervention effects. For the target estimand

ATT, our simulation results suggest to use either EB or CBPSJ. For target estimand ATE, we recommend the usage of EB. For target estimand ATO, all three studied methods showed excellent performance and, thus, all the three approaches can be recommended.
We recommend the corrected robust variance estimator for variance estimation for methods IPTW with target estimand ATE and ATT as well as for all methods estimating the ATO. For all other methods, we assessed the robust variance estimator only and researchers should be aware that the variance is overestimated for all methods in these cases.
The application of our recommendations to the described public health analyses (chapter 1.2.5) confirmed for the '10.000 steps' [46] study the intended method MW to be adequate. For the Chinese retirement study [44], our study suggests to use OW, MW or GMMW. Finally, for the children vaccination study [45] the authors aimed to assess the whole population (ATE) and therefore, our study suggests to use EB rather than the original used IPTW.

In the future, further research is needed to evaluate the performance of different variance estimation methods and, in particular, to compare the corrected robust variance estimator with the bootstrap estimator. In addition, it would be interesting to evaluate the performance of different PS weighting methods and variance estimators in the case of multiple indication groups and a survival outcome. Finally, it would be of great importance to develop PS weighting methods for continuous indication variables in the overlap population.
While these future directions are important for advancing the field, the results of this thesis already provide significant contributions, which will help to further improve the quality of propensity score analysis in public health studies by providing clear and valuable implications for the applied users.

References

[1] Glass, T. A., Goodman, S. N., Hernán, M. A., & Samet, J. M. (2013). Causal inference in public health. Annual review of public health, 34(1), 61-75.

[2] Hernán, M. A. (2021). Methods of public health research—strengthening causal inference from observational data. New England Journal of Medicine, 385(15), 1345-1348.

[3] Greenland, S., & Morgenstern, H. (2001). Confounding in health research. Annual review of public health, 22(1), 189-212.

[4] Skelly, A. C., Dettori, J. R., & Brodt, E. D. (2012). Assessing bias: the importance of considering confounding. Evidence-based spine-care journal, 3(01), 9-12.

[5] Rabarison, K. M., Timsina, L., & Mays, G. P. (2015). Community health assessment and improved public health decision-making: a propensity score matching approach. American journal of public health, 105(12), 2526-2533.

[6] Crump R. K., Hotz V. J., Imbens G. W., & Mitnik O. A. (2006). Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand. National Bureau of Economic Research.; Technical Report 330: Cambridge: MA.

[7] Chang, T. H., & Stuart, E. A. (2022). Propensity score methods for observational studies with clustered data: a review. Statistics in medicine, 41(18), 3612-3626.

[8] Gilmartin-Thomas, J. F., Liew, D., & Hopper, I. (2018). Observational studies and their utility for practice. Australian prescriber, 41(3), 82.

[9] Faraoni, D., & Schaefer, S. T. (2016). Randomized controlled trials vs. observational studies: why not just live together?. BMC anesthesiology, 16, 1-4.

[10] Lin, Y., Zhu, M., & Su, Z. (2015). The pursuit of balance: an overview of covariate-adaptive randomization techniques in clinical trials. Contemporary clinical trials, 45, 21-25.

[11] Thomas, L. E., Li, F., & Pencina, M. J. (2020). Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. Jama, 323(23), 2417-2418.

[12] Barton, S. (2000). Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. Bmj, 321(7256), 255-256.

[13] Hariton, E., & Locascio, J. J. (2018). Randomized controlled trials—the gold standard for effectiveness research. BJOG: an international journal of obstetrics and gynaecology, 125(13), 1716.

[14] Ross, J. S. (2014). Randomized clinical trials and observational studies are more often alike than unlike. JAMA internal medicine, 174(10), 1557-1557.

[15] Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. Journal of the American Statistical Association, 113(521), 390-400.

[16] Sacks, H., Chalmers, T. C., & Smith Jr, H. (1982). Randomized versus historical controls for clinical trials. The American journal of medicine, 72(2), 233-240.

[17] Miller, J. N., Colditz, G. A., & Mosteller, F. (1989). How study design affects outcomes in comparisons of therapy. II: Surgical. Statistics in medicine, 8(4), 455-466.

[18] Colditz, G. A., Miller, J. N., & Mosteller, F. (1989). How study design affects outcomes in comparisons of therapy. I: Medical. Statistics in medicine, 8(4), 441-454.

[19] Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. New England Journal of Medicine, 342(25), 1878-1886.

[20] Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. New England journal of medicine, 342(25), 1887-1892.

[21] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

[22] Anglemyer, A., Horvath, H. T., & Bero, L. (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database of Systematic Reviews, (4).

[23] Chattopadhyay, A., Hase, C. H., & Zubizarreta, J. R. (2020). Balancing vs modeling approaches to weighting in practice. Statistics in Medicine, 39(24), 3227-3254.

[24] Greenberg, A. (1954). Matched samples. Journal of Marketing, 18(3), 241-245.

[25] Cochran, W. G., & Chambers, S. P. (1965). The planning of observational studies of human populations. Journal of the Royal Statistical Society. Series A (General), 128(2), 234-266.

[26] Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics, 159-183.

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

[28] Baum, S., Ma, J., & Payea, K. (2010). Education Pays, 2010: The Benefits of Higher Education for Individuals and Society. Trends in Higher Education Series. College Board Advocacy & Policy Center.

[29] Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine, 34(28), 3661-3679.

[30] Kuss, O. (2013). The z-difference can be used to measure covariate balance in matched propensity score analyses. Journal of clinical epidemiology, 66(11), 1302-1307.

[31] Sheskin, D. J. (2020). Handbook of parametric and nonparametric statistical procedures. crc Press.

[32] Stuart, M. (1984). Understanding robust and exploratory data analysis.

[33] Shook-Sa, B. E., & Hudgens, M. G. (2022). Power and sample size for observational studies of point exposure effects. Biometrics, 78(1), 388-398.

[34] 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(1), 20-36.

[35] Kuss, O., Blettner, M., & Börgermann, J. (2016). Propensity score: an alternative method of analyzing Treatment Effects: Part 23 of a series on evaluation of scientific Publications. Deutsches Ärzteblatt International, 113(35-36), 597.

[36] Martens, E. P., de Boer, A., Pestman, W. R., Belitser, S. V., Stricker, B. H. C., & Klungel, O. H. (2008). Comparing treatment effects after adjustment with multivariable Cox proportional hazards regression and propensity score methods. Pharmacoepidemiology and drug safety, 17(1), 1-8.

[37] Pattanayak, C. W., Rubin, D. B., & Zell, E. R. (2011). Propensity score methods for creating covariate balance in observational studies. Revista Espanola de Cardiologıa (English Edition), 64(10), 897-903.

[38] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology, 49(12), 1373-1379.

[39] Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. Journal of clinical epidemiology, 48(12), 1503-1510.

[40] Austin, P. C., & Steyerberg, E. W. (2017). Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Statistical methods in medical research, 26(2), 796-808.

[41] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3), 399-424.

[42] Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. American journal of epidemiology, 173(7), 761-767.

[43] Zhao, Y., Strauss, J., Yang, G., Giles, J., Hu, P., Hu, Y., ... & Wang, Y. (2013). China health and retirement longitudinal study–2011–2012 national baseline users' guide. Beijing: National School of Development, Peking University, 2, 1-56.

[44] Peng, X., Yin, J., Wang, Y., Chen, X., Qing, L., Wang, Y., ... & Deng, D. (2022). Retirement and elderly health in China: Based on propensity score matching. Frontiers in Public Health, 10, 790377.

[45] Pati, S., Ladowski, K. L., Wong, A. T., Huang, J., & Yang, J. (2015). An enriched medical home intervention using community health workers improves adherence to immunization schedules. Vaccine, 33(46), 6257-6263.

[46] Fialho, P. M. M., Günther, L., Schmitz, E., Trümmler, J., Willemsen, S., Vomhof, M., ... & Pischke, C. R. (2022). Effects of the Population-Based "10,000 Steps Duesseldorf" Intervention for Promoting Physical Activity in Community-Dwelling Adults: Protocol for a Nonrandomized Controlled Trial. JMIR Research Protocols, 11(9), e39175.

[47] World Health Organization (2010). Global recommendations on physical activity for health. World Health Organization.

[48] Wang, Y., & Zubizaretta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. Biometrika. 107(1), 93-105.

[49] Amusa L., Zewotir, T., & North, D. (2019). Examination of entropy balancing technique for estimating some standard measures of treatment effects: a simulation study. Electron J Appl Stat. 12(2), 491-507.

[50] Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 20(1), 25-46.

[51] Zaho, Q., & Percival D. (2017). Entropy balancing is doubly robust. J Causal Inference. 5(1).

[52] Li, Y., & Li, L. (2021). Propensity score analysis methods with balancing constraints: A Monte Carlo study. Statistical methods in medical research, 30(4), 1119-1142.

[53] Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. Statistical methods in medical research, 29(12), 3721-3756.

[54] Austin, P.C., & Small, D.S (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. Stat Med. 33(24), 4306-19.

[55] Shu, D., Young, J. G., Toh, S., & Wang, R. (2021). Variance estimation in inverse probability weighted Cox models. Biometrics. 77(3), 1101-1117.

[56] Huling, J. D., & Mak, S. (2024). Energy balancing of covariate distributions. Journal of Causal Inference, 12(1), 20220029.

[57] Crump, R. K., Hotz, V. J., & Imbens, G. W., Mitnik O. A. (2006). Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand. National Bureau of Economic Research.; Technical Report 330: Cambridge: MA.

[58] Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. American journal of epidemiology, 188(1), 250-257.

[59] Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3), 337-346.

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

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

Acknowledgments

I want to thank Prof. Oliver Kuß, spokesperson of FC Bosporus, for hiring me and giving me the opportunity to pursue a doctorate on this interesting topic. I am especially grateful for his consistent support, not only for his expert knowledge but also for his encouragement during the challenging times of my doctorate.

I would like to express my gratitude to my co-supervisor, Prof. Holger Schwender, for sparking my interest in statistics and guiding me throughout the last 12 years—first as a Bachelor's student, then a Master's student, and as his student assistant.

I want to thank Prof. Matthias Schneider for giving me the opportunity to work as a statistician at his institute. He had confidence in my work right from the start, and I appreciate the trust he placed in me.

I would like to thank Prof. Jörg Distler for his trust in me and allowing me to work on various collaborations.

I must also thank all colleagues I worked with in the last years. In so many different ways, they made daily office life a lot of fun. I especially want to mention Alex, Lina, Tim A., Anna, Dennis and Lukas.

There are people in life for whom not enough pages exist. To my closest friends, thank you for your constant support, encouragement, and companionship throughout this journey. You all deserve your own section, but as space is limited, I will simply name you here: Ashith, Karsten, Kenny, Michael, Sarah and Saskia. Each of you means the world to me!

To my wonderful girlfriend, Alina, thank you for your unwavering support, patience, and love throughout this journey. I love you dearly, and I truly wouldn't be where I am today without you.

I would like to extend my deepest gratitude to my family- best yogurt-making grandma Margret, my incredible brother Marc, and the best parents in the world Marion and Rainer. Each of you supported me unconditionally and I love all of you from the bottom of my heart.