Applying Multiple Imputation for External Calibration to Propensity Score Analysis

Yenny Webb-Vargas[1,*], Kara E. Rudolph[2], David Lenis[1], Peter Murakami[2], Elizabeth A. Stuart[1,2]
[1]Department of Biostatistics and [2]Department of Mental Health,
Johns Hopkins Bloomberg School of Public Health., USA

Abstract

Although covariate measurement error is likely the norm rather than the exception, methods for handling covariate measurement error in propensity score methods have not been widely investigated. We consider a multiple imputation-based approach that uses an external calibration sample with information on the true and mismeasured covariates, Multiple Imputation for External Calibration (MI-EC), to correct for the measurement error, and investigate its performance using simulation studies. As expected, using the covariate measured with error leads to bias in the treatment effect estimate. In contrast, the MI-EC method can eliminate almost all the bias. We confirm that the outcome must be used in the imputation process to obtain good results, a finding related to the idea of congenial imputation and analysis in the broader multiple imputation literature. We illustrate the MI-EC approach using a motivating example estimating the effects of living in a disadvantaged neighborhood on mental health and substance use outcomes among adolescents. These results show that estimating the propensity score using covariates measured with error leads to biased estimates of treatment effects, but when a calibration data set is available, MI-EC can be used to help correct for such bias.

Keywords

Causal Inference; Measurement error; Multiple imputation; Propensity scores

1. Introduction

Propensity score methods are commonly used to estimate causal effects in non-experimental studies, as they help to ensure that treatment and comparison groups are similar with respect to observed covariates [1,2]. Nearly all existing propensity score methods assume that covariates are measured without error. However, in reality, covariate measurement error may be the rule, not the exception. Self-reported measures, latent variables measured using scales, and disease status definitions comprised of surrogate measures are all examples of variables measured with error. If only the error-prone

* Corresponding author; e-mail: yennywebb@gmail.com, address: 615 North Wolfe Street Room E3040, Baltimore, MD 21205, phone: (410)614-5126

version of the variable is available, but the assignment mechanism itself used the true covariate, then the true confounder remains unmeasured.

Preliminary investigations have shown that measurement error can have detrimental effects on the performance of propensity score methods. For example, Steiner et al.[3] show that measurement error can degrade the propensity score's ability to reduce bias, although Millimet[4] shows that the bias is fairly limited if there is relatively little measurement error and if the error itself is not related to treatment group or covariates. McCaffrey et al.[5] have shown formally that propensity scores based on direct use of error-prone covariates will not yield covariate balance on the underlying true covariate, and thus will not provide accurate treatment effect estimates.

A variety of methods for handling measurement error in regression settings have been developed. These include the method of moments[6], regression calibration[7], simulation-extrapolation (SIMEX[8]), and multiple imputation[9,10]. However, there has been very limited work extending these approaches to causal inference settings using propensity score methods, and the considerations may be quite different. For example, established methods are generally concerned with parameter estimation in a specific linear or non-linear model, where primary interest is in the coefficients of the variables that are measured with error. Meanwhile, in the propensity score context we care more about correct predicted values from the propensity score model, and the resulting treatment effect estimates obtained by applying those propensity scores in some way to the outcome analysis. In other words, we care less about how measurement error influences the coefficients on particular variables in a model, and more on how covariate measurement error influences how well we can estimate causal treatment effects.

There are just a few papers that do investigate covariate measurement error in propensity score settings. Sturmer et al.[11] consider an error-prone propensity score that is estimated from a model that ignores an important confounder (e.g., the true covariate, in our scenario) and propose a method called "propensity score calibration" to estimate an updated propensity score that accounts for this unobserved confounding. Propensity score calibration uses an approach related to regression calibration and relies on observing the missing confounders on a subset of the original sample and also assumes non-differential error across treatment groups. Meanwhile, McCaffrey et al.[5] propose a measurement-error bias-corrected inverse probability of treatment weighting estimator. A limitation of both of the Sturmer and McCaffrey approaches is that they assume that the error distribution is the same across treatment groups (i.e., the measurement error is non-differential across groups.) Another approach that addresses covariate imbalance and measurement error is to use a doubly robust method (as in Rotnitzky et al.[12]) and apply existing methods for measurement error to the outcome model but not the propensity score model. However, if the predicted propensity scores are incorrect (i.e., if the measurement error is not accounted for in the propensity score estimation), this would detract from the benefits of using propensity score methods.

Therefore, extensions of existing approaches that account for measurement error when using propensity score methods are needed. In the present work we investigate one such extension. We consider a scenario in which a calibration sample is available, and adapt Multiple Imputation for External Calibration (MI-EC[10]) to correct for covariate measurement error in propensity score estimation and use.

2. Propensity scores and the importance of accurate covariate measurement

The propensity score, first introduced by Rosenbaum and Rubin[1], is defined as the probability of receiving treatment given the observed covariates. Propensity score methods, such as matching, weighting, or subclassification[2], help ensure that the treatment and control groups being compared are as similar as possible on the observed characteristics, and they

often yield more reliable estimates of treatment effects than do traditional methods such as regression adjustment (e.g., see Martens et al.[13]).

We consider the problem of estimating the average treatment effect on an outcome Y, where the treatment assignment T ($T \in \{0, 1\}$) and the outcome Y are affected by a set of confounders (X, Z). First, we define a potential outcome $Y_i(t)$ as the outcome that we would observe if person i receives treatment t. With two treatments (control and treatment) there are two potential outcomes: $Y_i(0)$, the potential outcome if person i receives the control condition, and $Y_i(1)$, the potential outcome if person i receives the treatment. The average treatment effect is a comparision of these potential outcomes, such as $\Delta = E[Y_i(1) - Y_i(0)]$.

We assume that if we assign treatment t to person i we observe their potential outcome for treatment t (the consistency assumption, $[Y_i \mid T_i = t] = Y_i(t)$; for a discussion on the topic see Cole and Frangakis[14]). We also assume that the assigned treatment of one person does not influence the potential outcomes of another one, and that each treatment has only one version, known as the stable unit treatment value assumption (SUTVA[15]).

We focus on non-experimental study designs, in which we assume strong ignorability of the treatment assignment[1]. We assume 1) that each person has a positive probability of getting either treatment – called the positivity assumption, where $P(Z_i = z \mid X_i, Z_i) > 0 \quad \forall z$, and 2) that the treatment assignment is independent of the potential outcomes, given a set of observed covariates X, Z: that is, $T_i \perp\!\!\!\perp Y_i(1), Y_i(0) \mid X_i, Z_i$.

Given that the outcome Y is also influenced by the confounders X, Z, we would like to find groups of people that only differed by which treatment T they received, and did not differ in any other way. Since finding people with the same levels of all possible confounders is difficult in practice, a summary measure that also balances the distribution of confounders across treatment groups is helpful. The propensity score is such a balancing score and creating groups with similar propensity scores makes the distribution of X and Z similar across groups. As noted, there are different methods that use the propensity score to achieve balance across treatment groups[2].

The balancing property of the propensity score relies on having a correct model for the treatment assignment. To be correct, the model must include all the relevant confounders and must have the correct form. If one confounder is measured with error, the true confounder remains partially unobserved. Let W be a covariate that is correlated highly with X, and thus is essentially a measure of X with more error. If we use W in our propensity score model instead of using X, we would not achieve complete balance on X and our estimates would be subject to confounding[5]. Note that regression adjustment that uses W would also be subject to remaining confounding[16].

2.1. Inverse Probability of Treatment Weighting

Inverse Probability of Treatment Weighting (IPTW) is a method that uses the propensity scores to generate weights that, when used when estimating the treatment effect, result in having similar distributions of covariates among treated and control groups. For this work, we estimate the treatment effect using a weighted difference in sample means between treatment and control groups.

We start by posing a model for the treatment T assignment given the confounders X, Z. This model is only used for generating the predicted propensity scores. This allows us to use methods that are parametric – like generalized linear model with a logistic or probit link, or machine learning algorithms like generalized boosted regression models and classification and regression trees[17].

Let $\hat{p}_i = \hat{P}[T_i = 1 \mid X_i, Z_i]$ be the predicted propensity score for person i. IPTW generates weights based on the inverse of the probability of treatment assigned, which allows for estimation of the average treatment effect on the whole population (ATE). Let $\hat{u}_i$ be the weight for person i. To estimate the ATE, individuals in the treatment group receive a weight of $\hat{u}_i = \frac{1}{\hat{p}_i}$, whereas individuals in the control group receive a weight of $\hat{u}_i = \frac{1}{1-\hat{p}_i}$. The weights $\hat{u}_i$ generate a pseudo-population in which the distribution of covariates is the same in the treated and control groups, and the weighted difference estimates the average treatment effect Δ. In contrast, we could be interested in the average treatment effect among the treated population (ATT). In this case, the treated population have unit weights ($\hat{u}_i = 1$ if subject i is treated), while the population in the control group are weighted by the odds of being treated ($\hat{u}_i = \frac{\hat{p}i}{1-\hat{p}_i}$ if subject i is in the control group.)

Note that a linear model on the observed outcome Y that incorporates the correct variables in the correct form (including all non-linearities, and having no measurement error) would estimate Δ correctly, assuming the model for $\hat{p}_i$ is correct. The benefit of using IPTW with covariate adjustment in the outcome model is that this method is doubly robust. It will be unbiased if either (but not necessarily both) the propensity score model or the outcome model is correct[12].

3. Measurement error and propensity score methods

In this paper, we examine the consequences of using a covariate measured with error in the estimation of the average treatment effect using IPTW and we investigate the performance of MI-EC (described in further detail below) in correcting for bias due to this measurement error. We consider a classical measurement error model:

$$W = X + e, \tag{1}$$

where X is the true confounder, W is the error-prone covariate, and e has some distribution that does not depend on X nor Y, and has zero mean[16]. It is common to assume $e \sim N(0, \sigma^2)$. Given the measurement error model, we can define the reliability of the error-prone covariate X to be $r = \frac{var(X)}{var(W)}$.

Our setting relies on two samples (see Figure 1). The first is the main study sample, in which we observe the outcome of interest Y, the treatment assignment T, a set of confounders measured without error Z, and W, the version of the true confounder X that is measured with additive error. The second is the calibration sample, where only (X, W) are observed. This set-up is encountered when using measures that are calibrated against gold standards in studies external to the main study.

A common approach for handling measurement error in general is that of regression calibration[7,16]. This approach involves defining a regression calibration model and using it to predict X with information from W. However, in the calibration setting, we cannot apply regression calibration in a valid manner to our setting of propensity score estimation. For regression calibration to be valid, all confounders of the X and Y relation must be included in the regression calibration model. Since Z is correlated with X and predicts Y, it should be part of the model. But Z is not in the calibration sample, so a model that only uses X and W is not valid.

We consider two approaches for dealing with covariate measurement error in propensity score estimation: a naive method, and multiple imputation for external calibration (MI-EC).

Sample	Observed Data					
	W	*X*	*Z*	*T*	*Y(0)*	*Y(1)*
Calibration	✓	✓				
Main	✓		✓	✓	✓	✓

Fig. 1. Structure of the data. We focus on a setting in which there are two samples, one that contains external information (the calibration sample, with W and X), and one that contains information for the observational study (the main sample, with W, Z, T, Y). We observe $Y(0)$ for the people in the control group, and $Y(1)$ for the treated group, and the main sample can be much larger than the calibration sample.

3.1. The naive method

Tha basic naive method ignores the measurement error and uses the error-prone covariate W, instead of X, in the propensity score model. Specifically, it regresses T on W, Z to estimate the propensity scores. Then, it uses IPTW to get an estimate of the treatment effect Δ.

3.2. Multiple imputation for external calibration

Multiple imputation for external calibration (MI-EC;[10]) is an imputation-based approach to handling measurement error. In particular, MI-EC generates multiple imputations of the true covariate, X, in the main sample, using information on the relationship between W and X in the calibration sample, as well as information on Y, T, Z, W from the main sample. It thus generates imputations in a way that is congenial (the imputations use all of the variables used in the analysis model) and reflects the relationships between the covariates, the treatment, and the outcome, yet still corrects for the measurement error[18,19].

MI-EC relies on several assumptions. It assumes that the joint conditional distribution of $(X, Z, T, Y \mid W)$ is multivariate normal, and that this distribution is the same in both the main and calibration samples, while the distribution of W can vary between them. Furthermore, it assumes that the mean of the joint conditional distribution is linear in W and the covariance matrix is constant. Formally, we state this as:

$$f(Y_i, T_i, Z_i, X_i \mid W_i) \sim N(\alpha W_i, \Sigma) \tag{2}$$

for all i in main and calibration samples.

A common assumption in measurement error models is the non-differential measurement error – also called the standard surrogacy assumption. It is one that assumes that the distribution of W is ignorable once we condition on the true covariate X and other helpful covariates, and it is formally stated as $f(Y \mid X, T, Z, W) = f(Y \mid X, T, Z)$. In contrast, MI-EC assumes a stronger version of this, in which the distribution of Y, T, Z does not depend on W once we condition on a value of X:

$$f(Y, T, Z \mid X, W) = f(Y, T, Z \mid X). \tag{3}$$

As pointed out by Liao et al.[20], this assumption can be regarded as the common non-differential measurement error assumption requiring also that the measurement error is independent of Z and T given X (having $f(W \mid X, T, Z) = f(W \mid X)$). It implies that we assume that the amount and structure of the measurement error do not vary across levels of treatment T, levels of the final outcome Y, or across levels of the rest of the covariates Z.

Applying MI-EC to propensity score methods poses possible violations to the required assumptions. Assumption (2) is immediately violated because the treatment variable is binary, but Guo et al.[10] show that this violation does not impact the correction of bias and non-coverage of confidence intervals (and we further investigate this in the simulations described below). Meanwhile, assumption (3) may be violated when a covariate is measured differently between the treated and the control group, or when the error grows with respect to a covariate only measured in the main sample. See section 6 for more discussion on this topic.

Given assumptions (2) and (3), we can construct the posterior distribution $f(X \mid Z, T, Y, W)$. These assumptions are used in identification of the joint distribution $f(X, Z, T, Y \mid W)$ using the two samples, as we never observe (T, Y, Z, W, X) for any individual; we either observe $f(T, Y, Z \mid W)$ or $f(X \mid W)$. They further allow us to relate all five variables using linear regression coefficients and covariances. We then use these to construct a posterior distribution using the SWEEP operator[10,21].

Guo et al.[10] use Reiter's[22] two-stage imputation procedure, in which they draw m sets of parameters from the posterior distribution, then for each set, produce n samples of X. The method makes $m \times n$ predictions of X, which can be used in standard methods of analysis and the results combined using combining rules in[22].

In this paper we pair the MI-EC method with propensity score weighting with the goal of estimating $\Delta = E[Y_i(1) - Y_i(0)]$, the marginal average causal effect. A benefit of the MI-EC approach is that once X is multiply imputed, any propensity score approach (e.g., IPTW, matching, or subclassification) could be used. Similarly, a "doubly robust" approach that uses the covariates in both the propensity score and outcome models can also easily be used. The steps for applying MI-EC with propensity score methods are the following:

1. obtain nested imputations of the true covariate X using MI-EC;
2. for each imputation, calculate the propensity scores;
3. for each imputation, apply any propensity score method (we use IPTW) and obtain an estimated treatment effect $\hat{\Delta}^{(m,n)}$; and
4. use Reiter's combining rules to get the final estimate and confidence interval.

4. Simulation Study

Guo et al.[10] present simulations that show the performance of MI-EC when estimating parameters in a linear model. We extend the simulations conducted by[10] to include a binary treatment variable and its effect Δ, a model for the treatment assignment, propensity score estimation, and a model for the outcome. We use the simulation study to compare the bias, root mean square error, and confidence interval coverage of four methods. Code to implement these methods, as well as to conduct the simulation, is available in an online appendix (http://ywebbvar.github.io/PS_MIEC/).

4.1. Methods compared

We compared the methods described in section 3:

1. The naive method

2. The "true" method, using the true covariate X
3. Uncongenial MI-EC, which used $f(X \mid W, Z)$ (uncongenial because it does not include variables that are subsequently included in the treatment effect estimation[18])
4. Congenial MI-EC, which used $f(X \mid W, Z, T, Y)$

For the MI-EC methods, we used m = 12 draws from the parameter distribution and n = 3 samples for each m, following Guo et al.[10].

4.2. Data Generation - Normally distributed simulation

We use two samples, the main sample and the external calibration sample. We are interested in the effect of a treatment T on an a univariate, continuous outcome Y. However, the treatment assignment, as well as the outcome, depend on univariate confounding variables X and Z. Meanwhile, X is correlated with Z, and is measured with error as W. In the main sample, the vector (Y, T, Z, W) is observed, whereas in the external calibration sample, the vector (X, W) is observed.

The treatment values are assigned according to the following logistic regression model,

$$\log\left(\frac{P\left[T=1 \mid X, Z\right]}{1-P\left[T=1 \mid X, Z\right]}\right)=\gamma_0+\gamma_X X+\gamma_z Z, \tag{4}$$

where γ_X and γ_Z are the average change in log odds of receiving the treatment for a unit increase in X or Z, respectively. We generate X and Z to follow a multivariate normal distribution as:

$$f(X, Z) \sim N_2\left\{\binom{0}{0}, \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right\}, \tag{5}$$

we use the following classical measurement error model:

$$f(W \mid T, X, Z) \sim N(X, \sigma^2), \tag{6}$$

and we define the distribution of the potential outcomes as:

$$f(Y(T) \mid T, X, Z) \sim N(\Delta T+\delta_X X+\delta_Z Z, \tau^2), \tag{7}$$

where the errors of the potential outcomes $Y(1), Y(0)$ are independent, and with equal variance.

We considered three levels of correlation ρ between the missing covariate X and the other confounder measured without error Z. We also used two levels of association between the confounder X and the treatment assignment, expressed as γ_X. Finally, we varied the variance of W to achieve four levels of reliability r. The specific values used in the simulation are presented in Table 1.

	Parameter	Value		Parameter	Value
	N_{sim}	1500	W	low r	0.3
	n_{calib}	500		moderate r	0.6
	n_{main}	2500		high r	0.9
				very high r	0.999
T	γ_0	0			
	γ_Z	0.4	Y	Δ	2
	small γ_X	0.4		δ_X	0.5
	large γ_X	1.2		δ_Z	0.1
				τ^2	1
(X, Z)	low ρ	0.3			
	medium ρ	0.6			
	high ρ	0.9			

Table 1. Values of parameters used in simulations. N_{sim} is the number of simulations, n_{calib} is the number of observations in the calibration sample, and n_{main} is the number in the main sample. The rest of the parameters appear in the propensity score model in (4), the model for X and Z in (5), the measurement error model in (6), and the outcome model in (7).

We ran 1500 simulations for every combination of level of correlation ρ, level of association γ_X, and reliability r. In each simulation we: generated 2500 observations for the main sample and 500 for the calibration sample; we applied the different methods to impute the missing covariate X; and for each method, we used Inverse Probability of Treatment Weighted (IPTW) estimation for the average treatment effect.

We examined the performance of each method for estimating the average treatment effect. In particular, for each method we generated the estimated difference in outcomes between treatment and control groups, calculated with inverse probability of treatment weights. We then calculated, for each of the methods, the bias, root mean squared error, and the percentage of simulations in which the 95% confidence interval covered the true effect.

4.3. Data Generation - Non-normally distributed simulation

To assess sensitivity of the methods to some of the assumptions of MI-EC we also simulated settings in which either $Y(t)$ and/or X had a skewed distribution. For X, we defined $X_1 \sim F_{(200,100)}$, $X_2 \sim \chi^2_1$, and $X_3 \sim N(0, 1)$. Then, we defined the variable X as a linear combination of X_1, X_2, X_3, truncated at 4. We used a method based on principal components[23] to generate a variable Z with a predefined correlation with X. For the potential outcomes $Y(t)$, we defined an error ϵ as the linear combination of $\epsilon_1 \sim F_{(300,20)}$ and $\epsilon_2 \sim \Gamma(2, 2)$, and the potential outcomes as $Y(t) = \Delta T + \delta_X X, \delta_Z Z + \epsilon$. We ran simulations with:

1. $Y(t)$ coming from a normal distribution and X coming from a skewed distribution,
2. X coming from a normal distribution and $Y(t)$ coming from a skewed distribution,
3. both X and $Y(t)$ coming from skewed distributions.

4.4. Results

Normally-distributed simulation Results from the normally-distributed simulation studies appear in Figure 2. In summary, as expected, a naive method that simply uses the covariate measured with error leads to bias in the treatment effect estimate across settings with less than perfect reliability. Another approach that uses only the joint distribution of the true covariate and the error-prone covariate to multiply impute the true covariate, termed 'uncongenial MI-EC method', also leads to bias. In contrast, the congenial MI-EC method, which incorporates both the observed outcome and

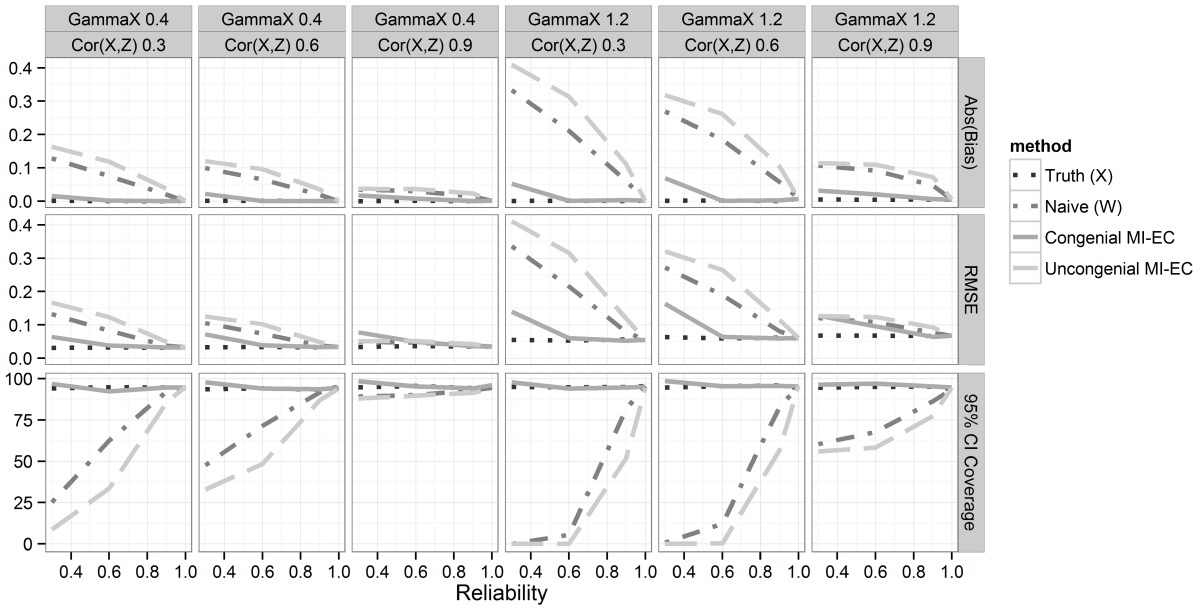

Fig. 2. Results from the normally-distributed simulation comparing the proposed congenial MI-EC method with the truth, the naive approach, and the uncongenial MI-EC. Increasing reliabilities of W are shown in the horizontal axes, while absolute bias (first row), root mean squared error (second row), and coverage of a 95% confidence interval (third row) for Δ are shown in the vertical axes. Columns define different levels of confounding – small confounding when the effect of X on treatment assignment (GammaX) is 0.4, and large confounding when it is 1.2; as well as different levels of correlation between the unobserved X and the observed Z (Cor(X, Z)).

the treatment assignment in the imputation, estimates the treatment effect almost as well as if the true covariate were available in the main data set.

As expected, there is substantial bias in the treatment effect estimate when using the covariate measured with error in the propensity score model (the naive method). This bias decreases as the reliability of the covariate increases. A reliability of 30% of the error-prone covariate can lead to bias in the treatment effect estimate that amounts to 0.3 standard deviation units (compared to $\Delta = 2$, it represents 15% of the treatment effect.) We found that the bias does not depend on the magnitude of Δ. Larger confounding by X (measured by its effect on the treatment assignment T) increases the bias in the estimate of Δ. Furthermore, a stronger correlation between X and Z reduces the bias. When the reliability is low and the association of X on T is large, the bias under low correlation between X and Z is about 0.3 units; meanwhile, if there is high correlation, the bias is about 0.1 standard deviation unit. Finally, the bias for the naive method increases with an increasing association of the error-prone covariate X on the treatment assignment T (γ_X), and it is scaled by the size of γ_X.

The uncongenial MI-EC approach that uses only the joint distribution of X and W to impute the true covariate X leads to bias greater than that of the naive method. Compared to the naive method, the bias ranges from a 7% increase under strong correlation between X and Z, to a 118% increase when there is low correlation between X and Z, and there is low reliability. In contrast, the congenial MI-EC method (which includes both the observed outcome and the treatment assignment in the imputation) estimates the treatment effect almost as well as if the true covariate X were available in the main data set. It decreases the bias by 85% in the worst-case scenario, when there is poor reliability, poor correlation between X and Z, and a large effect of X on the treatment assignment. In such a setting, the total bias of the congenial MI-EC method is 0.05.

The root mean squared error (RMSE) results are similar to those for bias. One exception is that the RMSE of the congenial MI-EC method is larger than that of the naive method under mild conditions of measurement error (under poor reliability, small confounding effect of X and large correlation of Z and X, located on the second row and third column.) This is due to the increase in variance from the imputation procedure. However, when there is a larger confounding effect of X, the RMSE of the the congenial MI-EC method is equal to or smaller than that of the naive method.

With regards to coverage, the MI-EC method yields 95% Wald confidence intervals with a coverage around 95% under all simulation scenarios. Under low reliability, the MI-EC method provides coverage of about 97%, leading to a small loss of power; meanwhile the naive method can have coverage of 25% under low reliability and small confounding, or 0% under low reliability and large confounding. Meanwhile, when there is high correlation between X and Z, the coverage of the naive method improves. It is about 80% when there is low confounding, and 60% under high confounding. Therefore, even when the bias in the treatment effect estimate is not very large for the naive method, the confidence intervals do not provide 95% coverage.

Non-normally-distributed simulation Results under violations of the joint normality assumption for (Y, X, Z) are similar, indicating that the MI-EC approach is not particularly sensitive to the assumption of multivarate normality. When only Y is misspecified (as binary, or as a highly skewed bimodal truncated continuous variable), while X, Z, W come from a joint normal distribution, the previous results hold. Whereas if X,Z, and W follow a highly skewed bimodal truncated continuous distribution, while Y follows either a binary, normal or skewed distribution, the magnitude of the bias increases for all methods, although the relative ranking of methods, and the general preference for the congenial MI-EC method, remains. Full details are provided in the e-appendix (available at http://ywebbvar.github.io/PS_MIEC/).

5. Illustrative Example

5.1. Overview and set-up

We now apply the MI-EC method to actual data, estimating the effect of living in a disadvantaged neighborhood on past-year substance use and mental health outcomes among adolescents using the National Comorbidity Survey Replication Adolescent Supplement (NCS-A). The NCS-A is a nationally representative survey of U.S. adolescent mental health (N=10,123), the methods and prevalence estimates of which have been described previously[24–26]. Participating adolescents gave informed assent and their parents or guardians gave informed consent. The Human Subjects Committees of Harvard Medical School and the University of Michigan approved recruitment and assent/consent procedures. Neighborhood disadvantage was defined using an established scale[27] that has been used previously in several epidemiological studies (e.g., Roux et al.[28]). Neighborhoods were classified as disadvantaged if they were in the lower tertile of the scale scores, as done in Rudolph et al.[29]. We consider two outcomes: 1) past-year substance (alcohol or drug) abuse or dependence, and 2) past-year anxiety or depressive disorder. These outcomes correspond to Diagnostic Statistical Manual IV (DSM-IV) diagnoses[30]. Because previous research suggests that the relationship between living in a disadvantaged neighborhood and mental health may differ by urbanicity[29], we restrict our analysis to the subset of NCS-A participants living in urban areas.

Maternal age at the birth of the adolescent is an important confounder of neighborhood-adolescent health associations, because it serves as a measure of family socioeconomic status. Ideally, maternal age at birth would be reported by the mother. For this example, we consider the mother's report of her age at the birth of the adolescent the true confounder, X. However, it is not always feasible to conduct interviews of both adolescents and their parents, so this confounder is

frequently reported by the adolescent. We consider the adolescent's report of maternal age at birth as W, a mismeasured version of X. For the purposes of this illustration, we restrict the urban subset of the NCS-A to those who have both X and W (n=1,926), as this allows us to use the true estimate (using X) as a reference.

The propensity score model includes gender, current age of the adolescent, race/ethnicity, region of the country, family income, family structure (i.e., the adolescent living her/his whole life with her/his mother and/or father), and maternal age at birth as main effects. We estimate the average effect of neighborhood disadvantage on prevalent substance abuse/dependence and prevalent anxiety and depression among those who currently live in disadvantaged neighborhoods (the average treatment effect on the treated, ATT), controlling for confounding through ATT weights (weighting by the odds). These disorders are considered prevalent if they were present in the 12 months prior to the diagnostic interview. For this illustrative example, we ignore the survey sampling design and weights and interpret the resulting effect estimates as the effect in the sample of adolescents in the NCS-A. However, the imputed values of X could be used in a survey design-based, weighted analysis to generate nationally representative estimates[31].

We divide the NCS-A sample into a calibration and study sample by randomly sampling 400 participants to use as the calibration sample and drop X from the remaining 1,526 to use as the study sample. As expected, adolescent-reported maternal age at birth is a noisier version of the mother-reported variable (the adolescent-reported variable has variance of 32.0 and the mother-reported variable has variance of 30.7 in the calibration sample). However, the two variables are highly correlated (0.94). Due to this high correlation, we add additional classical measurement error to adolescent-reported maternal age at birth to make two noisier versions of W. We compare the true (using mother-reported maternal age at birth, X), naive (using adolescent-reported maternal age at birth, W), congenial MI-EC (including the outcome, Y, exposure of living in a disadvantaged neighborhood, T, and vector of covariates, $\mathbf{Z}$) and uncongenial MI-EC (including $T, \mathbf{Z}$ but not Y) estimates of the ATT for three X, W correlation scenarios: 1) the true correlation, 0.94, 2) noisier measurement error with correlation 0.72, and 3) noisiest measurement error with correlation 0.30. As in the simulation, we use $m = 12$ draws from the posterior distribution and $n = 3$ imputations for each m.

5.2. NCS-A Results

Figure 3 shows the estimated ATT of living in a disadvantaged neighborhood on adolescent prevalent drug or alcohol abuse or dependence disorder and prevalent anxiety or depressive disorder. Specifically, the effect on the y-axis is the risk difference associated with living in a disadvantaged versus nondisadvantaged neighborhood for those who live in disadvantaged neighborhoods.

As seen in Figure 3, the effect estimates are slightly biased when the mismeasured version of maternal age at birth is used as if it were the truth. For example, naively using the version of adolescent-reported maternal age at birth with the most measurement error (correlation=0.30) would result in an estimate of the effect of living in a disadvantaged neighborhood on prevalent alcohol or drug abuse/dependence that is biased towards the null (-0.014 versus -0.016), and would result in an estimate of the effect of living in a disadvantaged neighborhood on prevalent anxiety or depression that is biased away from the null (0.082 versus 0.056). Although using the mismeasured covariate would not change inferences in the case of prevalent alcohol or drug use, in the case of prevalent anxiety or depression, using the version of adolescent-reported maternal age at birth with the most measurement error would result in a type-1 error—that is, we would conclude that the ATT is statistically significant when it is not.

We note that this example is used for illustrative purposes—substantive conclusions should not be drawn for several reasons. First, we are using a subset of urban NCS-A participants who have both maternal-reported and

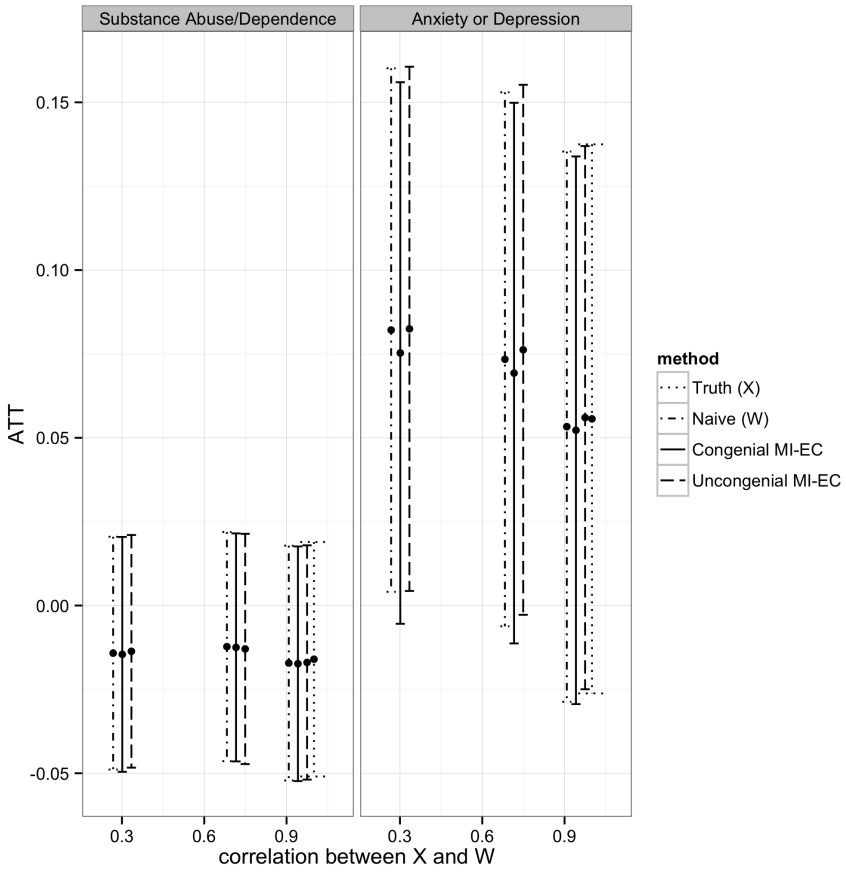

Fig. 3. Results from illustrative example. ATT estimates and 95% confidence intervals comparing the proposed MI-EC method with the truth, the naive approach, and the uncongeniel MI-EC for each of 3 X, W correlation scenarios. Estimates are for the average effect of living in a disadvantaged neighborhood on risk of having a current 1) substance abuse dependence disorder and 2) anxiety or depressive disorder for those who live in disadvantaged neighborhoods.

adolescent-reported maternal age at birth variables, which is not a meaningful population about which to draw inferences. Second, we are not incorporating the survey design and weights into this simple illustration. While we are using the sandwich estimator to calculate standard errors, it is not modified to incorporate the survey weights, sampling strata and clustering by neighborhood. Third, more complex models than what we fit here (e.g., including additional noncontinuous covariates) resulted in convergence problems with the current implementation of MI-EC. This is an area for future work. Fourth, positivity violations (cases where the probability of living in a disadvantaged or nondisadvantaged neighborhood is very small given some vector of covariate values) is frequently a concern when estimating neighborhood effects, but for this illustrative example, we have not addressed this issue.

6. Discussion and Conclusion

In the present work, we found that using a covariate measured with error in a propensity score method can lead to bias in the estimated treatment effect. However, a congenial MI-EC approach that includes the outcome, the treatment, and all confounders in the imputation model can be used to help correct for measurement error-induced bias.

The importance of congeniality has been discussed previously in the broader multiple imputation literature[18,19]. If a variable is used in the analysis procedure, it must be included in the imputation model for the approach to be congenial. If such a variable is absent, it implies that the absent variable is independent of the joint distribution defined in the imputation model.

On a similar note, Liao et al.[20] discussed how regression calibration, which can be regarded as a single imputation method, requires the inclusion of all confounders in the regression calibration model. Cefalu and Dominici[32] have shown an example of this result. In their case, they had a mismeasured exposure, a set of confounders, and an outcome. They found that unless the model used to predict the exposure included all confounders from the outcome model, the estimate of the exposure effect was biased. This speaks to having a congenial imputation model, but does not consider including the outcome or the treatment in the joint distribution from which to "impute". A limitation of the MI-EC approach is that the need to use the outcome in the imputation violates the separation of "design" from "analysis" that is important in the broader propensity score literature and should be done with caution[33]. Future work should investigate the costs of incorporating the outcome, and consider ways that concerns in utilizing the outcome could be addressed.

The MI-EC method by Guo et al. assumes a joint multivariate normal distribution. It is assumed because it makes it easier to construct a posterior distribution for the missing X when only two pieces available -the main and calibration samples - are available, using summary measures of $f(Y, Z \mid W)$ and $f(X \mid W)$. In their simulations, they showed that their method is robust to a binary Z or a mildly skewed X. On this note, Liao et al.[20] commented on the necessity of this assumption, since the multiple imputation method can be compared with a Monte Carlo integration of the distribution $f(Y, T, X, Z \mid W)$, over the observed data, which does not require assuming the joint multivariate normal distribution. Meanwhile, if an internal validation sample were available, one could use more flexible Bayesian methods for multiple imputation to construct $f(X \mid Y, T, Z, W)$.

The joint multivariate normal assumption is automatically violated in our case, because T is binary. Yet, the MI-EC method performs well, and it was also robust to having non-normal distributions for Y. However, we did observe convergence problems when we included many binary observed confounders, which may be due to a more extreme violation of the joint multivariate normal distribution assumption.

Finally, we have assumed W comes from a classical measurement error model, and that the error is non-differential with respect to treatment groups, and non-differential with respect to baseline covariates. The first assumption can partially be

relaxed, as the MI-EC method can handle a measurement error model that is linear in X. However, because the calibration sample only includes information on X and W, the measurement error model must be strongly non-differential[10]. By strongly non-differential, we mean that we assume that the measurement error distribution is the same for all levels of Y, Z and T (this last requirement represents a measurement error that is non-differential with respect to treatment groups, similarly to Sturmer and McCaffrey's methods.) Extensions of this work include: relaxing the assumption of a non-differential measurement model and comparing the performance of other propensity score methods using MI-EC.

In conclusion, we demonstrate that using a propensity score to control for confounding that is estimated as function of covariates measured with error leads to biased estimates of the treatment effect. However, when a calibration data set is available, MI-EC can be used to help correct for such bias.

7. Acknowledgments

This work was supported in part by the National Institute of Mental Health (R01MH099010; PI: Stuart.) KER's time was supported by the Drug Dependence Epidemiology Training program, (T32DA007292-21; PI: Deborah Furr-Holden). The National Comorbidity Survey Replication Adolescent Supplement (NCS-A) and the larger program of related National Comorbidity Surveys are supported by the National Institute of Mental Health [U01-MH60220 and ZIA MH002808-11] and the National Institute of Drug Abuse [R01 DA016558] at the NIH. The NCS-A was carried out in conjunction with the World Health Organization World Mental Health Survey Initiative.

The views and opinions expressed in this article are those of the authors and should not be construed to represent the views of any of the sponsoring organizations, agencies, or U.S. Government. The authors claim no conflicts of interest.

The authors wish to thank Kathleen Merikangas for support in providing the NCS-A data.

References

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

[2] Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb;25(1):1–21.

[3] Steiner PM, Cook TD, Shadish WR. On the Importance of Reliable Covariate Measurement in Selection Bias Adjustments Using Propensity Scores. J Educ Behav Stat. 2011 Feb;36(2):213–236.

[4] Millimet DL. The Elephant in the Corner: A Cautionary Tale about Measurement Error in Treatment Effects Models. In: Drukker DM, editor. Missing Data Methods Cross-sectional Methods Appl. Adv. Econom. Vol. 27A. vol. 27 of Advances in Econometrics. Bingley: Emerald Group Publishing; 2011. p. 1–39.

[5] McCaffrey DF, Lockwood JR, Setodji CM. Inverse probability weighting with error-prone covariates. Biometrika. 2013 Jun;100(3):671–680.

[6] Fuller WA. Measurement Error Models. John Wiley and Sons, Inc.; 1987.

[7] Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol. 1990 Oct;132(4):734–45.

[8] Cook J, Stefanski L. Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association. 1994;89(428):1314–1328.

[9] Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006 Aug;35(4):1074–81.

[10] Guo Y, Little RJ, McConnell DS. On using summary statistics from an external calibration sample to correct for covariate measurement error. Epidemiology. 2012 Jan;23(1):165–74.

[11] Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005 Aug;162(3):279–89.

[12] Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. J Am Stat Assoc. 1998;93(444):1321–1339.

[13] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol. 2008 Oct;37(5):1142–7.

[14] Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? Epidemiology. 2009 Jan;20(1):3–5.

[15] Rubin DB. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment. J Am Stat Assoc. 1980 Sep;75(371):591–593.

[16] Carroll RJ, Rupert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models. vol. 39; 2006.

[17] Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010 Feb;29(3):337–46.

[18] Meng Xl. Inferences with Uncongenial Sources of Input. Stat Sci. 1994;9(4):538–558.

[19] Schafer JL. Multiple Imputation in Multivariate Problems When the Imputation and Analysis Models Differ. Stat Neerl. 2003 Feb;57(1):19–35.

[20] Liao X, Spiegelman D, Carroll RJ. Regression calibration is valid when properly applied. Epidemiology. 2013 May;24(3):466–7.

[21] Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: J. Wiley & Sons; 1987.

[22] Reiter JP. Multiple imputation when records used for imputation are not used or disseminated for analysis. Biometrika. 2008 Nov;95(4):933–946.

[23] Goslee SC, Urban DL. The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software. 2007;22(7).

[24] Kessler RC, Avenevoli S, Costello EJ, Green JG, Gruber MJ, Heeringa S, et al. Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A). International Journal of Methods in Psychiatric Research. 2009;18(2):69–83.

[25] Kessler RC, Avenevoli S, Costello EJ, Green JG, Gruber MJ, Heeringa S, et al. National comorbidity survey replication adolescent supplement (NCS-A): II. Overview and design. Journal of the American Academy of Child & Adolescent Psychiatry. 2009;48(4):380–385.

[26] Merikangas KR, He JP, Burstein M, Swanson SA, Avenevoli S, Cui L, et al. Lifetime prevalence of mental disorders in US adolescents: results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). Journal of the American Academy of Child & Adolescent Psychiatry. 2010;49(10):980–989.

[27] Diez Roux AV, Kiefe CI, Jacobs Jr DR, Haan M, Jackson SA, Nieto FJ, et al. Area characteristics and individual-level socioeconomic position indicators in three population-based epidemiologic studies. Annals of Epidemiology. 2001;11(6):395–405.

[28] Roux AVD, Borrell LN, Haan M, Jackson SA, Schultz R. Neighbourhood environments and mortality in an elderly cohort: results from the cardiovascular health study. Journal of Epidemiology and Community Health. 2004;58(11):917–923.

[29] Rudolph KE, Stuart EA, Glass TA, Merikangas KR. Neighborhood disadvantage in context: the influence of urbanicity on the association between neighborhood disadvantage and adolescent emotional disorders. Social Psychiatry and Psychiatric Epidemiology. 2014;49(3):467—475.

[30] Kessler RC, Avenevoli S, Green J, Gruber MJ, Guyer M, He Y, et al. National Comorbidity Survey Replication Adolescent Supplement (NCS-A): III. Concordance of DSM-IV/CIDI Diagnoses With Clinical Reassessments. Journal of the American Academy of Child & Adolescent Psychiatry. 2009;48(4):386–399.

[31] Dugoff EH, Schuler M, Stuart EA. Generalizing observational study results: applying propensity score methods to complex surveys. Health Services Research. 2014 Feb;49(1):284–303.

[32] Cefalu M, Dominici F. Does exposure prediction bias health-effect estimation?: the relationship between confounding adjustment and exposure prediction. Epidemiology. 2014 Jul;25(4):583–90.

[33] Rubin DB. The design versus the analysis of observational studies for causal effects : Parallels with the design of randomized trials. Statistics in Medicine. 2007;26:20–36.