Harvard University

Harvard University Biostatistics Working Paper Series

Year 2016 *Paper* 206

A Cautionary Note on the Effect of Treatment Misclassification on the Average Treatment Effect

Danielle Braun[*] Corwin Zigler[†]

Malka Gorfine[‡] Francesca Dominici[**]

[*] Harvard University, dbraun@mail.harvard.edu

[†] Harvard University, czigler@hsph.harvard.edu

[‡] Tel-Aviv University, gorfinem@post.tau.ac.il

[**] Harvard University, fdominic@hsph.harvard.edu

A Cautionary Note on the Effect of Treatment Misclassification on the Average Treatment Effect

Danielle Braun, Corwin Zigler, Malka Gorfine, and Francesca Dominici

Abstract

Comparative effectiveness research often relies on large administrative data, such as claims data. Methods to estimate treatment effects assume that treatment assignment is error-free, but in reality the inaccuracy of procedural or billing codes frequently misclassifies patients into treatment groups. Propensity score methods are widely used to analyze observational studies in which patient characteristics might not be balanced by treatment group. We evaluate the impact of treatment misclassification on 1) propensity score estimation; 2) treatment effect estimation conditional on propensity score estimation and implementation. We focus on three common propensity score implementations: subclassification, matching, and inverse probability of treatment weighting (IPTW). We show in simulations that there is a clear relationship between the misclassification rate and the bias introduced to both the propensity score and treatment effect estimates, and that even when both specificity and sensitivity are relatively high (around 90%) the average treatment effect is biased. We briefly illustrate the impact of misclassification using SEER-Medicare data on brain cancer.

A Cautionary Note on the Effect of Treatment Misclassification on the Average Treatment Effect

ABSTRACT

Comparative effectiveness research often relies on large administrative data, such as claims data. Methods to estimate treatment effects assume that treatment assignment is error-free, but in reality the inaccuracy of procedural or billing codes frequently misclassifies patients into treatment groups. Propensity score methods are widely used to analyze observational studies in which patient characteristics might not be balanced by treatment group. We evaluate the impact of treatment misclassification on 1) *propensity score estimation*; 2) *treatment effect estimation conditional on propensity score estimation and implementation*. We focus on three common propensity score implementations: subclassification, matching, and inverse probability of treatment weighting (IPTW). We show in simulations that there is a clear relationship between the misclassification rate and the bias introduced to both the propensity score and treatment effect estimates, and that even when both specificity and sensitivity are relatively high (around 90%) the average treatment effect is biased. We briefly illustrate the impact of misclassification using SEER-Medicare data on brain cancer.

KEYWORDS

Causal Inference; Comparative Effectiveness Research; Measurement Error; Observational Data; Propensity Score; Validation Data.

INTRODUCTION

When comparing the effectiveness of clinical therapies as they are employed in routine practice, observational studies are often the only feasible option, but have limitations. Absent the benefits of randomization, differences in patient characteristics may confound treatment comparisons. Propensity score methods have gained widespread use for confounding adjustment in these settings (Rosenbaum and Rubin, 1983).

The propensity score can be used in various procedures for estimating causal effects. Analysis using propensity scores consists of three sequential stages (Harder *and others*, 2010); 1) estimating the propensity score, 2) implementing the propensity score to create pseudopopulations of observations where treatment groups have similar values of the estimated propensity score, and 3) comparing outcomes among treated and untreated observations following the implementation stage. In this work, we consider three common implementations; subclassification, matching, and inverse probability of treatment weighting (IPTW)

Existing methods assume that treatment assignment is measured without error, but in reality treatment assignment in observational studies is often measured with error, especially in analyses of claims data and in analyses of electronic medical records (Whittle *and others*, 1991; Du *and others*, 1999; Orrico, 2008). A literature review conducted by Jurek *and others* (2006), shows that measurement error in the exposure (in this context, treatment assignment) is often ignored. Causal inference regarding the effect of an exposure may be biased by these errors (Hernán and Cole, 2009).

Braun *and others* (2015) describe in detail the impact of the misclassified treatment on the three stages of the analysis, and discuss the complexities that arise due to the sequential nature of the three stages. They show that measurement error in T always impacts the propensity score estimation stage and the outcome analysis stage, but that it may or may not impact the implementation stage depending on how the propensity score is used. They do not evaluate the impact of the treatment misclassification rate on the propensity score and treatment effect estimates, which is the focus of this work.

The paper is structured as follows; the methods section describes in detail the notation and model formulations. The results section includes comprehensive simulation studies to evaluate the impact of the misclassification rate on the 1) propensity score estimation, and 2) treatment effect estimation, as well as an illustration of the impact of treatment misclassification in the context of brain cancer data. Finally, we summarize the main results in the discussion.

METHODS

Notations

Let Y denote the true outcome (ex: binary disease status which is the focus of this work, or a continuous, categorical, or survival outcome), T denote a true binary treatment (ex: surgical assignment based on SEER procedural codes), T_{ep} denote the error-prone binary treatment (ex: surgical assignment based on ICD9 billing codes from Medicare Part A), and $\mathbf{X}$ denote a vector of observed confounders measured without error (ex: age, co-morbidity, etc).

The target estimand is the average treatment effect (ATE); $\Delta = E[Y_1] - E[Y_0]$, where Y_0 is the outcome an individual would have had if he/she were untreated, and Y_1 is the outcome an individual would have had if he/she were treated. The observed outcome, Y, is defined as $Y = TY_1 + (1-T)Y_0$. The causal effect Δ can be estimated from the observed data provided that 1) $(T, Y, \mathbf{X})$ are measured without error, 2) each individual has a positive probability of being either treated or untreated: $0 < P(T = 1|\mathbf{X}) < 1$ for all $\mathbf{X}$, 3) the treatment assignment is ignorable, so that, conditional on $\mathbf{X}$, the potential outcomes are independent of T: $(Y_0, Y_1) \perp T|\mathbf{X}$. Under these assumptions, propensity scores can be used to estimate Δ with observed-data comparisons among treated and untreated observations with similar covariate profiles. Settings with treatment misclassification present a challenge since T_{ep} is observed instead of T, and therefore it is unknown as whether any given observed value of Y represents Y_0 or Y_1. This will lead to biased estimation of Δ.

Propensity score estimation

To model the true propensity score, we consider a Generalized Linear Model (GLM) relating T to $\mathbf{X}$, that is $PS_{true} = E(T|\mathbf{X} = \mathbf{x}, \gamma) = g^{-1}(\gamma_0 + \gamma_1^T \mathbf{x})$, where g is known. We consider this the gold-standard propensity score estimate, which could be estimated directly if T were known. Absent information on T, an analogous error-prone propensity score can be modeled with: $PS_{ep} = E(T_{ep}|\mathbf{X} = \mathbf{x}, \gamma_{ep}) = g^{-1}(\gamma_{0ep} + \gamma_{1ep}^T \mathbf{x})$. The propensity score estimates derived from these models will be denoted with $\widehat{PS_{true}}$ and $\widehat{PS_{ep}}$ respectively.

Treatment effect estimation

We consider outcome analyses that can be expressed via a likelihood function, where the likelihood function will be considered conditional on the propensity score implementation stage. To model the true outcome model, we consider a GLM relating Y to T and $\mathbf{X}$, that is $E(Y|T = t, \mathbf{X} = \mathbf{x}, \beta) = r^{-1}(\beta_0 + \beta_1 t + \beta_2^T \mathbf{x}), t = 0, 1, \mathbf{x} \in R^p$, where r is known. We consider this the gold-standard outcome model, which could be estimated directly if T were known. Absent information on T, an analogous error-prone outcome model can be modeled with: $E(Y|T_{ep} = t_{ep}, \mathbf{X} = \mathbf{x}, \beta_{ep}) = r^{-1}(\beta_{0ep} + \beta_{1ep} t_{ep} + \beta_{2ep}^T \mathbf{x}), t_{ep} = 0, 1, \mathbf{x} \in R^p$, where r is known.

For the implementations we consider, this GLM will be estimated either in subclasses defined by the estimated propensity scores or via a likelihood that is weighted by a function of the propensity score. Note that fitting this regression model to adjust for $\mathbf{X}$ in the outcome stage in addition to the adjustment for propensity score in the implementation stage is not necessary since the propensity score implementation would, in principle adjust for confounding due to $\mathbf{X}$. While we conduct all analyses assuming $\beta_2 = 0$ to exclude $\mathbf{X}$ from the outcome model, $\beta_2 \neq 0$ could be included to adjust for residual imbalances not captured by the propensity score implementation or to improve precision of causal estimates (Stuart, 2010).

RESULTS

Simulation design

Our goal is to evaluate the impact of the misclassification rate on the 1) propensity score estimation, and 2) treatment effect estimation, under three propensity score implementations; subclassification, matching, and IPTW.

We consider a measurement error model that can be written as $E(T_{ep}|T, \mathbf{X}) = h^{-1}(\eta_0 + \eta_1 T + \eta_2^T \mathbf{X})$, where h is known. We generate a dataset with $3,000$ individuals. For each we consider six confounders $\mathbf{X} = (1, X_1, \ldots, X_6)$ that include both continuous $(X_1 \sim N(0,1), X_2 \sim N(0,2), X_3 \sim N(0,3))$ and binary covariates $(X_4, X_5, X_6$ each generated from $Bern(0.5))$. $[T|\mathbf{X}, \gamma]$, $[T_{ep}|T, \mathbf{X}, \eta]$, and $[Y|T, \mathbf{X}, \beta]$ were generated as Bernoulli random variables, with
$\gamma = (\gamma_0, 0.3, -0.3, -0.3, 0.3, -0.3, 0.3)^T$, $\eta = (\eta_0, \eta_1, \eta_2, 0, 0, \eta_5, 0, 0)^T$, and $\beta = (0, -2, -1, 1, 1, -1, 1, 1)^T$.

We consider three different scenarios for $\gamma_0 = -1.55, -0.15, 1.2$ corresponding to three different prevalences of the true treatment, $P(T = 1) = 25\%, 50\%, 75\%$. For the evaluation of the impact on the propensity score estimation, we vary η_0 and η_1 in increments of 0.4 from -6 to 6 to obtain scenarios covering a wide range of sensitivity and specificity. Thus, we consider 961 different simulation scenarios for each of the prevalances. For evaluation of the impact on the treatment effect estimation (which is more computationally demanding), we selected 9 of these combinations of η_0 and η_1 representing the following sensitivity/specificity; $(0.1/0.1, 0.2/0.2, \ldots, 0.9/0.9)$. We consider two different scenarios; $\eta_2 = \eta_5 = -0.4$ and $\eta_2 = \eta_5 = -2$ corresponding to two different strengths of association between T_{ep} and $\mathbf{X}$. For the treatment effect estimation each of these 9 scenarios was simulated 500 times.

For subclassification we grouped observations into quartiles of the propensity score. For matching, we use full matching to obtain weights for each individual using the MatchIt R package. For IPTW, weights are calculated using the inverse propensity score and stabilized by multiplying the weights for the treated individuals by the expected value of being treated, and those for the untreated individuals by the expected value of being untreated (Robins *and others*, 2000). For the gold-standard approach, IPTW weights are stabilized based on the true treatment assignment, whereas for the error-prone they are stabilized based on the error-prone treatment assignments. All analysis was conducted in R.

Simulation results

We first evaluate the impact of the misclassification rate on the propensity score estimation, by plotting the absolute value of the difference between $\widehat{PS_{true}}$ and $\widehat{PS_{ep}}$ for each of the 961 simulations considered for $\eta_2 = \eta_5 = -0.4$ (Figure 1) (results for $\eta_2 = \eta_5 = -2$ are similar and not shown). For the first scenario, where $P(T = 1) = 25\%$, there are larger differences between the true and error-prone propensity scores when sensitivity is high and specificity is low, and smaller differences when sensitivity is low yet specificity is high. This is expected, as sensitivity impacts those who are treated whereas specificity impacts those who are untreated, and the proportion of treated observations is smaller than untreated. The reverse is seen for the third scenario, where $P(T = 1) = 75\%$, and there are larger differences when sensitivity is low and specificity is high. For the second scenario, where $P(T = 1) = 50\%$, there is symmetry of the impact of sensitivity and specificity, which is expected.

Next we evaluate the impact of the misclassification rate on the treatment effect estimate. For each of the three propensity score implementations considered we estimate the ATE for the 9 sensitivity/specificity combinations, as well as under the assumption of no measurement error (Figure 2, $\eta_2 = \eta_5 = -0.4$). We also compare the results to standard regression. As expected the ATE under no measurement error is very close to the true estimate (in red), across the three implementations and for standard regression. The bias in the ATE across all three implementations and for standard regression decreases as we increase the sensitivity and specificity across all three prevalences. The misclassification of treatment does not introduce additional bias when implementing using propensity score compared to standard regression. As we increase the dependence of T_{ep} on $\mathbf{X}$, Figure 3, $\eta_2 = \eta_5 = -2$, we see that this is no longer true. The misclassification of treatment increases the variability of the estimates for matching and IPTW compared to stratification and standard regression.

We selected 27 additional combinations corresponding to sensitivity/specificity of $(0.9/0.1, 0.8/0.2, \ldots, 0.1/0.9)$, $(0.5/0.1, \ldots, 0.5/0.9)$, $(0.1/0.5, \ldots, 0.9/0.5)$ which are described in detail in the supplementary materials. Briefly, for both $\eta_2 = \eta_5 = -0.4$ and $\eta_2 = \eta_5 = -2$, we see that for $(0.9/0.1, 0.8/0.2, \ldots, 0.1/0.9)$ the bias in the ATE remains constant across these sensitivity/specificity combinations, implying that the total error.. We see that as expected for a fixed sensitivity or specificity of 0.5 the bias decreases as we vary the specificity or sensitivity from 0.1 to 0.9. In addition, we consider $\beta = (0, -1, -2, 2, 2, -2, 2, 2)^T$ representing a weaker association between the outcome and the treatment and strong association between the outcome and confounders. As expected the bias introduced by the misclassification is not as substantial under this setting.

Data illustration

We illustrate the impact of misclassification using SEER-Medicare data to estimate the effect of biopsy vs. resection on 1-year mortality among patients diagnosed with brain tumors. ICD9 billing codes from Medicare Part A inaccurately reflect surgical treatment, but additional data from SEER is available which contains more accurate information regarding surgical treatment. This data is described in detail in Braun *and others* (2015), briefly, T_{ep} is based on ICD9 codes, and T is based on more accurate medical chart review. The majority of the patients that had a resection according to SEER procedural codes were billed as such according to ICD9 codes (sensitivity, $P(T_{ep} = 1|T = 1) : 96.8\%$), but patients with SEER procedural codes indicating a biopsy often have ICD9 codes indicating resection (specificity, $P(T_{ep} = 0|T = 0) : 26.2\%$).

Confounders with at least 2% prevalence in were selected to be included in the propensity score model (Braun *and others*, 2015). For stratification, the ATE based on T was $-0.02[-0.07, 0.01]$ and based on T_{ep} was $-0.11[-0.17, -0.04]$. For matching, the ATE based on T was $-0.12[-0.09, 0.05]$ and based on T_{ep} was $-0.14[-0.18, -0.02]$. For IPTW, the ATE based on T was $-0.03[-0.07, 0.02]$ and based on T_{ep} was $-0.11[-0.17, -0.04]$. Under this setting, with a high sensitivity and a low specificity, we see that under stratification and IPTW, we have substantial bias in the ATE. An ATE of -0.11 implies that the probability of dying within one year is 11% larger for those who received a biopsy compared to those who had a resection. Thus, under stratification, compared to truth the error-prone indicates that surgery is 9% more effective for preventing death within one year of diagnosis.

DISCUSSION

In this work we show there is a clear relationship between the misclassification rate and the propensity score estimation which depends on the prevalence of the true treatment. When estimating the treatment effect, substantial bias is introduced across all sensitivity/specificity combinations considered which increases as we increase the misclassification rates. For the treatment effect estimation the misclassification in treatment impacts the propensity score estimation, implementation, and the final estimation of the treatment effect conditional on the implementation. Even scenarios where there are relatively small differences in propensity score estimation, can yield large bias in treatment effect due to the sequential nature of these stages.

Treatment assignment is often measured with error. We have shown that this misclassification can introduce substantial bias in the treatment effect estimator. We illustrated using real data that this can lead to dramatically different conclusions. Based on these results, there is a clear need for methods to adjust for measurement error under this complex setting. Braun *and others* (2015) provide a likelihood-based approach, but other approaches may also be considered.

ACKNOWLEDGEMENTS

We gratefully acknowledge support from the National Cancer Institute at the National Institutes of Health [5T32CA009337-32] which supported work of Danielle Braun; [P01 CA134294] which supported work of Danielle Braun, Corwin Zigler, and Francesca Dominici; [R01 GM111339, R01 ES024332, R35CA197449] which supported work of Corwin Zigler and Francesca Dominici; and by the Agency for Healthcare Research and Quality [K18 HS021991] which supported work of Danielle Braun and Francesca Dominici. Work of Malka Gorfine is supported by the Israel Science Foundation (ISF) grant 2012898. *Conflict of Interest*: None declared.

References

Braun, Danielle, Gorfine, Malka, Parmigiani, Giovanni, Arvold, Nils A, Dominici, Francesca and Zigler, Corwin. (2015). Propensity scores with misclassified treatment assignment: a likelihood-based adjustment.

Du, Xianglin, Freeman, Jean L and Goodwin, James S. (1999). Information on radiation treatment in patients with breast cancer: the advantages of the linked medicare and seer data. *Journal of clinical epidemiology* **52**(5), 463–470.

Harder, Valerie S, Stuart, Elizabeth A and Anthony, James C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods* **15**(3), 234.

Hernán, Miguel A and Cole, Stephen R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology* **170**(8), 959–962.

Jurek, Anne M, Maldonado, George, Greenland, Sander and Church, Timothy R. (2006). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European journal of epidemiology* **21**(12), 871–876.

Orrico, Kathleen B. (2008). Sources and types of discrepancies between electronic medical records and actual outpatient medication use. *J Manag Care Pharm* **14**(7), 626–631.

Robins, James M, Hernán, Miguel Ángel and Brumback, Babette. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**(5), 550–560.

Rosenbaum, Paul R and Rubin, Donald B. (1983). The central role of the

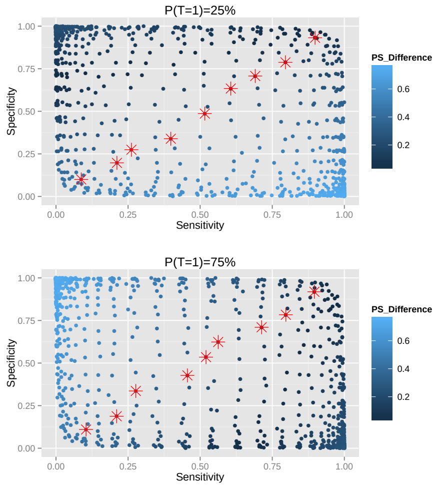

Figure 1: Difference between true and error-prone propensity score for various rates of sensitivity specificity, under three different prevalences, $\eta_2 = \eta_5 = -0.4$.

propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55.

Stuart, Elizabeth A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**(1), 1.

Whittle, Jeff, Steinberg, Earl P, Anderson, Gerard F and Herbert, Robert. (1991). Accuracy of medicare claims data for estimation of cancer incidence and resection rates among elderly americans. *Medical care*, 1226–1236.

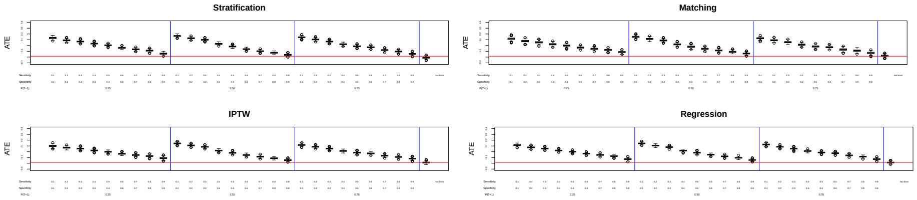

Figure 2: ATE estimator, across 500 simulations for sample size $N = 3,000$, based on varying rates of sensitivity and specificity. Also included are results when there is no error. The true ATE is marked in red. $\eta_2 = \eta_3 = -0.4$, $\beta = (0, -2, -1, 1, 1, -1, 1, 1)^T$. Sens/spec=0.1/0.1,...,0.9/0.9.

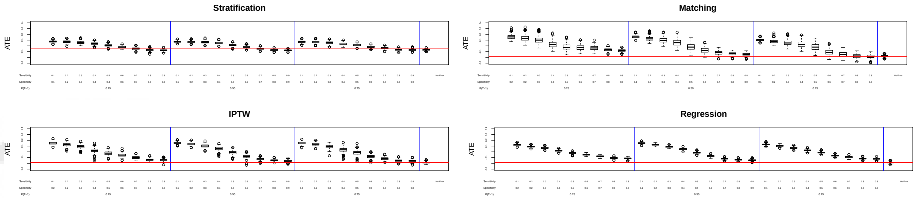

Figure 3: ATE estimator, across 500 simulations for sample size $N = 3,000$, based on varying rates of sensitivity and specificity. Also included are results when there is no error. The true ATE is marked in red. $\eta_2 = \eta_3 = -2$, $\beta = (0, -2, -1, 1, 1, -1, 1, 1)^T$. Sens/spec=0.1/0.1,...,0.9/0.9.