HHS Public Access

Author manuscript

Comput Stat Data Anal. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Comput Stat Data Anal. 2018 December ; 128: 48–57. doi:10.1016/j.csda.2018.05.003.

Measuring Model Misspecification: Application to Propensity Score Methods with Complex Survey Data

David Lenis[a,1], **Benjamin Ackerman**[a], and **Elizabeth A. Stuart**[a,b,c]

[a]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St, E3527, Baltimore, MD 21205, USA

[b]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, 8th Floor, Baltimore, MD 21205, USA

[c]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 624. N. Broadway, 4th Floor, Baltimore, MD 21205, USA

Abstract

Model misspecification is a potential problem for any parametric-model based analysis. However, the measurement and consequences of model misspecification have not been well formalized in the context of causal inference. A measure of model misspecification is proposed, and the consequences of model misspecification in non-experimental causal inference methods are investigated. The metric is then used to explore which estimators are more sensitive to misspecification of the outcome and/or treatment assignment model. Three frequently used estimators of the treatment effect are considered, all of which rely on the propensity score: (1) full matching, (2) 1:1 nearest neighbor matching, and (3) weighting. The performance of these estimators is evaluated under two different sampling designs: (1) simple random sampling (SRS) and (2) a two-stage stratified survey. As the degree of misspecification of either the propensity score or outcome model increases, so does the bias and the root mean square error, while the coverage decreases. Results are similar for the simple random sample and a complex survey design.

Keywords

Model Misspecification; Non-experimental study; Propensity Score Matching; Treatment on the Treated Weighting; Complex Survey Data; Causal Inference

Correspondence to: David Lenis.

[1]Present Address: Flatiron Health, 200 5th Ave, 8th Floor, New York, NY 10100, USA

1. Introduction

Model misspecification is a potential problem for nearly all methods that use parametric models, which leads to worries about incorrect inferences from misspecified models. However, to this point there has been relatively little formal investigation of model misspecification, including characterization of the extent of misspecification and how that might impact methods' performance. Here we propose a metric of misspecification and investigate the consequences of model misspecification within the context of causal inference in non-experimental studies, where there have been longstanding debates about whether misspecification of the treatment assignment model or the outcome model is more detrimental to estimation of treatment effects. Even though the measure for the degree of model misspecification presented in this article is used here in the context of causal inference, it can easily by applied to assess the impact of model misspecification in other model-based methods.

Randomized clinical trials (RCTs) are considered the gold standard for estimating causal effects. In an RCT, the researcher knows the treatment assignment mechanisms, allowing unbiased estimators of causal effects. Nevertheless, it is not unusual to find circumstances where a random assignment of the treatment is unfeasible or unethical. When this happens, researchers need to rely on non-experimental data.

A main drawback of non-experimental data is that the treatment assignment is not random, therefore there may be confounders that are related to the outcome and differ between treatment and comparison groups. Failure to address confounding will lead to biased estimators of causal effects. One way to mitigate confounding by observed characteristics is using the propensity score, which models the probability of being assigned to the treatment group given the set of confounders.

Non-experimental studies provide a particularly interesting case study for examining model misspecification because model misspecification can be an issue in two ways when using propensity score methods to estimate causal effects: first, in estimating the propensity score, and second, in the outcome model. Since the true treatment assignment mechanism is hardly ever known when working with non-experimental data, different approaches have been suggested to model and estimate the propensity score. While some authors have proposed nonparametric estimation procedures (Hahn, 1998; Imbens, 2004; Ho et al., 2011), it is common practice to estimate the propensity score parametrically via logistic regression.

Models are also used in the outcome analysis. Work by Cochran and Rubin (1973), Rubin (1973b), Carpenter (1977), Rubin (1979), Rosenbaum and Rubin (1984), Robins and Rotnitzky (1995), Heckman and Todd (2009), Rubin and Thomas (2000), Glazerman et al. (2003) Imai and Van Dyk (2004), Abadie and Imbens (2006) and Ho et al. (2007) suggested that adjusting for confounders in an outcome model may significantly improve inference on causal effects.

Thus, model assisted estimation of causal effects is a common practice in causal inference. However, there have been relatively few formal investigations of the consequences of model misspecification for different propensity score methods, and whether the degree of

misspecification of the treatment assignment model has greater ramifications on the bias or mean squared error of the estimate than that of the outcome model. Previous studies of model misspecification in the context of causal inference have grouped misspecified models in broad ad-hoc categories such as “incorrect model” or “wrong model” (Drake, 1993; Kang and Schafer, 2007; Robins et al., 2007). To our knowledge, this is the first attempt to systematically quantify the degree of model misspecification and evaluate its impact on two of the more commonly used estimation procedures (i.e., propensity score matching and weighting) under different survey designs.

Complex survey designs provide an extra layer of complexity when estimating causal effects. Non-experimental studies often use complex survey data, but there is relatively little guidance on how to incorporate the survey design in propensity score methods. Zanutto et al. (2005) and Zanutto (2006) discussed the use of propensity score subclassification with complex survey data, as illustrated in Hornik et al. (2001). Work by Austin et al. (2016) and Lenis et al. (2017) extended the use of propensity score matching to complex survey data. Similarly, Ridgeway et al. (2015) provided some insight on how to compute inverse probability of treatment weighting (IPTW) estimators using complex survey data; however, it is unclear whether model misspecification would have different implications in the complex survey context.

This paper is organized as follows: in Section 2, we present key definitions and assumptions needed for the estimation of causal effects in the context of non-experimental data. Section 3 reviews the methods implemented in our simulation study. Section 4 contains the details of our simulation study. In Section 5, the main results are presented, followed by the discussion and main conclusions in Section 6.

2. Definitions and Assumptions

2.1. The Causal Inference Framework

Traditionally, causal treatment effects are defined using the Rubin Causal Model (RCM) (Rubin, 1974). In the RCM, an individual treatment effect, associated with a binary treatment assignment T, is defined in terms of potential outcomes. For each unit i, $Y_i(t)$ with $t = 0, 1$, represents the outcome that would have been observed if unit i received the treatment t. Thus, the treatment effect for the i^{th} unit is equal to $Y_i(1) - Y_i(0)$. Notice that for any unit i, the pair $(Y_i(0), Y_i(1))$ is not observable - only one of the two potential outcomes is observed. Explicitly, the observed outcome, Y_i, is defined as:

$$Y_i = Y_i(1) \times T_i + Y_i(0) \times (1 - T_i). \quad (1)$$

Equation (1) is referred as the “consistency of the observed outcome assumption” (Hernan and Robins, 2017). Given that the unit level treatment effects cannot be estimated directly, we are often interested in estimating average treatment effects. At the population level, the most commonly defined average effects are: (1) the population average treatment effect (PATE) and (2) the population average treatment effect on the treated (PATT).

The PATE is defined as average effect across the population:

$$PATE = E\left[Y(1) - Y(0)\right]. \quad (2)$$

Under randomization of the treatment, units in the treated group and the units in the control group have similar distributions of covariates (observed and unobserved) and potential outcomes. In this way, the average outcome computed among the units in the treated group serves as a good counterfactual for the average outcome computed among the units in the control group. The differences between these two averages is an estimator of the population average treatment effect (PATE).

The PATT is defined as the average causal effect, computed only among those units in the population who were actually treated:

$$PATT = E\left[Y(1) - Y(0) \mid T = 1\right]. \quad (3)$$

When the treatment is randomized, it holds that the PATE is equal to the PATT. In non-experimental studies, where the treatment and comparison groups may be quite different from one another on confounders and effects, the PATT and the PATE can be different.

When randomization is not feasible, additional assumptions are required to identify and estimate treatment effects. In particular, a crucial assumption in the estimation of treatment effects is referred to as "ignorability" (Rosenbaum and Rubin, 1983). To further describe the implications of this assumption, we define for all i, $\mathbf{X}_i$ as q–dimensional vector of covariates, i.e., $\mathbf{X}_i = (X_{1,i}, \ldots, X_{q,i})$. Ignorability assumes that $\mathbf{X}$ contains all possible confounders: all variables related to treatment assignment and outcome. In other words, given the set of observed covariates $\mathbf{X}$, the treatment assignment is independent of the potential outcomes. The ignorability assumption means that the treatment assignment is random, conditionally on observed characteristics of the units in the sample. This implies that:

$$(Y_i(0); Y_i(1)) \perp\!\!\!\perp T_i \mid \mathbf{X}_i. \quad (4)$$

Another key assumption of the RCM is the Stable Unit Treatment Value Assumption (**SUTVA**). The implication of this assumption is twofold: (1) the treatment assignment of any unit does not affect the potential outcomes of other units (often referred to as non-interference) and (2) there is only one version of the treatment, implying that the treatment is comparable across units (Rubin, 1980).

2.2. PATT versus SATT

While many researchers are interested in estimating causal effects at a population level, data from a study sample can only be used to truly and consistently estimate a sample ATE (SATE). Estimation of the PATE requires one to have access to data on the full target population of interest, which is rare in practice. The SATE represents the difference in

average outcomes if everyone in the survey sample received the treatment versus everyone in the survey sample receiving the control condition. An unbiased estimator of the SATE (SATT) will correctly estimate the PATE (PATT) only when the sample distribution of the relevant variables is similar to its target population counterpart. One sampling design that guarantees this is a simple random sample (SRS), but this kind of sampling technique is hardly ever used. In general, most surveys are the result of complex sampling designs. Therefore, unless survey weights are used to weight the sample back to the population, using sample information to estimate a treatment effect will result in a consistent estimator for the SATE (SATT) but not for the PATE (PATT).

3. Propensity Score Methods

In this section we present three commonly used techniques to estimate population causal effects: (1) propensity score full matching, (2) 1:1 nearest neighbor propensity score matching, and (3) treatment on the treated weighting. While we focus on estimating the PATT in this paper, variations on these methods can also be used to estimate the PATE (Abadie and Imbens, 2006; Ridgeway et al., 2015).

3.1. Propensity Score Full and 1:1 Nearest Neighbor Matching

Matching estimators have been widely used in the context of non-experimental studies. They help reduce bias in the estimation of causal effects (Rubin, 1973a), are intuitive, and relatively easy to implement. Fundamentally, matching matches individual observations (i.e., comparison to treated units) on a set of observed covariates. The main goal of this matching approach is to generate a new sample (i.e., the matched sample), such that for every treated unit there is (at least) one comparison unit with similar values of observed covariates. The outcome of interest is then compared between the matched treated and matched comparison subjects to estimate the causal effect. One main disadvantage of this procedure resides in the fact that as the number of variables on which units are matched increases, the chances of finding matched pairs with similar observed characteristics decreases exponentially. Thus, matching directly on a set of covariates is only feasible in large samples and/or if a small set of covariates are used in the matching procedure. This is why propensity score matching can be useful. Rosenbaum and Rubin (1983) showed that a similar (or balanced) distribution of the observed characteristics can also be achieved when the matching procedure is based on the propensity score instead of the entire set of observed covariates. Guidelines regarding the implementation of propensity score matching in the context of complex survey data can be found in Austin et al. (2016) and Lenis et al. (2017)

We consider two types of matching methods: full matching and 1:1 nearest neighbor matching. Full matching (Stuart and Green, 2008) essentially forms a number of small subclasses based on the propensity score, where each subclass has at least one treated and at least one comparison subject, but may have multiple treated or multiple comparison. Weights according to the number of treated per subclass are used to estimate the effects to ensure that the comparison group better resembles the treated group. It can be thought of as an approach that is in between using 10 propensity score subclasses and weighting, where each subject is essentially its own subclass, described further below. Nearest neighbor 1:1

matching selects one comparison subject for each treated subject, based on propensity scores. Analysis proceeds by comparing outcomes between the matched treated and matched comparison subjects. Full details can be found in Stuart (2010).

3.2. Treatment on the Treated Weighting

An alternative approach to estimate causal effects is to compute a weighted estimator that weights subjects using a function of the propensity score. In the context of simple random samples (SRS), an IPTW estimator of the ATT requires, as a first step, the computation of propensity score based weights. The units in the comparison group receive a weight equal to the odds of receiving treatment, while the treated receive a weight equal to one (a different weighting strategy needs to be implemented when the goal is to estimate the ATE (Austin, 2011)). This serves to weight the comparison group to look similar to the treatment group, thus estimating the ATT (Robins et al., 2000; Harder et al., 2010).

After the propensity score weights are computed, a weighted difference in means (exposed versus unexposed) can be computed in order to estimate the ATT. Furthermore, weighted regression models can be fit to estimate causal effects (Joffe et al., 2004). This approach allows for the estimation of causal effects adjusting for relevant confounders. Ridgeway et al. (2015) developed a strategy to compute a propensity score weighted estimator using complex survey data.

3.3. Degree of Misspecification (DoM)

Relatively little previous work has explicitly addressed the level of misspecification in the propensity score and outcome model when assessing the impact of misspecification in the estimation of causal effects. Robins et al. (2007) evaluated the impact of model misspecification when implementing doubly robust estimators and concluded that “the relative performance of the estimators will very much depend on the data generating process and the nature of the model misspecifications” (p. 555). In a different but related context, Stuart and Jo (2015) examined violation of the assumptions underlying propensity score methods and instrumental variables methods and attempted to equate the extent of the violation of the key assumption of each approach.

In this article we propose a measure of the DoM of a model and explore how the DoM impacts the performance of the estimators considered. Controlling the DoM will allow us to: (1) evaluate how robust the considered estimators are to different levels of DoM and (2) assess whether the same level of DoM in each model (i.e., propensity score and outcome) has the same impact on the performance of the estimators considered.

Throughout this paper, we will use η to represent the DoM for a given model. For a given dependent variable, Z, we define μ_i as the mean of Z conditional on a set of predictors (i.e., $E[Z_i|\mathbf{X}_i]$) (here $\mathbf{X}_i$ represents the set of predictors. This set can also contain the treatment indicator. Notice that this is implies a slight abuse in notation since in Section 2 we defined $\mathbf{X}_i$ as a set of confounders that did not include the treatment indicator). We assume that there is a function g^C such that $\mu_i = g^C(\mathbf{X}_i)$. Thus, η can be defined as:

$$\eta = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,\hat{g}(\mathbf{X}_i) - \widehat{g^C}(\mathbf{X}_i)\,|}{\sigma_{\widehat{g^C}}} \quad (5)$$

Here N represents the population size, $\widehat{g^C}(\mathbf{X}_i)$ is the predicted conditional mean under the correct model specification for unit i with $i = 1, \ldots, n$, $\hat{g}(\mathbf{X}_i)$ is the predicted conditional mean under a given model specification for unit i with $i = 1, \ldots, n$. The symbol $\sigma_{\widehat{g^C}}$ represents the standard deviation of the predicted conditional means under the correct model specification. Thus when $g = g^C$, we have that $\eta = 0$ and when $g \neq g^C$ $\eta > 0$. Therefore we have that $\eta \in [0, \infty)$, and as η increases, so does the degree of misspecification of a given model.

This measure of DoM has some desirable properties: (1) is **unit independent**, which facilitates the comparisons across different working models and types of dependent variables (e.g., continuous, binary, categorical, etc.), (2) the magnitude is **informative** (i.e., higher values of η are associated higher degree of misspecification), (3) since η is computed in the population, it is not affected by sample size or the survey design.

Notice that η is defined as a parameter in our simulation study and since its computation requires knowledge of the true parametric model, it cannot be used in a real data analysis. Since our simulation study involves the estimation of the propensity score and outcome model, we have a DoM associated with the estimation of the propensity model (η_T) and a DoM related to the outcome model (η_Y).

The DoM of the propensity score model is defined as:

$$\eta_T = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,\hat{\pi}_i - \hat{\pi}_i^{\,C}\,|}{\sigma_{\hat{\pi}^C}}. \quad (6)$$

Here $\hat{\pi}_i$ is the predicted probability of being assigned to the treatment group under a given model specification, $\hat{\pi}_i^C$ is the predicted probability of being assigned to the treatment group under the correct model specification and $\sigma_{\hat{\pi}^C}$ is the standard deviation of the predicted probabilities of being assigned to the treatment group under the correct model specification. The DoM associated with the outcome of interest is defined as:

$$\eta_Y = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,\hat{Y}_i - \hat{Y}_i^{\,C}\,|}{\sigma_{\hat{Y}^C}}. \quad (7)$$

Here $\hat{Y}_i$ is the predicted observed outcome under a given model specification, $\hat{Y}_i^C$ is the predicted observed outcome under the correct model specification and $\sigma_{\hat{Y}^C}$ is the standard deviation of the predicted observed outcomes under the correct model specification.

3.4. Methods examined

In our simulation study (see Section 4) we compute three propensity score based methods to estimate the PATT, all using the propensity score: (1) full matching, (2) 1:1 matching, and (2) weighting. For all approaches we incorporate the survey weights in the estimation of the propensity score model, since Lenis et al. (2017) show that doing so could lead to more efficient estimators of the PATT.

Full matching is implemented as described in Stuart and Green (2008). The 1:1 nearest neighbor matching is done without replacement and using a greedy algorithm; results were nearly identical when an optimal algorithm was used instead. When the sample is obtained using a complex survey design, we follow Lenis et al. (2017) to implement the matching. Since we are assuming a non-response rate of 0% we do not implement the weight transfer described in Lenis et al. (2017). Weighting is implemented to estimate the ATT as described above. When the sample is the result of a complex survey design, we follow Ridgeway et al. (2015). That is, the survey weights are incorporated in the estimation of the propensity score model, and the final weights used in the outcome analysis are constructed by multiplying each survey weight by the propensity score based weights.

When survey models are used, we use the R package “survey” (Lumley, 2004, 2016) to account for the survey design and weights in the estimation procedure.

4. Simulation Study

4.1. The Data Generating Mechanism (DGM)

Our simulation study closely follows the one presented by Austin et al. (2016), with some modifications: (1) the PATT and the SATT differ from one another and (2) the degree of misspecification of the working models for the propensity score and outcome can be controlled.

As in Austin et al. (2016), we consider the case of a population of size $N = 1,000,000$, divided into 10 strata. Each strata contains 20 clusters, each composed of 5, 000 units.

We consider **two confounders** C_l with $l = 1,2$ and a data generating mechanism for the baseline covariates such that: (1) the probability density function is Normal, (2) the covariates are independent (i.e., correlation between the covariates is set equal to 0), (3) the standard deviation for each covariate is equal to 1 and, (4) the means vary across strata and cluster. Explicitly, for each strata j, the mean of the covariates deviates in μ_{lj} from 0, where μ_{lj} are obtained assuming that $\mu_{lj} \sim N(0, \tau^{stratum})$. Within each strata, the mean of each cluster (k) deviates from the strata specific mean by μ_{lk}, with $\mu_{lk} \sim N(0, \tau^{cluster})$. Thus the distribution of the l^{th} variable, in the j^{th} stratum, among the units of the k^{th} cluster is $C_{ljk} \sim N(\mu_{lj} + \mu_{lk}, 1)$. We set $\tau^{stratum} = 0.35$ and $\tau^{cluster} = 0.25$. The values for $\tau^{stratum}$ and $\tau^{cluster}$ are extracted from Austin et al. (2016)

The **treatment assignment** (T_i) model is defined as a Bernoulli random variable $T_i \sim Be(p_i)$ with

$$\text{logit}(p_i) = \alpha_0 + \sum_{l=1}^{2} \alpha_l C_{li} + \delta_d \alpha_3 C_{1i} C_{2i},$$

where $\alpha_0 = \log(0.20)$, $\alpha_1 = \log(2.00)$, $\alpha_2 = \log(2.50)$, and $\alpha_3 = \log(3.00)$. In this model, the multiplier δ_d with $d = 1, .., 11$ allows us to control the degree of misspecification of the working model (see Section 4.3) used to estimate the propensity score. The values of δ_d are selected such that the degree of misspecification (DoM) varies from 0.00 to 0.50.

The **potential outcomes** model under **control** is defined as $Y_i(0) \sim N(\mu_i^0, \sigma^2)$, with

$$\mu_i^0 = \beta_0 + \sum_{l=1}^{2} \beta_l C_{1i} + \Delta_m(\delta_d)\beta_3 C_{1i} C_{2i} + \sum_{j=2}^{10} \theta_j STR_{ji},$$

where $\beta_0 = \log(0.20)$, $\beta_1 = \log(2.50)$, $\beta_1 = -\log(2.00)$, $\beta_3 = \log(4.50)$, $\theta_j = \log(0.50)$ for $j = 2, .., 5$, and $\theta_j = \log(2.00)$ for $j = 6, \ldots, 10$. The term $\sum_{j=2}^{10} \theta_j STR_{j,i}$ ensures that the PATT and the SATT will be different. The variable $STR_{j,i}$ is a categorical variable that takes the value 1 if the sample unit i belongs to the j^{th} stratum. The parameter $\Delta(\delta_d)_m$ with $m = 1, \ldots, 11$ is indexed by δ_d to ensure that for every degree of misspecification in the propensity score model, the degree of misspecification of the outcome model also ranges from 0.00 to 0.50 by 0.05 increments. The potential outcome under **treatment** is defined by $Y_i(1) \sim N(\mu_i^1, \sigma^2)$, with $\mu_i^1 = \mu_i^0 + \gamma$, with $\gamma = \log(3.00)$. Recall that the observed outcome (Y_i) is defined as:

$$Y_i = T_i \times Y_i(1) + (1 - T_i) \times Y_i(0).$$

We model the outcome of interest as a continuous variable for two reasons. First, since the treatment effect is homogeneous, the PATT is equal to γ. Second, having a continuous outcome will allow us fit a model for the outcome of interest in the matched sample that will yield a consistent estimator of the PATT. This is due to the fact, as stated in Austin (2013), "that propensity score methods result in marginal estimates of effect, rather than conditional estimates of effect. When outcomes are continuous, a linear treatment effect is collapsible: the conditional and marginal estimates coincide. When the outcome is binary, regression adjustment in the propensity score matched sample will typically result in an estimate of the odds ratio. The odds ratio (like the hazard ratio) is not collapsible; thus the marginal and conditional estimates will not coincide."

4.2. Survey Designs

In our simulation study we consider two sampling schemes. First, we consider a simple random sampling scheme where 5, 000 units were randomly selected from the population without replacement. Second, we also implement a two stage stratified sample. As mentioned in Section 4.1, the target population consists of 10 strata, each with 20 clusters. Within each stratum, 5 clusters are selected randomly without replacement. Within each selected cluster, we draw a random sample without replacement of the final sampling units. Within each stratum, the same number of observations are selected among the sampled clusters. We allocate sample sizes to the 10 strata as follows: 750, 700, 650, 600, 550, 450, 400, 350, 300, and 250. Therefore, the final sample consists of 5, 000 units, which represents 0.5% of the target population. Survey weights are constructed to be equal to the inverse of the selection probability. Strata divide the population in mutually exclusive and exhaustive groups, and clusters within each stratum are randomly selected. Thus every strata is represented in the final sample, but not every cluster. For example, strata could be defined by states, while counties or street blocks define the clusters. In this example, every state will be represented in the final sample but not every county.

For simplicity, we assume a 0% non-response rate (work by Lenis et al. (2017) explored the consequences of the non-response in the estimation of population causal effects in the context of complex survey data).

We implement 1, 000 iterations in our simulation study. That is, under both sampling schemes 1, 000 samples are drawn from the population.

4.3. Analysis models

After the sample is obtained, the following propensity score model is estimated:

$$\text{logit}\,(p_i) = a_0 + \sum_{j=1}^{2} a_j C_{ij}\,. \quad (8)$$

Notice that by setting $\delta_1 = 0$ (see Section 4.1) the working model defined by equation 8 will be correctly specified, thus making the degree of misspecification equal to 0 ($\eta_T = 0$). The analysis outcome model is defined by the following equation:

$$m_i = b_0 + \sum_{j=1}^{2} b_j C_{j1} + \sum_{j=2}^{10} b_{j+2} STR_{ji} + b_{13} T_i\,. \quad (9)$$

Here m_i represents the model for the mean of the observed outcome, given the confounders, the strata identifier and the treatment assignment (i.e., $E[Y_i|C_{1i}, C_{2i}, STR_{2i}, \ldots, STR_{10i}, T_i]$). Again, by setting $\Delta_1(\delta_d) = 0$ for all δ_d (see Section 4.1), the working model defined by equation 9 will be correctly specified, making the associated degree of misspecification equal to 0 ($\eta_Y = 0$). Notice that the working models defined by equation 8 and equation 9 include all confounders, thus the assumption of no unmeasured confounders holds.

Therefore, the source of the misspecification in both models is the omission of the interaction term (i.e., C_1C_2).

5. Results

In this section we evaluate the performance of the estimator of γ as a function of the degree of misspecification using the following three metrics: (1) percentage bias (in absolute value), (2) empirical coverage of the 95% confidence interval and (3) root mean squared error.

Our main results are summarized in figures 1, 2 and 3. In these figures, the vertical axis displays the DoM in the outcome model (η_Y) while the horizontal axis shows the DoM in the propensity score model (η_T). The top three panels show the results associated with a SRS while the bottom three panels display the results associated with a two-stage stratified sample. The columns show results for full matching, 1:1 matching, and weighting, in that order (plots with value labels are available in the appendix).

Figure 1 shows how the absolute value of bias (in percentage) is affected by the DoM in both models. Lighter shades indicate less bias, while darker shades indicate higher levels of bias. From Figure 1 we can observe that results are similar for the simple random sample (top three panels) and a complex survey design (bottom three panels), and are remarkably similar across propensity score methods as well. As expected, the bias of the estimator increases as the DoM increases in both models. In fact, when the DoM is 0.50 in both models, the bias (in absolute value) can be as high as 200%. When the outcome model is correctly specified ($\eta_Y = 0$), both methods yield unbiased estimators of the PATT, regardless of the DoM associated with the propensity score model ($\eta_T \geq 0$). When the propensity score is correctly specified ($\eta_T = 0$) we observe that full matching and the weighting method (left and right panels) return an estimator that is unbiased regardless of the level of DoM associated with the outcome model ($\eta_Y \geq 0$). In the case of weighting this is due to the fact that the procedure used in the computation of the weighting estimator yields the doubly robust estimator attributed to Joffe (Robins et al., 2007). Doubly robust estimators (Scharfstein et al., 1999; Kang and Schafer, 2007) yield consistent estimators of the PATT when either the propensity score or the outcome model (but not necessarily both) are correctly specified (i.e., $\eta_T = 0$ or $\eta_Y = 0$). In the case of full matching, since the optimal stratification is determined using network optimization (see Rosenbaum (1991)), this procedure can create matched samples where the difference in the probability of receiving treatment (between treated and controls) are smaller compared to other matching algorithms (see Hansen (2004)). Thus, this method can be more successful in reducing model dependency than other matching algorithms. This feature translates into less bias even when the DoM in the outcome model is high. This same result does not hold for the 1:1 nearest neighbor matching estimator (top center panel), which has small bias when the DoM in the outcome model is high (even if the propensity score model is correctly specified). This can be seen when comparing the first column (in light gray) of the 1:1 matching figure, to the white columns for full matching and weighting. Across all three methods misspecifying the propensity score model, results in smaller biases than misspecifying the outcome model by the same degree. This result is consistent with the one obtained by Drake (1993).

Figure 2 displays the results associated with the empirical coverage of the 95% confidence interval. Lighter shades indicate higher coverage, while darker shades depict lower empirical coverage. Observe that there is a sharp fall in the coverage when the DoM exceeds 0.15 in both models. This is due to the fact that values of DoM higher than 0.15 are associated with bias larger than 10% (see Figure 1). Therefore, this pattern is expected, since the confidence intervals are centered at a value far from the true value of γ. Interestingly, full matching retains the highest coverage rates when models are misspecified, as compared to 1:1 matching or weighting.

Figure 3 summarizes the results for RMSE. Lighter shades indicate lower values of the RMSE, while darker shades show higher levels of RMSE. Notice that Figure 1 and Figure 3 display a similar pattern, indicating that there are no significant differences in the efficiency of the estimation procedures.

6. Discussion

In this paper, we explore how model misspecification affects the performance of three of the most commonly used methods to estimate the PATT: (1) propensity score full matching, (2) 1:1 nearest neighbor propensity score matching, and (2) treatment on the treated weighting. As noted in Section 4.3, an outcome model that adjusts for the confounders was used to estimate the PATT (i.e., γ).

One contribution of this paper is the careful quantitification of model misspecification. In Section 3.3 (see equation 6 and equation 7) we presented η, a metric of the degree of misspecification for a given model. Given that η is unitless, it can be used to compare the DoM of different models and different types of dependent variables. The fact that η is not affected by the sample size and survey design allowed us to evaluate the performance of the estimators in the context of complex survey data and simple random sampling. Futhermore, η can be easily used to evaluate the impact of model misspecification in other parametric models.

To our knowledge, this is the first attempt to systematically quantify the degree of model misspecification in the analysis models in order to evaluate its impact in the estimation of causal effects. Still, there are some limitations to our approach. The metric used to quantify the degree of misspecification (i.e., η) is computed at the population level and requires knowing the true model. Future work will focus on providing measures of the DoM that can be computed at a sample level. Additionally, we only explored the consequences of omitting the interaction term. In our simulation study, link functions are correctly specified and all relevant confounders are observed and measured without error. Future work will focus on assessing the impact of other types of model misspecification. This could include examining the implications of different confounding structures, such as when variables may or may not be true confounders and part of the misspecification may involve selecting incorrect variables to include for bias reduction.

Based on the metric of model misspecification, we evaluated the performance of methods for estimating the PATT in the presence of propensity score and/or outcome model

misspecification. Perhaps not surprisingly, but importantly, we found similar results across simple random samples and complex survey sample designs. This is useful guidance for researchers and implies that findings may be similar for a variety of study designs

All estimation procedures yield similar performance in terms of bias, coverage and RMSE when the outcome model is correctly specified ($\eta_Y = 0$), but the propensity score model is not ($\eta_T \geq 0$). When the propensity score model is correctly specified ($\eta_T = 0$), the weighting and full matching estimators are robust to different degrees of misspecification associated with the outcome model. Nevertheless, it is important to keep in mind that true models are rarely known. Thus in practice, it is very likely that both models (i.e., the propensity score and the outcome) will be misspecified (i.e., $\eta_T > 0$ and $\eta_Y > 0$). When this is the case, the performance of the estimation procedures are practically identical, which confirms results obtained by Kang and Schafer (2007).

In conclusion, as the degree of model specification increases, the performance of the estimators considered worsens. Under the more realistic scenario that both models (i.e., propensity score and outcome) present some degree of misspecification, the performance of the three estimators considered are practically identical. Thus, there is no methodological substitute for well informed and carefully planned model specification. While this is not a surprising result, having a metric of misspecification that can be used across models (e.g., treatment and outcome) can allow researchers to assess which models it is most important to specify correctly when using approaches that rely on multiple models together.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006; 74(1):235–267.

Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011; 46(3):399–424. [PubMed: 21818162]

Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Statistics in medicine. 2013; 32(16):2837–2849. [PubMed: 23239115]

Austin PC, Jembere N, Chiu M. Propensity score matching and complex surveys. Statistical Methods in Medical Research. 2016 0962280216658920.

Carpenter R. Matching when covariables are normally distributed. Biometrika. 1977:299–307.

Cochran WG, Rubin DB. Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A. 1973:417–446.

Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics. 1993:1231–1236.

Glazerman S, Levy DM, Myers D. Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science. 2003; 589(1):63–93.

Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998:315–331.

Hansen BB. Full matching in an observational study of coaching for the sat. Journal of the American Statistical Association. 2004; 99(467):609–618.

Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological methods. 2010; 15(3):234. [PubMed: 20822250]

Heckman JJ, Todd PE. A note on adapting propensity score matching and selection models to choice based samples. The econometrics journal. 2009; 12(s1):S230–S234. [PubMed: 20694053]

Hernan, MA., Robins, JM. Causal Inference. Boca Raton: Chapman & Hall/CRC. Forthcoming; 2017.

Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis. 2007; 15(3):199–236.

Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software. 2011; 42(8):1–28.

Hornik, R., Maklan, D., Judkins, D., Cadell, D., Yanovitzky, I., Zador, P., Southwell, B., Mak, K., Das, B., Prado, A., et al. Evaluation of the national youth anti-drug media campaign: Second semi-annual report of findings-april 2001. Rockville, MD: Westat; 2001.

Imai K, Van Dyk DA. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association. 2004; 99(467):854–866.

Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and statistics. 2004; 86(1):4–29.

Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. The American Statistician. 2004; 58(4):272–279.

Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science. 2007:523–539.

Lenis D, Nguyen TQ, Dong N, Stuart EA. Its all about balance: propensity score matching in the context of complex survey data. Biostatistics. 2017

Lumley T. Analysis of complex survey samples. Journal of Statistical Software. 2004; 9(1):1–19. R package verson 2.2.

Lumley, T. R package version 3.32. 2016. survey: analysis of complex survey samples.

Ridgeway G, Kovalchik SA, Griffin BA, Kabeto MU. Propensity score analysis with survey weighted data. Journal of Causal Inference. 2015; 3(2):237–249. [PubMed: 29430383]

Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. Statistical Science. 2007; 22(4):544–559.

Robins, JM., Hernan, MA., Brumback, B. Marginal structural models and causal inference in epidemiology. 2000.

Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association. 1995; 90(429):122–129.

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

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55.

Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American statistical Association. 1984; 79(387):516–524.

Rubin DB. Matching to remove bias in observational studies. Biometrics. 1973a:159–183.

Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics. 1973b:185–203.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology. 1974; 66(5):688.

Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association. 1979; 74(366a):318–328.

Rubin DB. Randomization analysis of experimental data: The fisher randomization test comment. Journal of the American Statistical Association. 1980; 75(371):591–593.

Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association. 2000; 95(450):573–585.

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association. 1999; 94(448):1096–1120.

Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics. 2010; 25(1):1. [PubMed: 20871802]

Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. Developmental psychology. 2008; 44(2):395. [PubMed: 18331131]

Stuart EA, Jo B. Assessing the sensitivity of methods for estimating principal causal effects. Statistical methods in medical research. 2015; 24(6):657–674. [PubMed: 21971481]

Zanutto E, Lu B, Hornik R. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. Journal of Educational and Behavioral Statistics. 2005; 30(1):59–73.

Zanutto EL. A comparison of propensity score and linear regression analysis of complex survey data. Journal of data Science. 2006; 4(1):67–91.

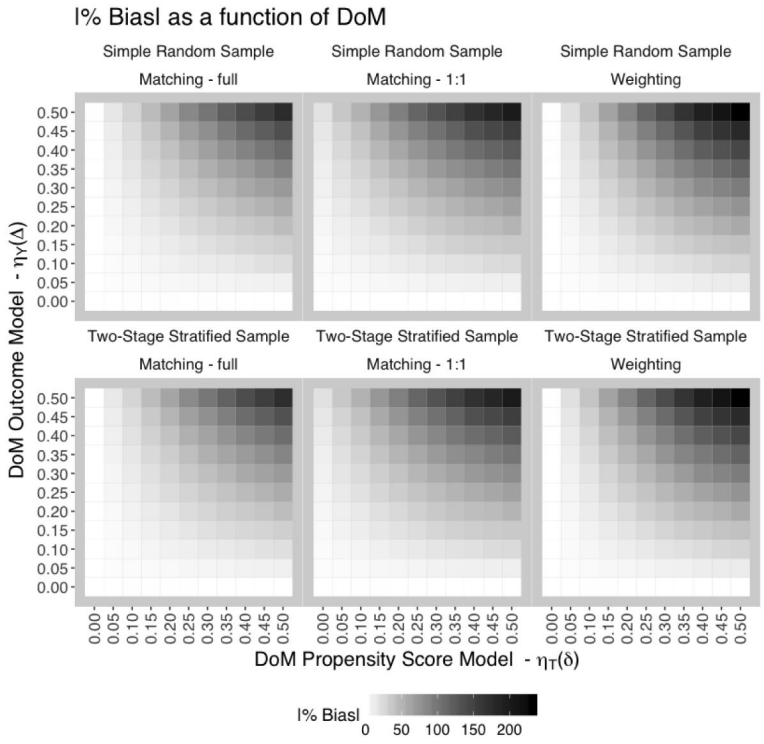

Figure 1. % Bias

% Bias in absolute value associated with the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_{δ}), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

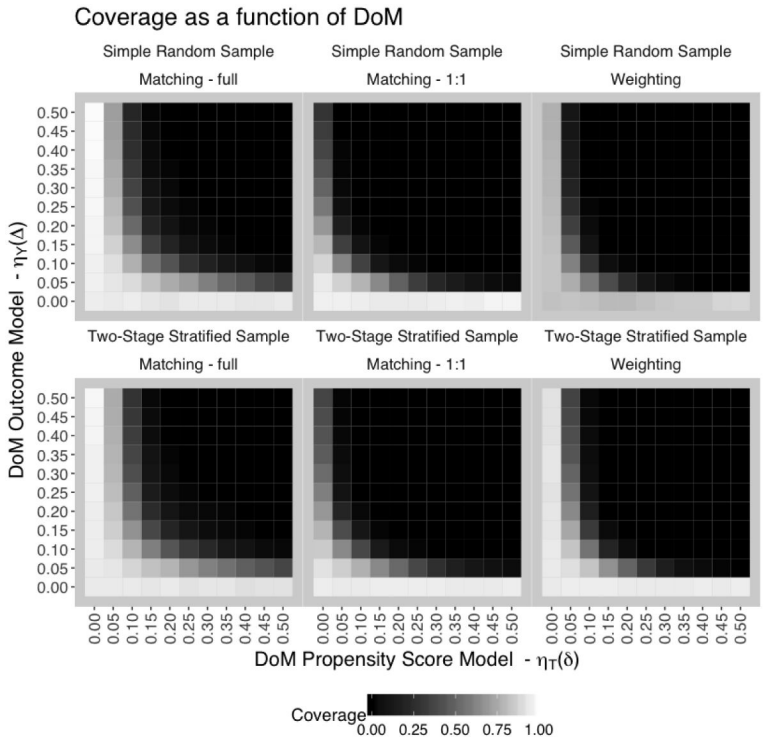

Figure 2. Coverage

Empirical coverage of the 95 interval in the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_{δ}), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).

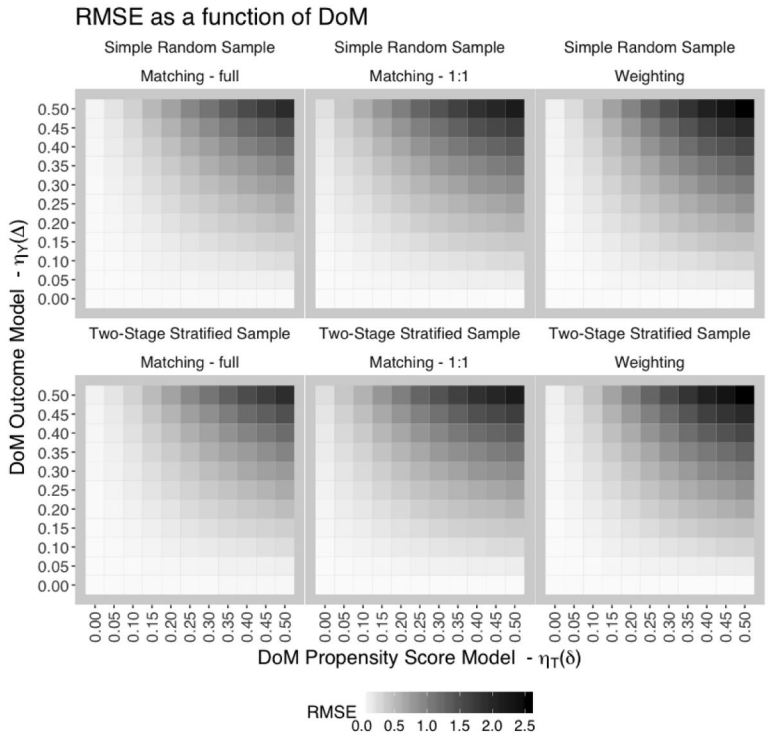

Figure 3. RMSE

RMSE associated with the estimation of γ as a function of the Degree of Misspecification of: (1) the Propensity Score Model (η_{δ}), and (2) the Outcome Model ($\eta_{\Delta(\delta)}$) (simulation study).