applied sciences

MDPI

Review

A Comprehensive Review and Tutorial on Confounding Adjustment Methods for Estimating Treatment Effects Using Observational Data

Amy X. Shi [1], Paul N. Zivich [2] and Haitao Chu [3,4,*]

[1] Late Stage Development, Cardiovascular, Renal and Metabolism (CVRM), Biopharmaceuticals R&D, AstraZeneca, Durham, NC 27703, USA
[2] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; pzivich@live.unc.edu
[3] Statistical Research and Data Science Center, Pfizer Inc., New York, NY 10001, USA
[4] Division of Biostatistics and Health Data Science, University of Minnesota at Twin Cities, Minneapolis, MN 55455, USA
* Correspondence: chux0051@umn.edu

Abstract: Controlling for confounding bias is crucial in causal inference. Causal inference using data from observational studies (e.g., electronic health records) or imperfectly randomized trials (e.g., imperfect randomization or compliance) requires accounting for confounding variables. Many different methods are currently employed to mitigate bias due to confounding. This paper provides a comprehensive review and tutorial of common estimands and confounding adjustment approaches, including outcome regression, g-computation, propensity score, and doubly robust methods. We discuss bias and precision, advantages and disadvantages, and software implementation for each method. Moreover, approaches are illustrated empirically with a reproducible case study. We conclude that different scientific questions are better addressed by certain estimands. No estimand is uniformly more appropriate. Upon selecting an estimand, decisions on which estimator can be driven by performance and available background knowledge.

Keywords: confounding; propensity score methods; outcome regression; doubly robust methods; observational data; covariate adjustment

Citation: Shi, A.X.; Zivich, P.N.; Chu, H. A Comprehensive Review and Tutorial on Confounding Adjustment Methods for Estimating Treatment Effects Using Observational Data. *Appl. Sci.* **2024**, *14*, 3662. https://doi.org/10.3390/app14093662

Academic Editor: Pentti Nieminen

Received: 1 March 2024
Revised: 22 April 2024
Accepted: 22 April 2024
Published: 25 April 2024

1. Introduction

Randomized controlled trials (RCT) continue to be the gold standard for evaluating the efficacy and safety of new medical interventions [1]. However, researchers sometimes use observational studies to estimate the effectiveness of treatments and exposures on health outcomes [2]. Direct unadjusted comparisons are misleading when the subjects receiving one treatment differ systematically from the subjects receiving another treatment. For rare diseases, single-arm trials are common due to the impracticability of conducting large randomized trials [3]. Instead, an external control arm is used for comparison under the assumption of no systematic differences across contexts. When there is imperfect compliance, randomized trials no longer estimate the effect of the actual "take up" of a treatment. In per-protocol analyses, the effectiveness of treatments on outcomes can be subject to confounding bias in which those who adhere are systematically different from those who do not; therefore, minimizing potential bias is critical [4,5]. Regulatory agencies have thus issued guidelines for the application of external data in drug development [3,6,7].

There is substantial discussion in the literature regarding estimating the causal treatment effects from observational data via the potential outcome model, in which exchangeability, positivity, and consistency are key assumptions [8–13]. Various ways of adjusting covariates are proposed for causal inference to reduce bias and increase precision (i.e., smaller

variance) of the estimator, such as the traditional regression models, g-computation, propensity score (PS), and doubly robust approaches [11]. Some researchers have made comparisons for certain methods [13,14]. Ding and Li (2018) [15] provided a systematic review of causal inference from the missing data perspective. However, much of the existing work compares those different methods for a single estimand; therefore, different estimands and the relative performance of various estimators have not been reviewed comprehensively. This work provides a review of common estimands and confounding adjustment approaches. We also discuss bias and precision, advantages and disadvantages, and software implementation for each method.

We start with outlining the causal inference framework and discussing related estimands for different populations in Section 2. Section 3 systematically reviews the most common statistical methods for confounding adjustment. We then briefly discuss diagnosis of checking variable balance in Section 4. An example using real data is then illustrated on how to use various estimators and causal methods with different software tools. This paper then concludes with a brief discussion.

2. Causal Framework and Estimands

Assume that there are N subjects and, for each subject, there is a binary treatment indicator A_i ($i = 1, 2, \ldots, N$), where $A_i = 0$ for the control and $A_i = 1$ for the active treatment. The observed endpoint variable is Y_i, and $Y_i(0)$ denotes the potential outcome under the control ($A_i = 0$) for subject i, whereas $Y_i(1)$ is for the potential outcome under the active treatment ($A_i = 1$). There are p covariates $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ for subject i, such as the baseline characteristics, demographic features, risk factors, etc. The covariates can be binary, categorical, or continuous.

The individual treatment effect is the difference between the two potential outcomes for a subject, $Y_i(1) - Y_i(0)$. The causal consistency assumption relates an observed outcome to the potential outcome [11].

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0),$$

which states that the observed response Y_i is equal to the potential outcome with a treatment level that matches the actual treatment level. It is not possible to observe both $Y_i(1)$ and $Y_i(0)$ for a single individual. Therefore, the individual treatment effect is not identifiable, and, instead, we focus on the causal effects averaged over subjects.

A commonly used population-level treatment effect is the average treatment effect (ATE), which is defined as the expected individual difference in potential outcomes as below:

$$\text{ATE} = E[Y(1) - Y(0)].$$

The ATE estimand gives the average effect of treatment in the population and is most relevant when we want to compute an average effect estimate that summarizes the effect for all members in the target population.

The treatment effect may be heterogeneous if it affects individuals differently. In this case, we can divide the population into subsets (e.g., male versus female) and contrast the average effects by subsets. This type of ATE is called the conditional average treatment effect (CATE) and is conditioned on covariates in the subset, as follows:

$$\text{CATE} = E[Y(1) - Y(0)|X = x].$$

In some scenarios, average treatment effects for subpopulations related to treatment are of interest. Two common subpopulations are the subjects in the active treatment group and the subjects in the control group, which are the basis for the average treatment effect for the treated (ATT) and the average treatment effect for the control (ATC), as follows:

$$\text{ATT} = E[Y(1) - Y(0)|A = 1],$$

$$\text{ATC} = E[Y(1) - Y(0)|A = 0].$$

ATT is the estimand that is most relevant when we want to evaluate the effect of a treatment among those who are treated, whereas ATC is most relevant when we want to consider the cost incurred by subjects who are not given treatment. The same approaches can be used to estimate the ATT and ATC. For this reason, we only discuss the ATT hereafter.

Moreover, if a matching method is used to pair subjects from the active treatment group to those in the control group, we have a matched subpopulation that may be different from the target (such as the whole or treated) population. The matching approach creates matched sets of treated and untreated subjects who have similar values of PS. The response variable is then compared between the treated and untreated subjects in the matched set. The corresponding estimand is called the average treatment for the matched set (ATM), as follows:

$$\text{ATM} = E[Y(1) - Y(0)|M = 1],$$

where $M = 1$ for matched observations.

Recently, another subpopulation is considered to be of interest, which is the set with the most overlap in the observed covariates between the control and treated [16–18]. This subpopulation contains subjects that are eligible to be recruited and assigned to either treatment arm with a similar probability, and thus may more closely mimic a population enrolled in a RCT [17]. The estimand for the overlap population is called ATO (Web appendix of Li et al. [18]):

$$\text{ATO} = \frac{E[e(X)(1 - e(X))(Y(1) - Y(0))]}{E[e(X)(1 - e(X))]},$$

where $e(X)$ is the propensity score, i.e., the probability of a subject being assigned to the active treatment arm given the covariates, $e(X) = \Pr(A = 1|X = x)$.

The choice of estimand(s) will depend on the research objectives. If the research question is about the effect on outcomes for all subjects, then the ATE is likely the best choice. If the research goal is to evaluate the effect on outcomes among subjects who received the active treatment, then the ATT would be most appropriate. If the research interest is regarding the subpopulation of those who have equal probability to be in either the active treatment or control group (e.g., the randomized controlled trial), then ATO is the best option.

3. Confounding Adjustment Methods

Once the estimand is selected, the next task is to express the estimand in terms of the observed data, referred to as 'identification'. Identification often relies on the assumptions of causal consistency and exchangeability (or ignorability) with positivity [19,20]. As described above, the causal consistency assumption links the potential outcomes to the observed outcomes. Exchangeability stipulates that the potential outcomes and treatment are independent marginally or conditionally. Conditional exchangeability is mathematically expressed as $E[Y(a) \mid X] = E[Y(a) \mid A = a, X]$ for all values a of treatment. In other words, exchangeability is the assumption that there are no unobserved common causes of the exposure and outcome. For exchangeability to be mathematically well defined, we further assume that the probability of receiving a treatment is non-zero for every covariate combination relevant for exchangeability. This assumption is referred to as positivity and written as $\Pr(A = a \mid X) > 0$ [21].

In an RCT, marginal exchangeability is met by design and covariates are balanced in expectation. However, chance imbalances can still occur, particularly when a sample size is small. Accounting for chance imbalances by covariates strongly predictive of the outcome can provide more precise estimates of causal effects [22]. Here, the goal of covariate adjustment is to improve precision and power in estimating causal effects.

If a study is not randomized, there can be systematic imbalance in covariates for different treatment arms (i.e., marginal exchangeability does not hold). Instead, one can

assume there is a sufficient adjustment set for confounding (i.e., conditional exchangeability). Confounding adjustment can then be achieved via different ways, such as traditional outcome regression, g-computation, PS adjustment, and doubly robust approaches.

Outcome regression methods were first developed to estimate conditional effects accounting for covariate imbalance between treatment arms [23,24]. However, standard adjustment by parametric regression models is sensitive to model mis-specification [8]. G-computation has been proposed as a way to estimate the marginal causal effect using an outcome regression model [25]. G-computation allows for a treatment effect to be different for different levels of the covariates, and it is also relatively robust to model mis-specification if there is no unmeasured confounding [26]. Alternatively, propensity score (PS) methods that use PS in different ways to control confounding include matching [10,27], stratification [28,29], weighting, and conditional adjustment using PS as a covariate [30]. For example, the inverse probability weights (IPW) can be applied to subjects in each treatment arm to balance the covariate distributions, and the comparison is made between the weighted outcomes [17,23,31–34]. Researchers [35] used a few large-scale cardiovascular observational studies to compare the performance of a conventional regression with PS methods. PS is the most widely applied causal inference tool for observational studies and it has theoretical advantages over conventional confounding adjustment using outcome regression. Another option is the doubly robust approach that combines PS methods and outcome regression. One of the doubly robust methods is the augmented IPW estimator, which can provide unbiased estimates if one of the models is mis-specified. We discuss each approach in more detail below.

3.1. Traditional Regression

For traditional regression, a model is fit for the response Y on the treatment indicator A, covariates X, and sometimes their interactions $A * X$. For example, a multiple linear regression for a continuous endpoint is set up as

$$\mathrm{Y}_i = \beta_0 + \beta_1 A_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \cdots + \varepsilon_i,$$

where $(\beta_0,\ \beta_1, \beta_2,\ \ldots)$ are the regression coefficients and ε_i is the random error. The least square estimator of β_1 can be taken as the estimator of CATE.

A logistic regression for a categorical endpoint is constructed as

$$\mathrm{logit}[\Pr(\mathrm{Y}_{\mathrm{i}}|A_i, X_i)] = \beta_0 + \beta_1 A_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \cdots .$$

The difference $\Pr(\mathrm{Y} = 1|A = 1, X = x) - \Pr(\mathrm{Y} = 1|A = 0, X = x)$ can be used as the estimator of CATE when $X = x$.

One concern of fitting a logistic regression is that the number of covariates may be very large compared to the number of events when the event of interest is rare. The rule of thumb is to have at least 10 events of the endpoint for each covariate in the regression, but that rule can be relaxed [36]. The regression model further assumes that the treatment effect measure is constant across the levels of covariates (or confounders) included in the model, but this is not often expected to be the case. Model mis-specification may lead to bias and impact precision in unbalanced designs with treatment effect heterogeneity [37]. With non-linear models, like logistic regression, the coefficient of treatment may not represent the marginal effect due to non-collapsibility [38]. For parameters such as odds ratios (OR), the subgroup-specific conditional treatment effects may be different from the unconditional treatment effect, even in the absence of confounding bias.

3.2. G-Computation

In 1986, a paper by Greenland and Robins demonstrated that, under the previous identification assumptions, a consistent estimate of the unconditional treatment effect can be obtained by using the g-computation formula [26,38,39]. G-computation is a generalization of standardization with respect to the covariates' distribution. G-computation takes the model

from traditional regression and computes $P(Y = 1|A = 1, X_i)$ and $P(Y = 1|A = 0, X_i)$ for all subjects, which are the two predicted probabilities of events for a subject's covariates vector X_i under both treatment and non-treatment. These predications can then be used to estimate the reviewed estimands by taking the corresponding mean. For ATE, we include all subjects in the prediction set; for ATT, only treated subjects are included. G-computation, based on the estimation of the components, allows for a treatment effect to vary by different covariates. More in-depth walkthroughs of how g-computation is applied can be found in the following references [14,39]. G-computation is fairly robust in regard to model misspecification for estimating the marginal (or adjusted population-averaged) risk difference if there is unmeasured confounding is not an option [38,40].

3.3. Propensity Score Methods

Propensity score (PS) methods are increasingly popular in observational studies as an alternative to traditional covariates adjustment via a regression model. PS is one of the most frequently used causal inference methods [9]. The PS methods are a set of statistical tools that seek to balance non-equivalent groups with non-randomized designs. Simply speaking, an individual's PS is their probability to have received a treatment conditional on a set of covariates, i.e., $e(X) = \Pr(A = 1|X = x)$. PS is commonly estimated by the standard logistic regression model. Other methods, such as nonparametric regression, generalized additive models, and machine learning methods, can be used to improve PS estimation [41].

In a randomized trial, the true PS is known by design, whereas, in an observational study, the PS must be estimated. The PS is typically estimated through a logistic regression that incorporates variables that are associated with treatment assignment. Because the PS summarizes all covariates as a single PS variable, it is able to mitigate problems of overfitting for the outcome model [35,42]. The PS approach would not help much if the outcome model is linear; however, tut the PS estimation may reduce overfitting for binary outcome models (especially with rare diseases), as the PS summarizes all other covariates into a single variable for the outcome model. Since the treatment assignment between the active and the control is often well balanced (which is relatively common in practice), one can flexibly model many covariates in the PS model. After the PS scores have been obtained, we use them to estimate the treatment effect. PS-based approaches separate the design and analysis in the sense that the PS model can be developed using only data regarding the covariates and treatment variables. Excluding the outcome from development of the PS estimation can avoid the "fishing expedition" of fitting models until favorable results are obtained [38].

There are various methods of using PS for estimating treatment effects: (1) matching; (2) stratification; (3) weighting; (4) using PS as a covariate in an outcome regression.

3.3.1. PS Matching

This approach creates matched sets of treated and untreated subjects with close PS values. The matching ratio can be different, with one-to-one pair matching being the most common. The response variable then is compared between the treated and untreated subjects in the matching set. The matching can be with or without replacement, and each untreated subject is paired to only one treated subject in matching without replacement, while each untreated subject can be matched to more than one treated subject in matching with replacement. Two common methods of matching are optimal matching and greedy matching. Optimal matching involves forming matched pairs to minimize the average within-pair difference in the PS. Greedy matching selects treated subjects sequentially for matching. For a given treated subject, select the closest untreated subject, even if that untreated subject would better serve as a "match pair" for a subsequent treated subject. Gu et al. [43] compared these two matching methods, concluding that optimal matching performed no better than greedy matching in producing balanced matched samples.

3.3.2. PS Stratification

PS stratification divides a dataset into several strata based on PS scores. A treatment effect is studied in each stratum and an overall treatment effect is computed using a weighted average across all strata. This method compares the outcome between treated and untreated subjects who are similar in their PS and thus also likely to be similar in the distribution of their measured covariates. Stratification uses weights that are proportional to the number of subjects in each stratum and allows for estimating the average treatment effect. The weight is $1/k$ when there are k equal strata. The ATT can be estimated if we weight by the stratum-specific number of treated subjects, while the ATE is estimated if we weight by the sum of stratum-specific numbers of treated and control subjects [28,29].

3.3.3. PS Weighting

PS weighting is another important tool in causal inference that can be implemented in different ways.

(1) Inverse Probability of Treatment Weighting (IPTW)

By IPTW, subjects are weighted by the inverse probability of being assigned to the treatment: $w_i = \frac{A_i}{e_i} + \frac{1-A_i}{1-e_i}$. However, we often need to estimate the weights in observational studies, so e_i would be estimated. Additionally, notice that weights can become quite large when e_i is near zero. For variance estimation, the distinction between the known weights and estimated weights is important [32]. The covariate imbalance between treatment groups is reduced by the weighting approach, and so an outcome can be compared directly between treated and untreated subjects in the weighted data. The IPTW can be used to estimate the ATE (Table 1).

Table 1. Summary of various types of population, corresponding estimands, and weights for both treated and control subjects.

Population	Estimand	Weight
Combined	ATE	$w_1 = \frac{1}{e_i(x)}, w_0 = \frac{1}{1-e_i(x)}$
Treated	ATT	$w_1 = 1, w_0 = \frac{e_i(x)}{1-e_i(x)}$
Overlap	ATO	$w_1 = 1 - e_i(x), w_0 = e_i(x)$
Matchable Treated	ATM	$w_1 = \frac{\min\{e_i(x), 1-e_i(x)\}}{e_i(x)}$, $w_0 = \frac{\min\{e_i(x), 1-e_i(x)\}}{1-e_i(x)}$

(2) ATT Weighting

Alternate weighting $w_i = A_i + \frac{(1-A_i)e_i}{1-e_i}$ is used to obtain ATT, that is $w_i = 1$ when a subject is in the treatment group and $w_i = \frac{e_i}{1-e_i}$ when a subject is in the control group (Table 1). A key requirement for both IPTW weighting and ATT weighting is the positivity assumption, meaning that PS should not be too close to 0 or 1. When this assumption fails, a small number of highly influential weights may lead to unstable weighting estimators. A few alternative methods have been proposed, including trimming and the overlap weighting [44].

(3) Overlap Weighting (OW)

The overlap weight is the probability that a subject is assigned to the opposite group, i.e., $1 - e_i$ for a subject in the treated group and e_i for a subject in the control group. The OW focuses on the causal effects on the population with the most overlap in covariates between two treatment groups [18]. Estimation under unconfounded or ignorable treatment assignment is often hampered by a lack of overlap in the covariates, which may result in imprecise estimates and lead to estimators that are sensitive to the choice of specification. The OW procedure involves the following a few steps: (1) estimating the e_i from a model;

(2) calculating the weights based on the estimated PS: $\hat{w}_1 = 1 - \hat{e}_i(x)$ if in the treatment group and $\hat{w}_0 = \hat{e}_i(x)$ if in the control group; (3) normalizing the weights so that the sum of the weights equals 1 within each group; (4) estimating the average treatment effect for the overlap population by the difference of the OW-weighted average outcomes between the groups.

Compared to the traditional IPTW weights and associated trimming methods, OW has several advantages: (1) there are no extreme weights; (2) it gives minimum variance of the weighted estimator of causal effects among all balancing weights (including IPTW); (3) the exact mean balance of covariates is achieved when PS is estimated via a logistic regression; and (4) there is no need to choose an artificial cutoff point, as required by trimming.

3.3.4. Use PS in Regression

Rosenbaum and Rubin [9] suggest to add the PS term in the regression model. For example, the estimated PS term ($\hat{e}(X_i)$) is added into a linear regression model:

$$Y_i = \beta_0 + \beta_1 A_i + \beta_2 \hat{e}(X_i) + \varepsilon_i,$$

where A_i indicates the treatment assignment for subject i and β_1 is the treatment effect conditional on the PS values that are calculated based on the covariates.

3.4. Doubly Robust Estimator

Doubly robust estimators combine models for the treatment and outcome in such a way that they provide unbiased estimates for the treatment effect as long as one of the models is correctly specified. Augmented Inverse Probability Weighting (AIPW) is a commonly used doubly robust estimator. The AIPW is built by combing an inverse probability weighting approach with g-computation [45]. For each treatment group, a separate model for the outcome variable is fitted by using a set of covariates, and the potential outcomes that correspond to each treatment assignment are predicted for all the subjects as follows:

$$\hat{Y}_i(0) = g^{-1}\left(X_i\hat{\boldsymbol{\beta}}_c\right),\ \hat{Y}_i(1) = g^{-1}\left(X_i\hat{\boldsymbol{\beta}}_t\right)$$

where g^{-1} is the inverse link function used in the generalized linear model for the outcome variable, $\hat{\boldsymbol{\beta}}_c$ is the regression coefficient estimate for the outcome regression model in the control group, and $\hat{\boldsymbol{\beta}}_t$ is the coefficient estimate for the outcome regression model in the treatment group. The AIPW estimates are given by [46]

$$\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-A_i)Y_i}{1-\hat{e}_i} + \frac{(A_i-\hat{e}_i)}{1-\hat{e}_i}\hat{Y}_i(0),$$

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i}{\hat{e}_i} - \frac{(A_i-\hat{e}_i)}{\hat{e}_i}\hat{Y}_i(1),$$

where $\hat{e}_i$ is the predicted PS for a subject from the treatment model.

A targeted maximum likelihood estimator (TMLE) is another doubly robust estimator that is based on a targeting step to optimize the bias-variance tradeoff for the parameters of interest [47]. Like the g-computation, TMLE estimates the predicted probabilities of potential outcomes for each subject. Then, g-computation calculates the difference in those two predicted probabilities over all levels of the covariates, whereas TMLE involves an extra targeting step incorporating the inverse probability weights prior to calculation of treatment difference.

4. Diagnostics of Covariate Balance

We outline the methods used for assessing balance in covariates suggested in Zhang et al. [48]. These diagnostics compare whether the distributions of relevant covariates are similar between the treated and untreated subjects [30,49].

The common measures of balance include the standardized differences and comparison of the distributions of covariates. The standardized mean difference (SMD) is defined as

$$\text{SMD}_{1\ vs\ 0} = \frac{|\ \overline{X_1} - \overline{X_0}\ |}{\sqrt[2]{(S_1^2 + S_0^2)/2}},$$

where $\overline{X_1}$, $\overline{X_0}$ are the sample means of the covariate for the two arms and S_1^2, S_0^2 are the corresponding sample variances. The SMD for the dichotomous variable is

$$\text{SMD}_{1\ vs\ 0} = \frac{|\ \overline{p_1} - \overline{p_0}\ |}{\sqrt[2]{(\ \overline{p_1}(1-\overline{p_1}) + \overline{p_2}(1-\overline{p_2}))/2}},$$

where $\overline{p_1}$, $\overline{p_0}$ are the sample proportions for the two treatment arms. This formula can be extended for categorical variables where the comparison can be employed at each level of the variable or the Mahalanobis distance can be used, as proposed by Dalton [50]. SMD is interpreted as the mean difference in a unit of the standard deviation and does not depend on sample sizes or units of the variable that is measured. Because SMD does not depend on the measurement unit, it allows for comparisons between variables with different units. SMD is often presented with a love plot that graphically displays a covariate balance before and after adjusting. Generally, 0.1 represent reasonable SMD thresholds for balance [48]. Other common balance measures include the Kolmogorov–Smirnov (KS) test statistics and t-statistics [51]. Variance is the second central moment of the mean and should also be compared. A variance ratio of 1 between treatment groups indicates a good balance, and a variance ratio between 0.5 and 2 is generally acceptable [48].

Moreover, we can look at the interactions, higher order terms, etc. The standard distribution tests can be employed, such as empirical cumulative distribution functions or non-parametric estimates of the density function of each covariate. Simple plotting, such as with side-by-side boxplots, Q-Q plots, and residual plots, is helpful and convenient.

Prognostic scores, defined as the predicted probability of the outcome in the control, can be used for assessing balance as well [52]. We can first regress the response on covariates only in the control group; then, that fitted model is employed to predict an outcome for all subjects. Simulation study has shown that the prognostic score approach can outperform mean differences and significance tests for assessing balance [52].

5. Software Tools for Implementation

Several software programs that implement confounding adjustment are available in many programming languages, including R, Python, SAS, and Stata. There are a number of R packages for PS methods, including twang [53], CBPS [54], PSW [55], ATE [56], WeightIt [57], causalweight [58], sbw [59], and PSWeight [60], as well as several packages that implement doubly robust estimators, including AIPW [61], CausalGAM [62], npcausal (nonparametric causal inference) [63], and tmle (targeted maximum likelihood estimation) [64]. The Comprehensive R Archive Network (CRAN) task view for causal inference provides a list of many more packages related to this topic [65]. Some packages are available in Python, such as zEpid [66], delicatessen [67], and PyWhy [68]. SAS provides many procedures for general confounding adjustment, such as PSMATCH, LOGISTIC, CAUSALGRAPH, CAUSALMED, and CAUSALTRT [46,69]. These causal procedures can be used to calculate PS, produce matched sets, estimate various estimands, and assess covariate balance. There are many ways of calculating standard errors in causal inference. The CAUSALTRT Procedure in the SAS/STAT® 15.3 User's Guide provides details and formulas for standard errors and confidence intervals [69]. Ding [13] gives good coverage of obtaining standard errors for various estimators.

6. Example

This example uses the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatment (SUPPORT), which was a multi-center observational study on

seriously ill and hospitalized patients to examine the effectiveness of Right Heart Catheterization (RHC) in the initial care of critically ill patients [70]. The dataset pertains to Day 1 of hospitalization. This example includes all 5735 subjects who were admitted to an ICU in the first 24 h after entering the study. RHC was coded as present if it was performed, and there were 2184 patients who received RHC. The outcome of interest is patients' chance of surviving by the end of first month. The original analysis by Connors et al. [70] used a binary logistic model to obtain PS to match RHC patients with No-RHC patients. Their results provided evidence that RHC patients had a decreased survival time.

Here, we apply the reviewed methods (i.e., regression with all covariates, g-computation, weighting, stratification, matching, and AIPW) and describe important steps for their implementation. In the Supplementary document, we provide code to replicate the analyses.

All the covariates (~50) are included, as suggested by Connors et al. [70]. The logistic regression is used to determine the PS, i.e., the probability of receiving RHC for each subject. The effectiveness of using PS to account for major covariates' imbalance is tested for differences between the two RHD groups (with and without RHC).

The PS distributions for the two RHD groups are displayed in Figure 1. The graph shows apparent differences in the PS distributions. There are many subjects with small PS values (around 0) and large ones (close to 1), indicating some covariate combinations were highly predictive of receiving RHC. The patients whose PS and RHC status do not agree may be atypical but received large weights.

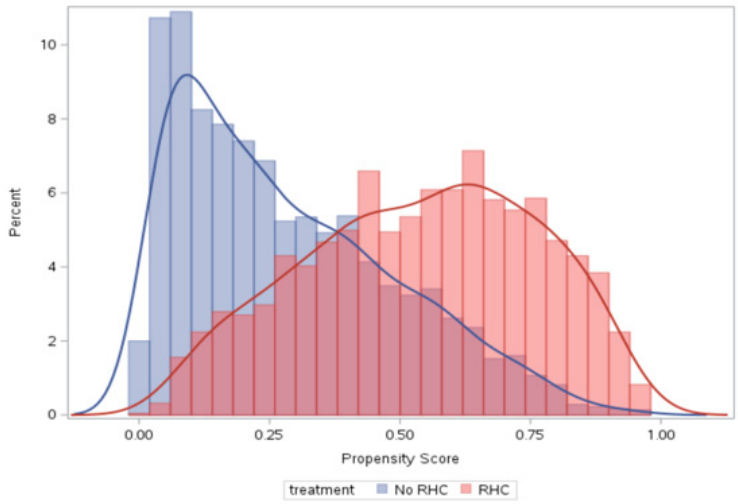

Figure 1. Histogram of the propensity score distributions for the two treatment groups. The blue and red curved lines are the probability density curves of the propensity score distributions for the two groups.

The love plot in Figure 2 compares the standardized differences of selected covariates included in the PS model: (1) blue for the original data in the regular regression; (2) purple for ATT weighting; (3) red for ATE weighting; (4) black for matching; (5) green for overlap weighting. We can see that the mean differences for most covariates are quite large in the traditional regression but are reduced with the adjustment methods. For all the methods, the largest SMD in absolute value is less than the recommended limit of 0.1 [48]. These indicate the effect of those methods on reducing the differences in covariates. The SMD with overlap weighting is the smallest (very close to zero in green). PS matching and weighting seem to remove a greater portion of systematic differences between the treated and untreated subjects compared with many other approaches, such as stratification and traditional regression, which is in agreement with Peter Austin's 2009 paper [71]. Among weighting approaches, overlap weighting seems to perform the best in terms of having the smallest mean difference, as expected.

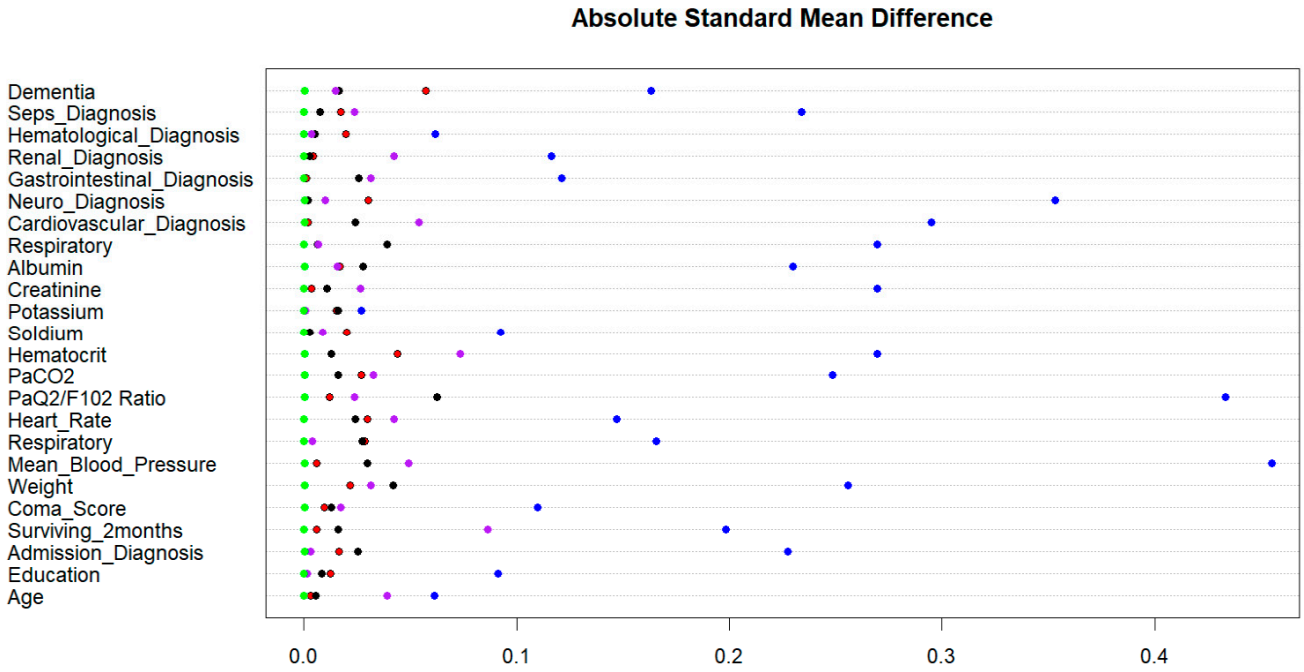

Figure 2. A love plot comparing the absolute standardized difference of selected covariates included in the model with different methods: (1) blue for original data in the traditional regression, (2) purple for ATT weighting, (3) red for ATE weighting with IPTW, (4) black for matching, and (5) green for overlap weighting.

The endpoint was surviving up to one month post-ICU admission. A binomial model with the logit link is used to obtain the risk difference and odds ratio for comparing the treatment effect of RHC versus No-RHC. We present the point estimates on the risk difference, standard errors, *p*-values, and corresponding confidence intervals for various methods. The results are displayed in Table 2, where the content is first arranged by the type of estimands, then by confounding adjust methods. In the Estimand column, we list the crude effect estimate, ATE, ATT, ATM, and ATO. For the crude effect, no covariates were included in the logistic model except for the treatment variable, and SAS's LOGISTIC procedure was used. For ATE, many methods were implemented, including the traditional regression with all covariates (SAS LOGISTIC), g-computation (SAS CAUSALTRT), PS stratification (SAS PSMATCH), PS weighting (SAS PSMATCH), and doubly robust estimators (SAS CAUSALTRT, the AIPW and PSWeight packages in R). For the remaining ATT, ATM, and ATO estimands, almost all estimators were employed as well. In Table 2, the results from using SAS are presented whenever there is a suitable procedure available in SAS, with the exception that the PSWeight package in R is used for the doubly robust estimator for ATT and overlapping weighting.

The crude estimate, without accounting for any covariates, and the CATE with a traditional regression, accounting for all covariates, give relatively larger risk differences (0.074 and 0.072, respectively) compared to most other methods (0.050–0.059). Most other methods, such as g-computation, PS weighting, etc., produce similar results in terms of having comparable point estimates and confidence intervals. The risk difference between the RHC and No-RHC control groups is all positive, indicating that patients with RHC had an increased rate of 30-day mortality after adjusting for treatment selection bias. This result confirms with the original analysis by Connors et al. that RHC patients had decreased survival rates compared to No-RHC patients [70].

Table 2. Analysis results for the binary outcome of passing the first month: The point estimates on risk difference in having 30-day mortality between RHC and No-RHC groups, associated standard errors, *p*-values, and confidence intervals for various methods. All methods except for the crude model include all available covariates.

Estimand	Methods	Risk Difference	Standard Error	*p*-Value	95% CI of Risk Difference
Crude	Regression without covariates	0.074	0.013	<0.0001	[0.048, 0.099]
CATE	Traditional regression with all covariates	0.072	0.021	<0.0001	[0.031, 0.114]
ATE	G-computation	0.059	0.014	<0.0001	[0.032, 0.087]
	PS stratification (5 strata)	0.055	0.032	0.1622	[−0.008, 0.119]
	PS weighting: IPTW	0.053	0.010	<0.0001	[0.035, 0.070]
	Doubly robust	0.059	0.014	<0.0001	[0.032, 0.087]
ATT	G-computation	0.056	0.014	<0.0001	[0.028, 0.085]
	PS stratification (5 strata)	0.055	0.030	0.1623	[−0.008, 0.118]
	PS weighting	0.053	0.014	<0.0001	[0.025, 0.081]
	Doubly robust	0.062	0.016	<0.0001	[0.029, 0.094]
ATM	PS matching	0.049	0.017	0.0035	[0.016, 0.08]
ATO	Overlap weighting	0.057	0.012	<0.0001	[0.031, 0.083]
	Overlap weighting with doubly robust	0.059	0.012	<0.0001	[0.033, 0.085]

Note: The results from using SAS are presented whenever there is a suitable procedure available to use in SAS. For the doubly robust estimator for ATT and overlapping weighting, the PSWeight package in R was used.

PS stratification can be unstable when the number of strata is large or the data size is small. In this example, stratification with five strata showed some signs of instability, with the first stratum and the last stratum having a relatively large standard error; additionally, it would be even worse in terms of having estimates that are very different from stratum to stratum when the number of strata increases to 10.

The ATM risk difference obtained from the matched observations is the smallest (0.049), proving to be a little different than most ATE and ATT estimates. This is because ATM represent the matched observation pool, a subset of the whole population that most ATE and ATT is derived from.

In general, it seems that the ATO with overlap weighting and its doubly robust version have the smallest standard errors and, therefore, the smaller confidence intervals, as claimed in Li et al. [17]. However, the performance of any method must be considered in terms of the estimand and that estimand's relevance to the motivating scientific question.

7. Discussions

In randomized control trials, it is reasonable to assume there are no systematic differences in covariates between treatment groups. Therefore, the causal effect of a treatment can be directly estimated by comparing the observed outcome in the active treatment group versus that in the control group. To evaluate the treatment effect from observational data, additional effort must be made to remove the impact of confounding variables, which are related to both the treatment and outcome.

For observational data, making causal inference needs to elect an appropriate estimand to start. The choice of estimands relies on the research objectives. For example, if we want

to make an inference about the effect on outcomes for all subjects, then the ATE should be the preferred estimand. If the population of interest is among subjects who selected treatment, then the ATT should be used. If one is interested in comparing causal effects estimated from an observational study to that from a randomized trial, then ATO might be more appropriate.

We reviewed a wide range of confounding adjustment methods. Each method has its own pros and cons. Table 3 summarizes the advantages and disadvantages of different methods.

Table 3. Comparison of different confounding adjustment methods.

Method	Advantage	Disadvantage
Traditional regression	• Simplest and easy to fit • Provides a prognostic model for outcome of interest	• Model mis-specification decreases precision in unbalanced design • May be implausible with a large number of covariates • Treatment effect is assumed to be the same for different levels of the covariates
G-computation	• Treatment effect can be different for different levels of the covariates • Robust to model mis-specification	• Needs more time to compute confidence intervals
PS Stratification	• Keeps all the data and divides into strata • Gives estimates for all strata • Explores interactions of treatment and PS	• Has bias when a small number of strata are used • Unstable when the number of strata is large and/or the data size is small
PS Matching	• Simple to perform and interpret • Often provides good balance among matched pairs in most cases	• Loses data due to unmatching • May not be precise • Need to choose matching algorithm
PS Weighting	• Keeps all the data • Often provides good balance in covariates • Easy to implement and intuitive to understand	• May be unstable with the presence of extreme weights • Needs to decide whether to take out very large and small weights
Traditional Regression with PS as a Covariate	• Simple to implement	• May not be necessary • Removes less difference compared with matching and weighting.
Doubly Robust Estimator	• Unbiased if one of the treatment model and response model is correct	• More difficult to implement • Takes a longer time
Overlap Weighting	• Weights are bounded between 0 and 1 • Minimize the asymptotic variance • Define a population of substantial clinical relevance and policy interest • Doubly robust estimator can be applied	• Not available in some software languages except for PSWeight package in R

Traditional regression models are among the easiest to implement due to many software packages being available. However, these models may be less efficient in reducing confounding bias and more difficult to interpret when there are many covariates. Specifically, the treatment effect is assumed to be the same for all levels of the covariates included in a model [36]. Moreover, the conditional effect and marginal effect may no longer be in the same direction due to non-collapsibility [38]. Model mis-specification may also impact precision in unbalanced designs with treatment-effect heterogeneity [37]. When g-computation

is applied, treatment effects can be differentiated by covariates. G-computation is effective in reducing confounding bias and balancing covariates. However, variance estimation is more complex, relying on either bootstrapping or the empirical sandwich variance estimator [72,73].

PS stratification has the advantage of keeping the whole data and exploring possible interactions between the treatment variable and PS. It tends to work well with small to moderate covariate imbalances [35]. If there are not many strata, residual confounding within strata may cause bias. To mitigate this bias, we can increase the number of strata. The more strata used, the closer the stratification will be to the matching method. However, stratification may perform poorly when the data size is small by giving inconsistent results across different strata, as shown in our example. To choose the proper number of strata, we should consider both the need for confounding control and the need of having enough events. Previous research has shown that five strata may reduce confounding bias by up to 90% [35].

PS-based matching is simple to use and often performs well in most cases. It provides nice covariate balance among the matched subjects. However, it occasionally tends to provide imprecise estimates as a result of the fact that some data without matches are dropped. Matching can result in large variance in estimates if a great deal of data is taken out. There are a number of matching techniques in the literature but little guidance to how to select among them in practice; the primary advice seems to select the one with the best balance [23]. Multivariate matching with the Mahalanobis distance or coarsened exact matching [74] are competitive, if not preferable, to PS matching in their ability to reduce imbalance and estimation bias.

PS weighting keeps all the data and often provides good balance among groups in most case [35]. When a covariate imbalance is large, the PS will be close to either 0 or 1, meaning that some subjects can have very large weights. It produces unbiased estimates but can have large variances if many subjects have large weights. Trimming can be employed in the case of many large weights. Yang and Ding (2018) [75] provided asymptotic theories for trimming, pointing out that trimming may introduce extra arbitrariness to the process while stabilizing PS weighting.

Treating PS as a covariate in a regression model is very convenient to achieve and is efficient in balancing covariates. A disadvantage is that it requires that the regression is correctly specified [71]. Researchers have also demonstrated that confounding adjustment using PS removes less of the systematic differences if compared with other approaches, such as the PS matching and weighting [71].

Doubly robust estimators, like AIPW and TMLE, offer clear advantages over g-computation and PS methods [76,77]. First, doubly robust estimators are unbiased as long as either the treatment model or outcome model is correctly specified. Second, doubly robust estimators are more efficient than IPW when the outcome model is not grossly mis-specified. Third, the variance estimator based on the influence curve has a closed-form. Lastly, some doubly robust methods allow for valid use of machine learning to estimate the PS and outcome models, unlike g-computation and PS methods [78,79].

Overlap weighting does not involve issues related to large weight problems, unlike standard inverse probability weights, as overlap weights are bounded between 0 and 1. Overlap weighting can obtain the exact mean balance of any covariates and minimize the asymptotic variance, as shown in the example. The variance estimators of the OW estimator of the marginal treatment effect have a closed-form, whereas the bootstrap or simulation used to estimate the variances with non-linear estimators do not [37].

As an alternative to PS methods, covariate balancing can be achieved by multivariate reweighting methods such as entropy balancing [80]. It can exactly match the first, second, and possibly higher moments of specified covariates. These balance improvements can potentially reduce model dependence for the subsequent estimation of treatment effects.

We have carried out a comprehensive review of common confounding adjusting approaches, including outcome regression models, g-computation, PS methods, and doubly

robust methods. Estimands and target population should be considered in determining which methods produce the most valid results. Each method has its own advantages and disadvantages. We conclude that there are considerable differences between estimands and corresponding estimators; however, none of them have proven to be uniformly better.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app14093662/s1, The SAS code which was used to replicate the analyses is displayed in Supplementary File.

Author Contributions: Conceptualization, A.X.S. and H.C.; methodology, A.X.S., P.N.Z. and H.C.; software, A.X.S. and P.N.Z.; validation, H.C.; formal analysis, A.X.S.; investigation, H.C.; resources, H.C.; data curation, A.X.S. and H.C.; writing—original draft preparation, A.X.S.; writing—review and editing, P.N.Z. and H.C.; visualization, A.X.S.; supervision, P.N.Z. and H.C.; project administration, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Fan Li at the Duke University for her insightful comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. A.X.S. and H.C. are employed by AstraZeneca and Pfizer, and may own stocks of their companies. However, all of the contents in this article are strictly educational, instructive, and methodological.

References

1. Friedman, L.M.; Furberg, C.D.; DeMets, D.L.; Reboussin, D.M.; Granger, C.B. *Fundamentals of Clinical Trials*; Springer: Berlin/Heidelberg, Germany, 2015.
2. Yang, H.; Yu, B. *Real-World Evidence in Drug Development and Evaluation*; CRC Press: Boca Raton, FL, USA, 2021.
3. ICH. *E10 Choice of Control Group and Related Issues in Clinical Trials*; ICH: Geneva, Switzerland, 2001.
4. Lu, H.; Cole, S.R.; Hall, H.I.; Schisterman, E.F.; Breger, T.L.; KEdwards, J.; Westreich, D. Generalizing the per-protocol treatment effect: The case of ACTG A5095. *Clin. Trials* **2019**, *16*, 52–62. [CrossRef] [PubMed]
5. Cole, S.R.; Edwards, J.K.; Zivich, P.N.; Shook-Sa, B.E.; Hudgens, M.G.; Stringer, J.S. Reducing Bias in Estimates of Per Protocol Treatment Effects: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Netw. Open* **2023**, *6*, e2325907. [CrossRef] [PubMed]
6. FDA. *Rare Diseases: Natural History Studies for Drug Development*; Guidance for Industry; FDA: Silver Spring, MD, USA, 2019.
7. FDA. *Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products*; FDA: Silver Spring, MD, USA, 2021.
8. Rubin, D.B. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Am. Stat. Assoc.* **1979**, *74*, 318–328.
9. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
10. Rosenbaum, P.R. Optimal matching for observational studies. *J. Am. Stat. Assoc.* **1989**, *84*, 1024–1032. [CrossRef]
11. Hernan, M.A.; Robins, J.M. *Causal Inference: What If*; CRC Press: Boca Raton, FL, USA, 2023.
12. Igelström, E.; Craig, P.; Lewsey, J.; Lynch, J.; Pearce, A.; Katikireddi, S.V. Causal inference and effect estimation using observational data. *J. Epidemiol. Community Health* **2022**, *76*, 960–966. [CrossRef]
13. Ding, P. A first course in causal inference. *arXiv* **2023**, arXiv:230518793.
14. Smith, M.J.; Mansournia, M.A.; Maringe, C.; Zivich, P.N.; Cole, S.R.; Leyrat, C.; Belot, A.; Rachet, B.; Luque-Fernandez, M.A. Introduction to computational causal inference using reproducible Stata, R, and Python code: A tutorial. *Stat. Med.* **2022**, *41*, 407–432. [CrossRef] [PubMed]
15. Ding, P.; Li, F. Causal inference: A missing data perspective. *Stat. Sci.* **2018**, *33*, 214–237. [CrossRef]
16. Crump, R.K.; Hotz, V.J.; Imbens, G.W.; Mitnik, O.A. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **2009**, *96*, 187–199. [CrossRef]
17. Li, F.; Morgan, K.L.; Zaslavsky, A.M. Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* **2018**, *113*, 390–400. [CrossRef]
18. Li, F.; Thomas, L.E.; Li, F. Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.* **2019**, *188*, 250–257. [CrossRef] [PubMed]
19. Hernán, M.A.; Robins, J.M. Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* **2006**, *60*, 578–586. [CrossRef] [PubMed]
20. Cole, S.R.; Frangakis, C.E. The consistency statement in causal inference: A definition or an assumption? *Epidemiology* **2009**, *20*, 3–5. [CrossRef] [PubMed]

21. Zivich, P.N.; Cole, S.R.; Westreich, D. Positivity: Identifiability and estimability. *arXiv* **2022**, arXiv:220705010.
22. Morris, T.P.; Walker, A.S.; Williamson, E.J.; White, I.R. Planning a method for covariate adjustment in individually randomised trials: A practical guide. *Trials* **2022**, *23*, 328. [CrossRef] [PubMed]
23. Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **1998**, *66*, 315–331. [CrossRef]
24. Heckman, J.J.; Ichimura, H.; Todd, P. Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* **1998**, *65*, 261–294. [CrossRef]
25. Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **1986**, *7*, 1393–1512. [CrossRef]
26. Greenland, S.; Robins, J.M. Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **1986**, *15*, 413–419. [CrossRef]
27. Abadie, A.; Imbens, G.W. Large sample properties of matching estimators for average treatment effects. *Econometrica* **2006**, *74*, 235–267. [CrossRef]
28. Stuart, E.A. Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **2010**, *25*, 1–21. [CrossRef] [PubMed]
29. Guo, S.; Fraser, M.W. *Propensity Score Analysis: Statistical Methods and Applications*; SAGE Publications: New York, NY, USA, 2014.
30. Austin, P.C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **2011**, *46*, 399–424. [CrossRef]
31. Robins, J.M.; Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* **1995**, *90*, 122–129. [CrossRef]
32. Robins, J.M.; Hernan, M.A.; Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology* **2000**, *11*, 550–560. [CrossRef] [PubMed]
33. Hirano, K.; Imbens, G.W. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* **2001**, *2*, 259–278. [CrossRef]
34. Hirano, K.; Imbens, G.W.; Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **2003**, *71*, 1161–1189. [CrossRef]
35. Elze, M.C.; Gregson, J.; Baber, U.; Williamson, E.; Sartori, S.; Mehran, R.; Nichols, M.; Stone, G.W.; Pocock, S.J. Comparison of propensity score methods and covariate adjustment: Evaluation in 4 cardiovascular studies. *J. Am. Coll. Cardiol.* **2017**, *69*, 345–357. [CrossRef] [PubMed]
36. Vittinghoff, E.; McCulloch, C.E. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am. J. Epidemiol.* **2007**, *165*, 710–718. [CrossRef]
37. Zeng, S.; Li, F.; Wang, R.; Li, F. Propensity score weighting for covariate adjustment in randomized clinical trials. *Stat. Med.* **2021**, *40*, 842–858. [CrossRef]
38. FDA. *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products*; FDA: Silver Spring, MD, USA, 2023.
39. Snowden, J.M.; Rose, S.; Mortimer, K.M. Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *Am. J. Epidemiol.* **2011**, *173*, 731–738. [CrossRef] [PubMed]
40. Freedman, D.A. Randomization does not justify logistic regression. *Stat. Sci.* **2008**, *23*, 237–249. [CrossRef]
41. Lee, B.K.; Lessler, J.; Stuart, E.A. Improving propensity score weighting using machine learning. *Stat. Med.* **2010**, *29*, 337–346. [CrossRef] [PubMed]
42. Senn, S.; Graf, E.; Caputo, A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Stat. Med.* **2007**, *26*, 5529–5544. [CrossRef] [PubMed]
43. Gu, X.S.; Rosenbaum, P.R. Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graph. Stat.* **1993**, *2*, 405–420. [CrossRef]
44. Zhou, Y.; Matsouaka, R.A.; Thomas, L. Propensity score weighting under limited overlap and model misspecification. *Stat. Methods Med. Res.* **2020**, *29*, 3721–3756. [CrossRef] [PubMed]
45. Bang, H.; Robins, J.M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **2005**, *61*, 962–973. [CrossRef] [PubMed]
46. Lamm, M.; Yung, Y.-F. Estimating Causal Effects from Observational Data with the CAUSALTRT Procedure. In Proceedings of the SAS Global Forum 2017 Conference, Orlando, FL, USA, 2–5 April 2017; SAS Institute Inc.: Cary, NC, USA, 2017. Available online: http://support.sas.com/resources/papers/proceedings17/SAS0374-2017.pdf (accessed on 21 April 2024).
47. Schuler, M.S.; Rose, S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am. J. Epidemiol.* **2017**, *185*, 65–73. [CrossRef]
48. Zhang, Z.; Kim, H.J.; Lonjon, G.; Zhu, Y. Balance diagnostics after propensity score matching. *Ann. Transl. Med.* **2019**, *7*, 16. [CrossRef]
49. Austin, P.C. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol. Drug Saf.* **2008**, *17*, 1218–1225. [CrossRef]
50. Yang, D.; Dalton, J.E. A unified approach to measuring the effect size between two groups using SAS. In *SAS Global Forum*; 2012; pp. 1–6. Available online: https://support.sas.com/resources/papers/proceedings12/335-2012.pdf (accessed on 29 February 2024).
51. Austin, P.C.; Stuart, E.A. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **2015**, *34*, 3661–3679. [CrossRef] [PubMed]

52. Stuart, E.A.; Lee, B.K.; Leacy, F.P. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J. Clin. Epidemiol.* **2013**, *66*, S84–S90.e1. [CrossRef] [PubMed]
53. Ridgeway, G.; McCaffrey, D.F.; Morral, A.R.; Cefalu, M.; Burgette, L.F.; Pane, J.D.; Griffin, B.A. *Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the R TWANG Package*; RAND Corporation: Santa Monica, CA, USA, 2022.
54. Fong, C.; Ratkovic, M.; Imai, K.; Hazlett, C. Package 'cbps'. 2022. Available online: https://cran.r-project.org/web/packages/CBPS/CBPS.pdf (accessed on 21 April 2024).
55. Mao, H.; Li, L.; Greene, T. Propensity score weighting analysis and treatment effect discovery. *Stat. Methods Med. Res.* **2019**, *28*, 2439–2454. [CrossRef]
56. Haris, A.; Chan, G. ATE: Inference for average treatment effects using covariate balancing. *R Package Version 0.2.0* **2015**.
57. Greifer, N.; Greifer, M.N. Package 'WeightIt'. CRAN. 2019. Available online: https://cran.r-project.org/web/packages/WeightIt/index.html (accessed on 21 April 2024).
58. Bodory, H.; Huber, M. The Causal Weight Package for Causal Inference in R. 2018. Available online: https://cran.r-project.org/web/packages/causalweight/causalweight.pdf (accessed on 21 April 2024).
59. Zubizarreta, J.R.; Li, Y.; Kim, K.; Allouah, A.; Greifer, N. Package 'sbw'. 2019. Available online: https://cran.r-project.org/web/packages/sbw/sbw.pdf (accessed on 21 April 2024).
60. Zhou, T.; Tong, G.; Li, F.; Thomas, L.E. PSweight: An R package for propensity score weighting analysis. *arXiv* **2020**, arXiv:201008893. [CrossRef]
61. Zhong, Y.; Kennedy, E.H.; Bodnar, L.M.; Naimi, A.I. AIPW: An r package for augmented inverse probability–weighted estimation of average causal effects. *Am. J. Epidemiol.* **2021**, *190*, 2690–2699. [CrossRef]
62. Glynn, A.; Quinn, K.; Quinn, M.K.; Quinn, K.M.; Estimator, P.W. Package 'CausalGAM'. 2017. Available online: https://cran.r-project.org/web/packages/CausalGAM/CausalGAM.pdf (accessed on 21 April 2024).
63. Kennedy, E.H. Nonparametric causal effects based on incremental propensity score interventions. *J. Am. Stat. Assoc.* **2019**, *114*, 645–656. [CrossRef]
64. Gruber, S.; Van Der Laan, M. tmle: An R package for targeted maximum likelihood estimation. *J. Stat. Softw.* **2012**, *51*, 1–35. [CrossRef]
65. Mayer, I.; Zhao, P.; Greifer, N.; Huntington-Klein, N.; Josse, J. Cran Task View: Causal Inference. 2022. Available online: https://misscausal.gitlabpages.inria.fr/misscausal.gitlab.io/files/ctv/CausalInference.html (accessed on 21 April 2024).
66. Zivich, P.N.; Davidson-Pilon, C.; Diong, J.; Reger, D. Pzivich/zEpid: v0.9.1 (v0.9.1). 2022. Available online: https://zenodo.org/records/7242696 (accessed on 21 April 2024).
67. Zivich, P.N.; Klose, M.; Cole, S.R.; Edwards, J.K.; Shook-Sa, B.E. Delicatessen: M-estimation in Python. *arXiv* **2022**, arXiv:220311300.
68. Sharma, A.; Kiciman, E. DoWhy: A Python Package for Causal Inference. 2019. Available online: https://github.com/py-why/dowhy (accessed on 21 April 2024).
69. SAS. *The CAUSALTRT Procedure in in SAS/STAT®15.3 User's Guide*; SAS Institute Inc.: Cary, NC, USA, 2023.
70. Connors, A.F.; Speroff, T.; Dawson, N.V.; Thomas, C.; Harrell, F.E.; Wagner, D.; Desbiens, N.; Goldman, L.; Wu, A.W.; Califf, R.M.; et al. The effectiveness of right heart catheterization in the initial care of critically III patients. *JAMA* **1996**, *276*, 889–897. [CrossRef]
71. Austin, P.C. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med. Decis. Mak.* **2009**, *29*, 661–677. [CrossRef] [PubMed]
72. Ren, J.; Cislo, P.; Cappelleri, J.C.; Hlavacek, P.; DiBonaventura, M. Comparing g-computation, propensity score-based weighting, and targeted maximum likelihood estimation for analyzing externally controlled trials with both measured and unmeasured confounders: A simulation study. *BMC Med. Res. Methodol.* **2023**, *23*, 18. [CrossRef] [PubMed]
73. Stefanski, L.A.; Boos, D.D. The calculus of M-estimation. *Am. Stat.* **2002**, *56*, 29–38. [CrossRef]
74. Iacus, S.M.; King, G.; Porro, G. Causal inference without balance checking: Coarsened exact matching. *Political Anal.* **2012**, *20*, 1–24. [CrossRef]
75. Yang, S.; Ding, P. Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. *arXiv* **2017**, arXiv:170400666. [CrossRef]
76. Funk, M.J.; Westreich, D.; Wiesen, C.; Stürmer, T.; Brookhart, M.A.; Davidian, M. Doubly robust estimation of causal effects. *Am. J. Epidemiol.* **2011**, *173*, 761–767. [CrossRef] [PubMed]
77. Robins, J.; Sued, M.; Lei-Gomez, Q.; Rotnitzky, A. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Stat. Sci.* **2007**, *22*, 544–559. [CrossRef]
78. Zivich, P.N.; Breskin, A. Machine learning for causal inference: On the use of cross-fit estimators. *Epidemiology* **2021**, *32*, 393. [CrossRef]
79. Naimi, A.I.; Mishler, A.E.; Kennedy, E.H. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *Am. J. Epidemiol.* **2023**, *192*, 1536–1544. [CrossRef]
80. Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Anal.* **2012**, *20*, 25–46. [CrossRef]