Harvard University
Harvard University Biostatistics Working Paper Series

Year 2016 *Paper* 201

Using Validation Data to Adjust the Inverse Probability Weighting Estimator for Misclassified Treatment

Danielle Braun* Corwin Zigler†
Francesca Dominici‡ Malka Gorfine**

*Harvard University, dbraun@mail.harvard.edu
†Harvard University
‡Harvard University
**Tel-Aviv University

Using Validation Data to Adjust the Inverse Probability Weighting Estimator for Misclassified Treatment

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

Abstract

The inverse probability weighting (IPW) estimator is widely used to estimate the treatment effect in observational studies in which patient characteristics might not be balanced by treatment group. The estimator assumes that treatment assignment, is error-free, but in reality treatment assignment can be measured with error. This arises in the context of comparative effectiveness research, using administrative data sources in which accurate procedural or billing codes are not always available. We show the bias introduced to the estimator when using error-prone treatment assignment, and propose an adjusted estimator using a validation study to eliminate this bias. In simulations, we explore the impact of the misclassified treatment assignment on the estimator, and compare the performance of our adjusted estimator to an estimate based only on the validation study. We illustrate our method on a comparative effectiveness study assessing surgical treatments among Medicare beneficiaries, diagnosed with brain tumors. We use linked SEER-Medicare data as our validation data, and apply our method to Medicare Part A hospital claims data where treatment is based on ICD9 billing codes, which do not accurately reflect surgical treatment.

Using Validation Data to Adjust the Inverse Probability Weighting Estimator for Misclassified Treatment

Danielle Braun[1,2,*], Corwin Zigler[2], Francesca Dominici[2], and Malka Gorfine[3]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[3]Department of Statistics, Tel-Aviv University, Tel-Aviv, Israel.

Running title: Adjusting IPW for Misclassified Treatment

Contact person:

Danielle Braun

Harvard T.H. Chan School of Public Health

677 Huntington Avenue, SPH2, 4th Floor

Boston, MA 02115

United States

(617)-582-7228

dbraun@hsph.harvard.edu

Abstract

The inverse probability weighting (IPW) estimator is widely used to estimate the treatment effect in observational studies in which patient characteristics might not be balanced by treatment group. The estimator assumes that treatment assignment, is error-free, but in reality treatment assignment can be measured with error. This arises in the context of comparative effectiveness research, using administrative data sources in which accurate procedural or billing codes are not always available. We show the bias introduced to the estimator when using error-prone treatment assignment, and propose an adjusted estimator using a validation study to eliminate this bias. In simulations, we explore the impact of the misclassified treatment assignment on the estimator, and compare the performance of our adjusted estimator to an estimate based only on the validation study. We illustrate our method on a comparative effectiveness study assessing surgical treatments among Medicare beneficiaries, diagnosed with brain tumors. We use linked SEER-Medicare data as our validation data, and apply our method to Medicare Part A hospital claims data where treatment is based on ICD9 billing codes, which do not accurately reflect surgical treatment.

Keywords

Causal Inference; Comparative Effectiveness Research; IPW Estimator; Measurement Error; Propensity Score; Treatment Assignment; Validation Data.

1 Introduction

There is a lot of interest in estimating causal treatment effects. Ideally, randomized control studies would be used to study treatment effects, but not always these are feasible due to ethical reasons, cost, time constraints, and compliance (among other reasons). Observational studies are more widely available, however, involve some limitations. Since subjects are not randomized by treatment, their characteristics might not be balanced by treatment group. In order to overcome this limitation, propensity score based methods have been proposed (Rosenbaum and Rubin, 1983).

The propensity score is defined as the probability that an individual has been assigned to treatment given their covariates. Various propensity score methods have been introduced including subclassifying, matching, or weighing individuals by their propensity scores. Rosenbaum and Rubin (1984) introduce a method that stratifies individuals based on their propensity score, and averages the treatment effect across strata. Matching individuals by their propensity scores attempts to create treated and control groups that have similar covariate values. Propensity scores can also be used to weigh individuals observations (Rosenbaum, 1987). The focus of this work is on the inverse probability weighting (IPW) estimator, which estimates the treatment effect, by weighing treated individuals by their inverse propensity score and untreated individuals by the inverse of one minus their propensity score (Rosenbaum, 2005).

The IPW estimator assumes that treatment assignment is measured without error, but in reality treatment assignment in observational studies could be measured with error. Treatment assignment in this context, can be thought of as the exposure variable. Standard techniques adjusting for measurement error in exposures cannot be applied directly to this setting. Misclassification of treatment assignment will lead to both error in the exposure variable directly, as well as error in the propensity score estimates. Previous literature using propensity score methods has largely focused on measurement error in confounders and missing confounders (McCaffrey *and others*, 2013; McCandless *and others*, 2012; Stürmer *and others*, 2005; Webb-Vargas *and others*, 2014).

Measurement error in the treatment assignment has been previously considered by Braun *and others* (2014). They focus on likelihood-based methods to estimate the treatment effect, and con-

sider three propensity score implementations (subclassificaiton, matching, and inverse probability weighting). Under each of these implementations they describe in detail the impact of the misclassified treatment on the three stages of the analysis (propensity score estimation, propensity score implementation, and outcome analysis) and propose a likelihood-based approach to adjust for the misclassification.

In contrast to Braun *and others* (2014) the focus of this work is on estimating the treatment effect using the IPW estimator. Babanezhad *and others* (2010) consider a similar setting of mismeasured exposures. Using estimating equations they derive the asymptotic bias for the IPW and doubly robust estimators. This work differs from theirs in three main ways: 1) we do not approach the problem using estimating equations as they do, instead we look at the estimator directly; 2) their goal was to estimate the causal effect of a binary exposure, T and for an outcome, Y, they focused on estimating the expected contrast $\beta^* = E(Y_1 - Y_0)$ (where $Y_1 = Y$ when $T = 0$ and $Y_0 = Y$ when $T = 1$) based on the marginal structure mean model (Hernán *and others*, 2002): $E(Y_t) = \beta^* t + \alpha^* t = 0, 1$ where $\alpha^* = E(g(x))$ and $g(X)$ is an unknown function of the covariates X. In contrast, our proposed approach does not use any model and estimates the conditional effect $\Delta = E(Y|T = 1, X) - E(Y|T = 0, X)$ directly. Thus, while they use a linear model to relate outcome Y and the exposure T, our theoretical analysis is not restricted to a specific type of outcome or model, and is therefore applicable to any outcome model.

The motivating data application is previously discussed in Braun *and others* (2014). Briefly, our interest is in studying the treatment effect of two types of surgery (resection versus biopsy) among Medicare beneficiaries diagnosed with malignant neoplasm of brain (brain tumors). We are able to obtain Medicare Part A Hospital claims data, which is a large data set (41,971 individuals) containing information on our outcome of interest (1-year mortality), the treatment assignment (the determination of whether resection or a biopsy was preformed), as well as many confounders (including age, gender, co-morbidities, etc). However, treatment assignment in this data set, is based on ICD9 billing codes which are inaccurate. For a subset of individuals (5,463 individuals), we are able to obtain data from SEER-Medicare, a cancer surveillance database with detailed clinical information. Treatment assignment based on the procedural codes from SEER-Medicare is

assumed to be more accurate, gold-standard (Chawla *and others*, 2014; Cooper *and others*, 2000; Du *and others*, 2000). The SEER registry data managers review patient clinical charts to ascertain these sort of differences, and do not use claims data in order to appropriately assign treatment (biopsy vs. resection) received. The SEER-Medicare is our internal validation study.

Using a validation study, we propose an adjusted IPW estimator to estimate the treatment effect in the main study. In Section 2 we introduce general notation, and then define the IPW estimator and proposed adjusted estimator in Section 3. We perform simulations in Section 4, and apply our method to data application in Section 5. Finally, we summarize the main results in Section 6.

2 General Notation

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

We assume the main study (ex: Medicare Part A enrollees) consists of $i = 1, ..., N_m$ individuals, for which we observe Y, T_{ep} and $\mathbf{X}$. We assume the validation study (ex: SEER-Medicare enrollees) consists of $j = 1, ..., N_v$ individuals, where typically $N_v < N_m$, for which we observe $T, T_{ep}, \mathbf{X}$ and Y. We define $N = N_m + N_v$. Note that our work assumes settings of either an internal or external validation study in which Y is observed.

The true propensity score is modeled by 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. Similarly, the error-prone propensity score is modeled with: $PS_{ep} = E(T_{ep}|\mathbf{X} = \mathbf{x}, \gamma_{ep}) = g^{-1}(\gamma_{0ep} + \gamma_{1ep}^T \mathbf{x})$. Propensity scores estimated from these models will be denoted by $\widehat{PS_{true}}$ and $\widehat{PS_{ep}}$ respectively. To model the measurement error model, we consider a GLM relating T to T_{ep} and $\mathbf{X}$, that is $E(T|T_{ep}, \mathbf{X}) = h^{-1}(\theta_0 + \theta_1 T_{ep} + \theta_2^T \mathbf{X})$ and let $\pi_{m|n,\mathbf{X}} = P(T = n|T_{ep} = m, \mathbf{X})$ $n, m = 0, 1$.

3 IPW Estimator

Our interest is in estimating the true average treatment effect (ATE) on the outcome Y for the entire population; $\Delta = E[Y|T=1,\mathbf{X}] - E[Y|T=0,\mathbf{X}]$. Δ can be estimated from the observed data under the following three assumptions: 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: $(Y_0, Y_1) \perp T|\mathbf{X}$ (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).

3.1 Gold-Standard IPW Estimator

The IPW estimator gives an estimate of the treatment effect, by weighing treated individuals by their inverse propensity score and untreated individuals by the inverse of one minus their propensity score (Lunceford and Davidian, 2004; Rosenbaum, 2005). If the true treatment assignment, T, is known, the IPW estimator would be:

$$\hat{\Delta}_{IPW_{true}} = N^{-1}\sum_{k=1}^{N}\frac{T_k Y_k}{\widehat{PS}_{true,k}} - N^{-1}\sum_{k=1}^{N}\frac{(1-T_k)Y_k}{1-\widehat{PS}_{true,k}}$$

The expected value of this estimator is $E(Y_1) - E(Y_0)$ (Lunceford and Davidian, 2004), and thus it provides an unbiased estimate of the true treatment effect.

3.2 Error-prone IPW Estimator

When the true treatment assignment is unknown in the main study, and only information on the error-prone treatment assignment, T_{ep}, is available, IPW estimator is error-prone and can be written as:

$$\hat{\Delta}_{IPW_{ep}} = N^{-1}\sum_{k=1}^{N}\frac{T_{ep,k} Y_k}{\widehat{PS}_{ep,k}} - N^{-1}\sum_{k=1}^{N}\frac{(1-T_{ep,k})Y_k}{1-\widehat{PS}_{ep,k}}$$

We first show that this estimator is biased, and in the following section propose an approach to adjust for the bias in the main study using the validation study. We begin by defining Y in terms of potential outcomes, Y_0 the outcome an individual would have had if he/she were untreated, and Y_1 the outcome an individual would have had if he/she were treated. Y can be written as $Y = TY_1 + (1-T)Y_0$. It follows that $T_{ep}Y = T_{ep}TY_1 + T_{ep}(1-T)Y_0$ and $(1-T_{ep})Y = (1-T_{ep})TY_1 + (1-T_{ep})(1-T)Y_0$.

The expectation of each of the two components of the IPW estimator is taken separately. The expectation of the first component of the estimator is:

$$\begin{aligned}
\mathrm{E}(\frac{T_{ep}Y}{PS_{ep}}) &= \mathrm{EE}(\frac{T_{ep}Y}{PS_{ep}}|Y_1,Y_0,X) = \mathrm{EE}(\frac{T_{ep}TY_1 + T_{ep}(1-T)Y_0}{PS_{ep}}|Y_1,Y_0,X) \\
&= \mathrm{EE}(\frac{T_{ep}TY_1}{PS_{ep}}|Y_1,X) + \mathrm{EE}(\frac{T_{ep}(1-T)Y_0}{PS_{ep}}|Y_0,X) \\
&= E(\frac{Y_1}{PS_{ep}}P(T_{ep}=1,T=1|Y_1,X)) + E(\frac{Y_0}{PS_{ep}}P(T_{ep}=1,T=0|Y_0,X))
\end{aligned}$$

Similarly, the expectation of the second component of the estimator is:

$$\begin{aligned}
\mathrm{E}(\frac{(1-T_{ep})Y}{PS_{ep}}) &= \mathrm{EE}(\frac{(1-T_{ep})Y}{1-PS_{ep}}|Y_1,Y_0,X) = \mathrm{EE}(\frac{(1-T_{ep})TY_1 + (1-T_{ep})(1-T)Y_0}{1-PS_{ep}}|Y_1,Y_0,X) \\
&= \mathrm{EE}(\frac{(1-T_{ep})TY_1}{1-PS_{ep}}|Y_1,X) + \mathrm{EE}(\frac{(1-T_{ep})(1-T)Y_0}{1-PS_{ep}}|Y_0,X) \\
&= E(\frac{Y_1}{1-PS_{ep}}P(T_{ep}=0,T=1|Y_1,X)) + E(\frac{Y_0}{1-PS_{ep}}P(T_{ep}=0,T=0|Y_0,X))
\end{aligned}$$

Thus, overall, the expectation of the error-prone IPW estimator is:

$$\begin{aligned}
E(\Delta_{IPW_{ep}}) &= E(\frac{Y_1}{PS_{ep}}P(T_{ep}=1,T=1|Y_1,X)) + E(\frac{Y_0}{PS_{ep}}P(T_{ep}=1,T=0|Y_0,X)) \\
&- E(\frac{Y_1}{1-PS_{ep}}P(T_{ep}=0,T=1|Y_1,X)) - E(\frac{Y_0}{1-PS_{ep}}P(T_{ep}=0,T=0|Y_0,X))
\end{aligned}$$

Under two additional assumptions; 1) that the measurement error model is independent of outcome; $P(T=n|T_{ep}=m,Y_m,\mathbf{X}) = P(T=n|T_{ep}=m,\mathbf{X}), n=0,1\ m=0,1,\ 2)$ that the

error-prone treatment assignment is ignorable, so that, conditional on $\mathbf{X}$, the potential outcomes are independent of T_{ep}: $(Y_0, Y_1) \perp T_{ep}|\mathbf{X}$. , the expectation of the error-prone IPW estimator in the main study reduces to:

$$E(\Delta_{IPW_{ep}}) = E(Y_1\pi_{1|1,\mathbf{X}}) + E(Y_0\pi_{0|1,\mathbf{X}}) - E(Y_1\pi_{1|0,\mathbf{X}}) - E(Y_0\pi_{0|0,\mathbf{X}}) \quad (1)$$

This estimator is clearly biased, and would be unbiased only when $\pi_{1|1,\mathbf{X}} = 1$ and $\pi_{0|0,\mathbf{X}} = 1$, that is if there is no misclassification. Note, the work considered by Babanezhad *and others* (2010) is a special case of Equation 1, under the assumption that $E(Y_1 - Y_0) = \beta^*$, where β^* is the expected contrast.

3.3 Adjusted IPW Estimator

One proposed approach to eliminate the bias caused by using the error-prone treatment, shown in Equation (1), involves two steps. First, the division of the first component of the estimator in Equation (1) by $\pi_{1|1,\mathbf{X}}$ and the second component of the estimator by $\pi_{0|0,\mathbf{X}}$. The expected value of this new estimator would be:

$$\begin{aligned} &E(\frac{Y_1\pi_{1|1,\mathbf{X}}}{\pi_{1|1,\mathbf{X}}}) + E(\frac{Y_0\pi_{0|1,\mathbf{X}}}{\pi_{1|1,\mathbf{X}}}) - E(\frac{Y_1\pi_{1|0,\mathbf{X}}}{\pi_{0|0,\mathbf{X}}}) - E(\frac{Y_0\pi_{0|0,\mathbf{X}}}{\pi_{0|0,\mathbf{X}}}) \\ &= E(Y_1) + E(\frac{Y_0\pi_{0|1,\mathbf{X}}}{\pi_{1|1,\mathbf{X}}}) - E(\frac{Y_1\pi_{1|0,\mathbf{X}}}{\pi_{0|0,\mathbf{X}}}) - E(Y_0) \end{aligned}$$

This first step is not sufficient to eliminate the bias. Two components, $E(\frac{Y_0\pi_{0|1,\mathbf{X}}}{\pi_{1|1,\mathbf{X}}})$ and $E(\frac{Y_1\pi_{1|0,\mathbf{X}}}{\pi_{0|0,\mathbf{X}}})$ contribute to this remaining bias. The proposed second step involves estimating these two bias components in the validation data, and subtracting them from the overall estimator.

Thus, the proposed adjusted estimator in the main study can be written as:

$$\hat{\Delta}_{IPW_{adj-main}} = N_m^{-1}\sum_{i=1}^{N_m}\frac{T_{ep,i}Y_i}{\widehat{PS}_{ep,i}\widehat{\pi_{1|1,\mathbf{X}}}} - N_m^{-1}\sum_{i=1}^{N_m}\frac{(1-T_{ep,i})Y_i}{(1-\widehat{PS}_{ep,i})\widehat{\pi_{0|0,\mathbf{X}}}}$$

$$
\begin{aligned}
&- \quad N_v^{-1} \sum_{j=1}^{N_v} \frac{(1-T_j)Y_j}{(1-\widehat{PS}_{true,j})} \frac{\widehat{\pi_{0|1,\mathbf{X}}}}{\widehat{\pi_{1|1,\mathbf{X}}}} \\
&+ \quad N_v^{-1} \sum_{j=1}^{N_v} \frac{T_j Y_j}{\widehat{PS}_{true,j}} \frac{\widehat{\pi_{1|0,\mathbf{X}}}}{\widehat{\pi_{0|0,\mathbf{X}}}}
\end{aligned} \tag{2}
$$

The expected value of this estimator is $E(Y_1) - E(Y_0)$, and thus this estimator is unbiased. The first two sums in Equation (2) are calculated for the main study, however the measurement error model $\pi_{1|1,\mathbf{X}}$ and $\pi_{0|0,\mathbf{X}}$ will be estimated in the validation study. Thus, the proposed estimator relies on the transportability assumption. This is a common assumption in the measurement error literature assuming that the measurement error model is transportable from the validation study to the main study.

The overall adjusted IPW estimator for the entire population can be written as:

$$
\begin{aligned}
\hat{\Delta}_{IPW_{adj}} &= \omega(N_m^{-1} \sum_{i=1}^{N_m} \frac{T_{ep,i} Y_i}{\widehat{PS}_{ep,i} \widehat{\pi_{1|1,\mathbf{X}}}} - N_m^{-1} \sum_{i=1}^{N_m} \frac{(1-T_{ep,i})Y_i}{(1-\widehat{PS}_{ep,i}) \widehat{\pi_{0|0,\mathbf{X}}}} \\
&- N_v^{-1} \sum_{j=1}^{N_v} \frac{(1-T_j)Y_j}{(1-\widehat{PS}_{true,j})} \frac{\widehat{\pi_{0|1,\mathbf{X}}}}{\widehat{\pi_{1|1,\mathbf{X}}}} + N_v^{-1} \sum_{j=1}^{N_v} \frac{T_j Y_j}{\widehat{PS}_{true,j}} \frac{\widehat{\pi_{1|0,\mathbf{X}}}}{\widehat{\pi_{0|0,\mathbf{X}}}}) \\
&+ (1-\omega) \left(N_v^{-1} \sum_{j=1}^{N_v} \frac{T_j Y_j}{\widehat{PS}_{true,j}} - N_v^{-1} \sum_{j=1}^{N_v} \frac{(1-T_j)Y_j}{1-\widehat{PS}_{true,j}} \right) \\
&= \omega \hat{\Delta}_{IPW_{adj-main}} + (1-\omega) \hat{\Delta}_{IPW_{true-val}}
\end{aligned} \tag{3}
$$

where $\hat{\Delta}_{IPW_{true-val}}$ is the IPW estimator based on the true treatment assignment in the validation study. One common choice for ω is N_m/N, however, we propose instead to use an $\hat{\omega}$ that minimizes the variance of $\hat{\Delta}_{IPW_{adj}}$. Let V_1 be the variance of $\hat{\Delta}_{IPW_{adj-main}}$, V_2 be the variance of $\hat{\Delta}_{IPW_{true-val}}$, and V_{12} indicate the covariance of the two estimators ($\rho\sqrt{V_1 V_2}$, where ρ is the correlation of the two estimators).

$$
var(\omega \hat{\Delta}_{IPW_{adj-main}} + (1-\omega) \hat{\Delta}_{IPW_{true-val}}) = \omega^2 V_1 + (1-\omega)^2 V_2 + 2\omega(1-\omega) V_{12} = V_\omega \tag{4}
$$

Minimizing Equation 4, we obtain $\hat{\omega} = \frac{V_2 - V_{12}}{V_1 + V_2 - 2V_{12}}$. Therefore we propose using the adjusted estimator in Equation 3 with $\hat{\omega} = \frac{\hat{V_2} - \hat{V_{12}}}{\hat{V_1} + \hat{V_2} - 2\hat{V_{12}}}$. $\hat{V_1}$, $\hat{V_2}$, and $\hat{V_{12}}$ are obtained by bootstrapping with 500 bootstrap samples.

4 Simulations

The goal of our simulations is to study the effects of measurement error in the treatment assignment on the IPW estimator. Performance of our proposed estimator is evaluated by comparing treatment effect estimates for the validation study, main study, and overall population based on true treatment assignment (gold standard, Equation (1)), error-prone treatment assignment (naive, Equation (1)), and the proposed adjusted estimator (Equation (3)).

4.1 Simulation Characteristics

For each simulation scenario we simulated two data sets in a similar manner, one to be used as the main study and one to be used as validation data. We simulated the main study with N_m individuals and the validation study with N_v individuals. We generate two continuous confounders $\mathbf{X} = (1, X_1, X_2)$. T_{ep} was generated as Bernoulli according to the error-prone propensity score $logit(P(T_{ep} = 1|X_1)) = \gamma_0 + \gamma_1 X_1$. T was generated as Bernoulli according to the measurement error model $logit(P(T = 1|T_{ep}, X_2)) = \eta_0 + \eta_1 T + \eta_2 X_2$. Y was generated as Bernoulli according to the outcome model $logit(P(Y = 1|T, X)) = \beta_0 + \beta_1 T + \beta_2 X_2$. $\beta = (\beta_0, \beta_1, \beta_2)^T = (-0.3, 10, -0.6)^T$, so that $\beta_1 = 10$. We consider $\gamma = (\gamma_0, \gamma_1)^T = (0.5, -0.3)^T$. For the measurement error model, we consider $\eta = (\eta_0, \eta_1, \eta_2)^T = (0.2, 10, -0.1)$. In addition, six sample sizes were considered 1) $N_m = 3,000$ and $N_v = 1,500$, 2) $N_m = 5,000$ and $N_v = 1,000$, 3) $N_m = 4,000$ and $N_v = 1,000$, 4) $N_m = 3,000$ and $N_v = 1,000$, 5) $N_m = 2,000$ and $N_v = 1,000$, and 6) $N_m = 2,000$ and $N_v = 500$, . After the data was generated, we applied the analysis strategies described in previous sections. 100 repetitions were performed for each scenario. For each of the 100 repetitions, 500 bootstrap samples were used to obtain estimates of V_1, V_2, and V_{12}.

4.2 Simulation Results

Figure 1 evaluates the effects of measurement error using the IPW estimator. We present detailed results for one sample size combination, $N_m = 5,000$ and $N_v = 1,000$ (Figure 1), the remaining results are reported in the Web Appendix (Web Figures 1-5). From left to right, we plot results based on the validation study, main study, and overall study population. For each study population we calculate the ATE based on the true treatment assignment (GS), and based on the error-prone treatment assignment (EP). For the overall study population, we also calculate the ATE based on the proposed adjustment (Equation 3).

Across all study populations, the ATE is unbiased when using the gold standard treatment assignment, and biased when using the error-prone treatment assignment. Our main interest is in two potential estimators, the first using the gold standard based only on the validation study, and the second is our proposed adjustment using both the main and validation studies (Overall: Adj in Figure 1). Using the gold standard based only on the validation study, the ATE is unbiased, and the variance is 0.001331. The proposed adjusted is also unbiased, and the variance reduces to 0.000775. Thus, our proposed adjustment provides a 42% decrease in variance.

5 Data Application

We apply the adjusted estimator proposed in this paper, to determine the treatment effect of resection versus biopsy on one year mortality for Medicare beneficiaries ages 65 and older, diagnosed with malignant neoplasm of brain between 1999 and 2007. We use the Medicare dataset as our main study, and the SEER-Medicare dataset as our internal validation study. We focus on a subset of the population which had either a resection or biopsy performed. Patients in the SEER-Medicare dataset are also included in the Medicare Part A dataset, but linkage across sources is unidentifiable, so we cannot determine the Part A record corresponding to each SEER-Medicare record. For this reason, we decide to limit the Medicare Part A dataset to individuals diagnosed in even years, and the SEER-Medicare dataset to individuals diagnosed in odd years, thus ensuring that the same individual is not included in both datasets. In addition, to make the two study populations more

comparable, we decide to limit our analysis to patients in Medicare Part A that are in the subset of states included in SEER-Medicare. We also exclude patients diagnosed with other forms of cancer, allowing us to more accurately evaluate the effect of this specific treatment on mortality. The final Medicare Part A dataset used for the analysis consisted of 11,036 patients, and the final SEER-Medicare consisted of 1,582 patients.

We select confounders with at least 2% prevalence in both cohorts for the analysis. This is described in detail in Braun *and others* (2014), and summaries of patient characteristics of the two cohorts are shown in Table 1 (this table also appears in (Braun *and others*, 2014)). Error-prone resection rates are higher in SEER-Medicare compared to Medicare Part A (92.0% vs. 62.8%, $p < 2.2 \times 10^{-16}$), and mortality rates are lower in SEER-Medicare compared to Medicare Part A (72.6% vs. 81.5%, $p < 2.2 \times 10^{-16}$). Patient characteristics across some of the confounders also differ across the cohorts.

Our proposed adjustment assumes the measurement error model is transportable from the validation study to the main study. We hypothesize that errors in the appropriate coding of neurosurgical biopsy vs. resection may be related to important but subtly distinct definitions of biopsy vs. subtotal resection within the brain. These differences are likely not the focus of physicians or administrators handling claims, and thus any erroneous or misclassified ICD-9 codes used for these two procedures should arise similarly for the Medicare component of the Part A dataset as well as the SEER-Medicare linkage. Therefore, we believe that this transportability assumption likely holds since the mechanism driving the error is likely similar in the two populations.

5.1 Data Analysis Results

The measurement error model, $E(T|T_{ep}, \mathbf{X}) = h^{-1}(\theta_0 + \theta_1 T_{ep} + \theta_2^T \mathbf{X})$, where h is $logit$, is estimated in the SEER-Medicare data. We then apply our correction the Medicare Part A dataset. Results from this analysis are shown in Table 2. We compare the ATE based on the error-prone IPW estimator to the proposed adjusted IPW estimator. We see that the ATE stays the same after the adjustment (from -0.11 95% CI: $(-0.12, -0.10)$ to -0.10 95% CI: $(-0.17, -0.03)$. In addition, we conducted a sensitivity analysis to assess how sensitive the proposed adjustment is to the trans-

portability assumption. We estimated $\hat{\theta}$ in the SEER-Medicare validation study, and then sampled 100 different samples, $\widehat{\theta_{sens}}$, from a normal distribution with mean $\hat{\theta}$ and standard deviation equal to standard error estimates of $\hat{\theta}$ multiplied by one, two, or five. These results are presented in Table 3 and show when there is small variability in the estimates of θ (standard error estimates of $\hat{\theta}$ multiplied by one) the adjustment yields similar results, but as variability in θ increases results quickly become less consistent.

6 Discussion

In this paper we present an approach to adjust for measurement error in treatment assignment in observational studies. We derive the bias in the IPW estimator, propose an adjusted estimator, and evaluate the performance of the adjusted estimator in simulations with finite sample sizes.

Simulation studies show that, when we adjust for confounding using IPW, the proposed estimator (defined in Equation (3)) is appropriate and performs as well as the gold standard in terms of bias. In addition, the proposed estimator provides a variance reduction compared to the naive approach of using the gold standard based on the validation study only.

The proposed estimator relies on a transportability assumption, which requires careful thought. In our data application, it is reasonable to assume that the mechanism driving the error is the same for the two populations, as discussed in Section 5. In our data application we assume that treatment assignment based on SEER-Medicare is the gold standard, however, surgical treatment assignment in patients with malignant neoplasm of brain is complex. Surgeons will often decide whether to perform a more extensive tumor resection vs. only a biopsy, which includes removing less tumor tissue, only once they are in the operating room, based on nearby critical brain regions. In addition, the line between these two procedures is often ambiguous, and even among surgeons performing the same procedure there could be discrepancies in coding. Thus, even the SEER-Medicare procedural codes might not be an absolute gold-standard, which is a limitation. However, these codes do provide more accurate information compared to ICD-9 codes, are the best treatment assignment variables available and permit an illustration of our methods showing the impact of the

measurement error adjustment.

The IPW estimator is widely used in the literature. Measurement error in exposure, treatment assignment, is often ignored (Jurek *and others*, 2006). This paper proposes a direct approach to estimate the bias, and eliminate it. It is easy to implement, and can be implemented in a wide range of scenarios in which validation data is available.

Acknowledgments

We gratefully acknowledge support from the National Cancer Institute at the National Institutes of Health [5T32CA009337-32 to Parmigiani] [P01 CA134294 to Lin] and by the Agency for Healthcare Research and Quality [K18 HS021991 to Dominici]. Work of Malka Gorfine is supported by the Israel Science Foundation (ISF) grant 2012898. *Conflict of Interest*: None declared.

References

Babanezhad, Manoochehr, Vansteelandt, Stijn and Goetghebeur, Els. (2010). Comparison of causal effect estimators under exposure misclassification. *Journal of Statistical Planning and Inference* **140**(5), 1306–1319.

Braun, Danielle, Gorfine, Malka, Zigler, Corwin, Dominici, Francesca and Parmigiani, Giovanni. (2014). Adjustment for mismeasured exposure using validation data and propensity scores.

Chawla, Neetu, Yabroff, K Robin, Mariotto, Angela, McNeel, Timothy S, Schrag, Deborah and Warren, Joan L. (2014). Limited validity of diagnosis codes in medicare claims for identifying cancer metastases and inferring stage. *Annals of epidemiology* **24**(9), 666–672.

Cooper, Gregory S, Yuan, Zhong, Stange, Kurt C, Dennis, Leslie K, Amini, Saeid B and Rimm, Alfred A. (2000). Agreement of medicare claims and tumor registry data for assessment of cancer-related treatment. *Medical care* **38**(4), 411–421.

Du, Xianglin, Freeman, Jean L, Warren, Joan L, Nattinger, Ann B, Zhang, Dong and Goodwin, James S. (2000). Accuracy and completeness of medicare claims data for surgical treatment of breast cancer. *Medical care* **38**(7), 719–727.

Hernán, Miguel A, Brumback, Babette A and Robins, James M. (2002). Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in medicine* **21**(12), 1689–1709.

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.

Lunceford, Jared K and Davidian, Marie. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* **23**(19), 2937–2960.

McCaffrey, D. F., Lockwood, J. R. and Setodji, C. M. (2013). Inverse probability weighing with error-prone covariates. *Biometrika* **100**(3), 671–680.

McCandless, Lawrence C, Richardson, Sylvia and Best, Nicky. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association* **107**(497), 40–51.

Rosenbaum, Paul R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**(398), 387–394.

Rosenbaum, Paul R. (2005). Propensity score. *Encyclopedia of biostatistics*.

Rosenbaum, Paul R and Rubin, Donald B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1), 41–55.

Rosenbaum, Paul R and Rubin, Donald B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**(387), 516–524.

Stürmer, Til, Schneeweiss, Sebastian, Avorn, Jerry and Glynn, Robert J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology* **162**(3), 279–289.

Webb-Vargas, Yenny, Rudolph, Kara E, Lenis, D, Murakami, Peter and Stuart, Elizabeth A. (2014). Applying multiple imputation for external calibration to propensty score analysis.

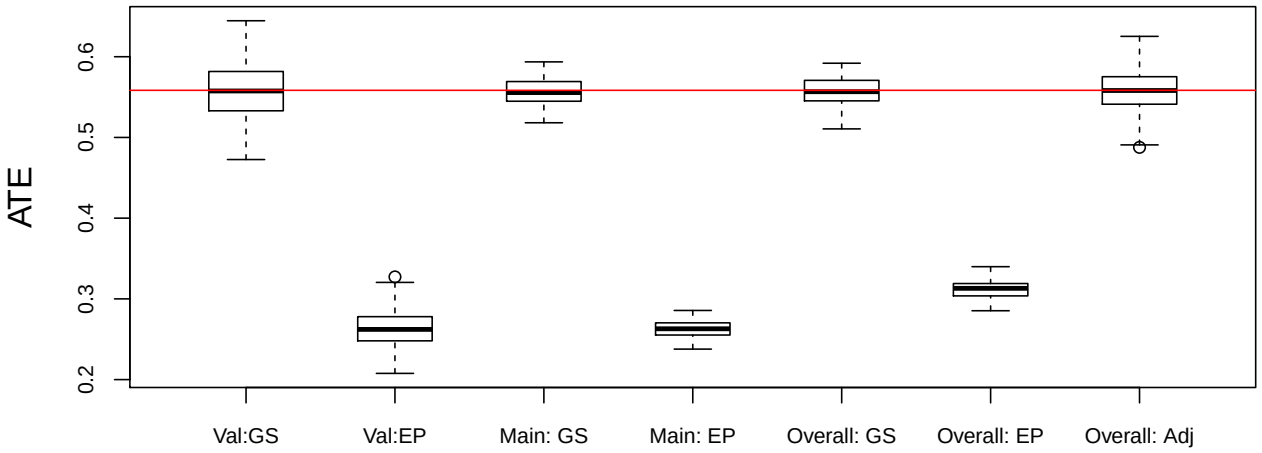

Figure 1: ATE estimator, across 100 simulations for sample size $N_m = 5,000$ and $N_v = 1,000$. GS (Gold Standard) is based on the true treatment assignment. EP (Error Prone) is based on the error-prone treatment assignment. Adj is based on the error-prone treatment assignment with the proposed adjusted estimator, with $\hat{\omega}$. V_1, V_2, V_{12} were estimated based on 500 bootstrap samples for each simulation. The true ATE is marked in red.

Table 1: Medicare Part A and SEER-Medicare Population Characteristics

	Medicare Part A	SEER-Medicare	P-value
Population Size	11,036	1,582	–
1-year mortality events	8,989 (81.5%)	1,148 (72.6%)	$< 2.2 \times 10^{-16}$
Error-prone Resection based on ICD9 codes	6,928 (62.8%)	1,453 (92.0%)	$< 2.2 \times 10^{-16}$
True Resection based on SEER-Medicare	-	1,242 (78.5%)	–
Age	74.77	73.67	8×10^{-12}
Female	47.1%	46.6%	0.69
Dual Eligible	7.4%	6.1%	0.06
Brain MRI	9.1%	15.1%	4×10^{-14}
Head CT	6.6%	6.7%	0.93
Radiotherapy Inpatient ICD9	2.9%	8.2%	$< 2 \times 10^{-16}$
Atherosclerosis	20.5%	14.7%	4×10^{-8}
Substance Abuse	8.0%	6.2%	0.01
Hypertension	60.0%	57.7%	0.09
COPD	12.0%	9.4%	0.002
Dementia	10.2%	5.6%	1×10^{-8}
Trauma	4.2%	3.1 %	0.05
Psychological Disorder	3.7%	2.8%	0.06
Depression	7.0%	5.6%	0.04
Seizure Disorder	20.4%	21.7%	0.24
Asthma	3.1%	2.3%	0.08
Valvular/Rheumatic Heart Disease	5.7%	4.0%	0.005
Diabetes	18.0%	16.4%	0.11
Region: West	15.8%	45.3%	$< 2.2 \times 10^{-16}$
Region: Midwest	26.9%	15.2%	$< 2.2 \times 10^{-16}$
Region: Northeast	19.3%	22.6%	0.002
Region: South	28.0%	16.8%	$< 2.2 \times 10^{-16}$
Admission Type: Emergency	34.9%	33.8%	0.37
Admission Type: Urgent	23.5%	24.2%	0.54
Admission Type: Elective	41.6%	42.0%	0.74
Admission Source: Clinic	52.5%	51.6%	0.51
Admission Source: HMO	4.0%	4.4%	0.42
Admission Source: SNF	10.0%	10.0%	0.96
Admission Source: Court/Law	31.2%	31.9%	0.55
Admission Source: Other	2.3%	2.1%	0.55

Table 2: ATE of resection vs. biopsy in Medicare Part A/SEER-Medicare Data Application, Outcome 1-year mortality

	ATE [95% CI]
SEER/Medicare: true trt	-0.03 [-0.07, 0.03]
Medicare Part A: ep trt	-0.11 [-0.13, -0.10]
Medicare Part A: adj	-0.10 [-0.17, -0.05]

Table 3: Sensitivity analysis for estimating ATE of resection vs. biopsy in Medicare Part A Data Application, Outcome 1-year mortality. After estimating $\hat{\theta}$ in the SEER-Medicare validation study, 100 different samples, $\widehat{\theta_{sens}}$ were sampled from a normal distribution with mean $\hat{\theta}$ and standard deviation equal to standard error estimates of $\hat{\theta}$ multiplied by one, or two.

	Sensitivity Analysis
	ATE [95% CI]
Medicare Part A: Adj. σ	-0.09 [-0.30, 0.12]
Medicare Part A: Adj. $*2\sigma$	0.39 [-2.11, 1.99]