Research Article

Michael D. Regier*, Erica E. M. Moodie, and Robert W. Platt

The Effect of Error-in-Confounders on the Estimation of the Causal Parameter When Using Marginal Structural Models and Inverse Probability-of-Treatment Weights: A Simulation Study

Abstract: We performed an empirical study to evaluate the effect of mismeasured continuous confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weighting. By executing an extensive simulation using 500 randomly generated parameter value combinations within a defined space, we observed the well-understood effects of attenuation and augmentation, and two unanticipated effects: null effects and sign reversals. We implemented a secondary empirical study to further investigate the sign reversal effect. We use the results of our study to identify conceptual similarities between the analytic and empirical results for multivariable linear and logistic regression, and our empirical results. Through this synthesis, we have been able to suggest feasible directions of research as well as outline the form of expected results.

Keywords: causal inference, censored data, inverse probability weighting, marginal structural model, measurement error

*Corresponding author: Michael D. Regier, West Virginia University, Morgantown, WV, USA, E-mail: mregier@hsc.wvu.edu
Erica E. M. Moodie: E-mail: erica.moodie@mcgill.ca, Robert W. Platt: E-mail: robert.platt@mcgill.ca, McGill University, Montréal, QC, Canada

1 Introduction

Marginal structural models (MSMs) estimated via inverse probability-of-treatment weighting (IPTW) have become a standard tool for epidemiologists and researchers studying longitudinal processes [1, 2]. In the context of causal modeling, there has been interest and debate regarding the interpretation of effects of “exposures” that are not directly measurable, such as adiposity as measured via body-mass index [3–5]. The more general problem of measurement error in the context of directly but poorly measured quantities remains an area of interest.

Cole et al. [6] consider the problem of consistently estimating the net effect of HAART treatment on the first diagnosis of clinical AIDS or death from any cause when treatment is imperfectly measured. They integrate regression calibration, using validated treatment data, with inverse probability-of-treatment and censoring that is combined with a pooled logistic regression approach for marginal structural Cox models. The empirical results from Cole et al. [6] demonstrate attenuation of the estimated effect when treatment is mismeasured and the ability of regression calibration to produce an unbiased estimate.

Babanezhad et al. [7] investigate the problem of treatment measurement error, or more generally treatment mismeasurement. They study the impact of mismeasurement across several models which aim to quantify the same estimand. Focusing on the misclassification of a dichotomous treatment, they consider ordinary least squares, g-estimation, the naive IPTW, and doubly robust IPTW estimation. The asymptotic bias is discussed within the context of a linear model where it is shown, analytically and empirically, that the four estimators have the same bias when the probability of misclassification does not depend on covariates.

Recently, Ogburn and VanderWeele [8] explore the problem of nondifferential misclassification of a binary confounder. Following on work by Greenland [9], they show that if the effect of the confounder on the outcome is in the same direction for both the treated and the untreated, formalized as monotonicity [8, Result 1], then adjustment reduces the bias due to the confounder; the adjusted effect lies between the crude and the true measures. If the monotonicity assumption is violated, then the adjusted value may not fall between the crude and the true measures.

In this article, we study the problem of confounder mismeasurement and empirically investigate its impact on the estimation of the causal effect of treatment using MSMs with IPTW, focusing on point-treatment study. In Section 2, we review MSMs, the stabilized inverse probability-of-treatment weight, and the measurement error model. Section 3 describes the simulation study with the results presented in Section 4.

2 The models

2.1 Counterfactuals and the marginal structural model

Counterfactuals, or potential outcomes, are a useful tool for understanding MSMs and often, they are presented in the context of time-varying covariates and treatments [10–12]. With our focus on MSMs for point-treatment studies, we have a single interval setting where data are collected at the initiation of the study t_0 , and at a successive point in time, t_1 . We let L denote the vector of baseline covariates that were measured at t_0 before follow-up or final assessment at t_1 . We define the treatment decisions A to be the treatment assignment occurring between times t_0 and t_1 ; throughout, we shall take $A \in \{0, 1\}$. Finally, we define Y as the continuous outcome measured at t_1 .

We want a model describing the mean of Y for each treatment history, a , so we let Y_a denote the potential response that would be observed if the subject followed treatment a . We begin with the model

$$E(Y_a) = g(a : \zeta)$$

where ζ parameterizes the model and the function g is a link function. In general, the past treatments can be incorporated into the systematic component using some function of the treatment history, $f(a)$, thus $g(a : \zeta) = \zeta_0 + \zeta_1 f(a)$. The observed data for the i th subject are $O_i = \{A, L, Y_A\}$. We note that the point-treatment study is generalizable to k intervals [13].

Under the assumptions of consistency, exchangeability [14], positivity [15], and time ordering where treatment precedes the outcome [16], we can obtain an unbiased estimate of ζ . We accomplish this by fitting a model to the observed data where, under the conditions of consistency and exchangeability, the marginal expectation of the counterfactual is equal to the conditional expectation of the response given the treatment, $E(Y_a) = E(Y|A = a)$. Additionally, we assume that the n subjects are a random sample from some superpopulation of subjects, which is nearly infinite in size, about whom we wish to make inference [14].

2.2 The inverse probability-of-treatment weight

The assumption of no unmeasured confounders is an essential condition that permits causal inference from observational data [12]. Under this assumption, and in the absence of model misspecification, statistical exogeneity implies causal exogeneity [14]. A measure of statistical exogeneity for each subject can be used to weight each observation so that each subject represents that which would have been observed in the superpopulation with an ancillary treatment process [12].

The weight, called the stabilized weight, can be interpreted as the inverse of the incremental effect of the time-varying confounders on the current treatment beyond other treatment determinants including previous history and baseline covariates [1, 2]. When subject information is not observable, such as loss to

follow-up, a censoring process runs concurrent with both the treatment and the covariate processes. Following Hernán et al. [17], we define the censoring process at time t_0 , as C . It occurs between times t_0 and t_1 and precedes treatment allocation; $C \in \{0, 1\}$ with $C = 0$ defining data observation. The weight for point-treatment studies is specified as $w = w^A \times w^C$, where

$$w^A = \frac{\Pr(A = a|C = 0)}{\Pr(A = a|L, C = 0)}$$

$$w^C = \frac{\Pr(C = 0)}{\Pr(C = 0|L)} \quad (1)$$

Robins [12] shows that the weight corrects the regression estimator for the effect of confounding and censoring due to L . It is naturally expressed by considering the true joint density of the factual and counterfactual data, and it encodes information ensuring that the treatment history is causally exogenous, thus permitting the equivalence $E(Y_{a,c=0}) = E(Y|A = a, C = 0)$.

2.3 Classical unbiased nondifferential measurement error

We apply the classical, unbiased, additive, nondifferential measurement error model to confounders. For the i th subject and k th confounder, we have $L_{ik}^* = L_{ik} + \varepsilon_{ik}$ where L_{ik}^* is the mismeasured version of L_{ik} , $E(\varepsilon_{ik}) = 0$, and the errors are independent, $\varepsilon_{ik} \perp \varepsilon_{il}$ for $k, l \in \{1, \dots, p\}$, $k \neq l$ and $i, r \in \{1, \dots, n\}$, $i \neq r$. Under the assumption that $L^*|L, Y = L^*|L$ or equivalently L^* and Y are independent given L , the model is nondifferential [18, 19]. Furthermore, we assume that if L_k and L_l are mismeasured, then the associated errors are independent, $\varepsilon_{ik} \perp \varepsilon_{il}$.

For this situation, the relevant point-treatment directed acyclic graph (DAG) for classical measurement error is shown in Figure 1, where ε_L denotes exogenous measurement error noise. Here, L d-separates L^* from $\{A, Y\}$, thus by the probabilistic implications of d-separation [20, Theorem 1.2.4], L^* is independent of $\{A, Y\}$ conditional on L in every distribution compatible with the specified DAG (Figure 1). Furthermore, for all compatible distributions, this conditional independence implies surrogacy and nondifferential measurement error [18, 21].

When L^* , rather than L , is observed, then the inverse probability-of-treatment weight with a censoring process is

$$w = \frac{\Pr(A = a|C = 0)}{\Pr(A = a|L^*, C = 0)} \frac{\Pr(C = 0)}{\Pr(C = 0|L^*)}. \quad (2)$$

Using L^* rather than L , the causal estimator may be biased. We are solving $0 = \sum_{i=1}^n w_i^* m(A, Y; \zeta)$ and not $0 = \sum_{i=1}^n w_i m(A, Y; \zeta)$, where $m(A, Y; \zeta)$ is a function of the treatment history and response parameterized by ζ .

Figure 1 Nondifferential unbiased confounder measurement error for point-treatment directed acyclic graph, where ε_L denotes the measurement error noise such that $L^* = f(L, \varepsilon_L) = L + \varepsilon_L$

3 Simulation study

Attenuation and augmentation of parameter estimates are well-documented effects of measurement error. Attenuation occurs when the estimator is smaller in magnitude than the true value of the parameter. For simple linear regression, this is expressed as the reliability ratio [19, section 3.2.1] or equivalently by the attenuation factor [21, section 2.1]. Gustafson [21] shows that additional correlated covariates worsen this effect.

Augmentation, also known as *reversal*, occurs when the estimator is larger in magnitude than the true value of the parameter. In the setting of multiple linear regression, both attenuation and augmentation can occur, under additive measurement error, if there are multiple mismeasured covariates [21]. Recognizing these well-known effects of measurement error in linear regression, we are motivated to investigate whether the same set of effects will be observed for MSMs using inverse probability weights.

3.1 Primary study

For our investigation, we focus on the point-treatment case. Figure 2(a) and 2(b) are the DAGs corresponding to the data generating mechanism with and without censoring, respectively. The dashed line (—) in Figure 1 indicates that the censored data indicator was not included as a covariate in the data generating model, but affects what is observed.

Table 1 gives the six scenarios we considered. Scenarios S1–S3 had no censoring, using the data generating mechanism in Figure 2(a). Scenarios S4–S6 had censored data (Figure 2b). Scenarios S2 and S5 had only one mismeasured confounder, whereas scenarios S3 and S6 had both confounders mismeasured (Table 1).

Figure 2 DAGs for the simulation study data generation mechanisms, where $L^* = f(L, \varepsilon_L) = L + \varepsilon_L$ and ε_L denotes the measurement error noise: (a) without censoring and (b) with censoring. The dashed line indicates that the censored data indicator was not included as a covariate in the data generating model, but that it affects what is observed

Table 1 Simulation study scenarios

Scenario	Censoring	L_1 mismeasured	L_2 mismeasured
S1	No	No	No
S2	No	Yes	No
S3	No	Yes	Yes
S4	Yes	No	No
S5	Yes	Yes	No
S6	Yes	Yes	Yes

The vector of confounders, $L_i = (L_{i1}, L_{i2})$, $i = 1, \dots, n$, is assumed bivariate Normal, $L_i \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, \mu_2) = (0, 0)$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where $\sigma_1^2 = \sigma_2^2 = 1$ and $\rho = 0$, the correlation between L_1 and L_2 . We chose uncorrelated confounders to eliminate, or at least minimize, any induced augmentation [21, section 2.4]. We let A be binary, with $A = 1$ denoting the event of interest (e.g. new treatment) and $C = 1$ indicating missing data.

Table 2 identifies the observed data vectors for the six scenarios. Logistic regression was used to estimate the probability models given in Table 2. For example, we used $\text{logit}\{P(A = 1|L)\} = \alpha_0 + \alpha_1 L_1 + \alpha_2 L_2$ for scenario S1 and for scenario S4, we had the additional model $\text{logit}\{P(C = 0|L)\} = \gamma_0 + \gamma_1 L_1 + \gamma_2 L_2$ for the censoring mechanism. These two models have the basic structure except for the specification of mismeasured covariates (Table 2). Inverse probability-of-treatment weights were estimated by incorporating the probability models specified in Table 2 into eqs (1) and (2) for a point-treatment process, for example, scenario S4 used the stabilized weight of eq. (1). These weights were used to estimate the causal effect of A on Y .

Table 2 Observed data and associated probability models for the six simulation study scenarios

Scenario	Observed data	Probability model
S1	$O = (L, A, Y)$	$Pr(A = a L)$
S2	$O = (L_1^*, L_2, A, Y)$	$Pr(A = a L_1^*, L_2)$
S3	$O = (L^*, A, Y)$	$Pr(A = a L^*)$
S4	$O = (L, C, (1 - C)A, (1 - C)Y)$	$Pr(A = a L, C = 0)Pr(C = 0 L)$
S5	$O = (L_1^*, L_2, C, (1 - C)A, (1 - C)Y)$	$Pr(A = a L_1^*, L_2, C = 0)Pr(C = 0 L_1^*, L_2)$
S6	$O = (L^*, C, (1 - C)A, (1 - C)Y)$	$Pr(A = a L^*, C = 0)Pr(C = 0 L^*)$

The response is continuous and specified as $Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 A + \varepsilon_y$, $\varepsilon_y \sim N(0, \sigma_y^2)$ and $\sigma_y^2 = 1$. Given the distribution of the confounders, the marginal model parameters are

$$\begin{aligned} E(Y_A) &= E_{L|A}[E_{Y|L,A}(Y|L, A)] \\ &= \beta_0 + \beta_1 \mu_1 + \beta_2 \mu_2 + \beta_3 A \\ &= \zeta_0 + \zeta_1 A, \end{aligned}$$

where $E(L) = \mu = (\mu_1, \mu_2) = (0, 0)$, thus $\zeta_0 = \beta_0$ and $\zeta_1 = \beta_3$. We let L be mismeasured such that $L_1^* = L_1 + \varepsilon_1$, where $\varepsilon_1 \sim N(0, \tau_1^2)$, and $L_2^* = L_2 + \varepsilon_2$, where $\varepsilon_2 \sim N(0, \tau_2^2)$, such that $\tau_1 = \tau_2 = 0.5$ and $\varepsilon_1 \perp \varepsilon_2$.

We implemented an extensive simulation study, using 500 randomly generated parameter value combinations within a defined space, to investigate the effect of measurement error on the causal parameter. Parameters were selected from the space defined by $\{\gamma_0, \gamma_1, \gamma_2, \alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3\}$. We set $\gamma_0 = \alpha_0 = \beta_0 = 0$ and randomly selected the remaining parameters from the uniform distribution, $U[-3, 3]$, rounded to three decimal places. By design, the product space associated with the seven free parameters permitted independent sampling within each dimension. For each of the 500 locations in the parameter space, a simulation study was implemented using $S = 5,000$ simulated data sets with each data set having a sample size of $n = 10,000$.

We estimate the standard error of the bias using the empirical sampling distribution of biases, $\{\hat{B}_1, \dots, \hat{B}_S\}$ and the bootstrap. The estimated bias is $\hat{B}(\hat{\theta}) = \hat{E}(\hat{\theta}) - \theta_j$ where $\bar{\theta} = \hat{E}(\hat{\theta}) = \frac{1}{S} \sum_{s=1}^S \hat{\theta}_s$ and $\theta \in \{\zeta_0, \zeta_1\}$. The 95% simulation based confidence intervals for ζ_0 and ζ_1 , across the six scenarios and

four cases, were derived using the respective empirical distributions. The simulation study was conducted using R [22] with 4-GB RAM at 1,067 MHz DDR3 and a 2.8-GHz Intel Core 2 Duo processor.

3.2 Secondary study

We conducted additional simulations to investigate further the results of the primary study. We wanted to see if the findings of Gustafson [21] applied to the estimation of a MSM causal parameter. Specifically, we investigated the effects of confounder measurement error variability and confounder correlation. We chose the sign reversal scenario, as discussed in Section 4, where $(\gamma_0, \gamma_1, \gamma_2, \alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3) = (0, 1.037, 0.893, 0.186, -0.683, 0, 2.924, 2.661, -0.455)$. This parameterization indicates that $\zeta_0 = 0$ and $\zeta_1 = -0.45$. We varied the parameters τ_1^2 , τ_2^2 , and ρ , permitting τ_1 and τ_2 , the standard deviations associated with the measurement error models, to take values in the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and specified correlations between L_1 and L_2 , $\rho \in \{0, 0.5, 0.75, 0.9\}$.

4 Results

Tables 3 and 4 describe four representatives from the 500 locations in the parameter space, with mean values taken over the 5,000 simulations, that illustrate the four types of effects for four specific sets of $\{\zeta_0, \zeta_1\}$. When the covariates are accurately and precisely measured, the estimated regression coefficients are unbiasedly estimated (Tables 3 and 4, and scenarios S1 and S4, left-most panels of Figures 3–6). In these figures, the solid line (—) is the regression line $Y = \zeta_0 + \zeta_1 A$. The dashed line is the regression line $E(Y|A) = \zeta_0^* + \zeta_1^* A$, where ζ_0^* and ζ_1^* denote the intercept and slope obtained using the weights w^* . The dark gray lines denote simulation regressions that were attenuated, $|\hat{\zeta}_{1s}| < |\zeta_1|$, and the light gray lines denote augmented regressions, $|\hat{\zeta}_{1s}| > |\zeta_1|$, $s = 1, \dots, 5,000$.

Table 3 Estimated bias and associated standard error of the causal effect of treatment for the four representative measurement error effects (columns) and the six scenarios (rows), where ζ_j , $j = 0, 1$ is the true value of the parameter

Scenario	Attenuation	Augmentation	Null effect	Sign reversal
$\hat{B}(\zeta_0)$ (SE)				
ζ_0	0	0	0	0
S1	0.045(0.007)	0.000(0.001)	0.000(0.001)	0.001(0.002)
S2	0.390(0.004)	0.048(0.001)	0.006(0.001)	−0.443(0.001)
S3	0.879(0.001)	0.037(0.001)	0.004(0.001)	−0.289(0.001)
S4	0.099(0.009)	0.000(0.001)	0.000(0.001)	0.000(0.002)
S5	0.349(0.006)	0.050(0.001)	−0.004(0.001)	−0.448(0.001)
S6	0.908(0.002)	0.040(0.001)	−0.003(0.001)	−0.329(0.001)
Bias(ζ_1) (SE)				
ζ_1	2.56	−0.33	−2.31	−0.45
S1	−0.134(0.010)	0.000(0.001)	0.000(0.001)	0.001(0.003)
S2	−0.896(0.007)	−0.118(0.001)	−0.015(0.001)	1.007(0.001)
S3	−1.980(0.003)	−0.091(0.001)	−0.011(0.001)	0.678(0.001)
S4	−0.197(0.012)	0.000(0.001)	0.000(0.001)	0.002(0.003)
S5	−0.955(0.008)	−0.117(0.001)	−0.014(0.001)	0.992(0.001)
S6	−1.959(0.003)	−0.091(0.001)	−0.009(0.001)	0.663(0.001)

Table 4 Summary of simulation study results for the four measurement error effects (columns) and six scenarios (rows), where ζ_j , $j = 1, 2$ is the true value of the parameter, $\bar{\zeta}_j$, $j = 0, 1$ is the simulation based average for the i th regression coefficient. The simulation based interval, (p_l, p_u) , is defined by the 2.5 and 97.5 percentiles, respectively

	Attenuation	Augmentation	Null effect	Sign reversal
$\bar{\zeta}_0(p_l, p_u)$				
Scenario/ ζ_0	0	0	0	0
S1	0.05(−1.04, 0.52)	0.00(−0.03, 0.04)	0.00(−0.03, 0.03)	0.00(−0.17, 0.25)
S2	0.40(−0.29, 0.72)	0.05(0.01, 0.08)	0.00(−0.02, 0.03)	−0.43(−0.56, −0.30)
S3	0.88(0.64, 1.04)	0.04(0.00, 0.07)	0.01(−0.02, 0.03)	−0.29(−0.42, −0.15)
S4	0.10(−1.25, 0.66)	0.00(−0.03, 0.04)	0.00(−0.03, 0.03)	0.00(−0.20, 0.31)
S5	0.35(−0.62, 0.75)	0.05(0.02, 0.09)	0.00(−0.03, 0.03)	−0.45(−0.58, −0.29)
S6	0.91(0.55, 1.12)	0.04(0.01, 0.08)	0.00(−0.03, 0.03)	−0.33(−0.46, −0.17)
$\bar{\zeta}_1(p_l, p_u)$				
Scenario/ ζ_1	2.56	−0.33	−2.31	−0.45
S1	2.43(1.62, 4.27)	−0.33(−0.38, −0.29)	−2.31(−2.35, −2.27)	−0.44(−0.84, −0.17)
S2	1.66(1.12, 2.77)	−0.45(−0.50, −0.40)	−2.32(−2.36, −2.28)	0.56(0.37, 0.73)
S3	0.58(0.33, 0.98)	−0.43(−0.47, −0.38)	−2.32(−2.36, −2.28)	0.23(0.03, 0.40)
S4	2.36(1.43, 4.42)	−0.33(−0.38, −0.29)	−2.31(−2.37, −2.25)	−0.44(−0.88, −0.16)
S5	1.61(0.96, 2.95)	−0.45(−0.50, −0.40)	−2.32(−2.37, −2.27)	0.55(0.34, 0.72)
S6	0.60(0.30, 1.09)	−0.43(−0.47, −0.38)	−2.32(−2.37, −2.27)	0.22(0.00, 0.39)

Figure 3 Attenuation effect of confounder measurement error on the estimate of $\zeta = \{\zeta_0, \zeta_1\}$. The solid line is the true regression line for the effect of A on Y. The dashed line is the mean simulation based, estimated effect of A on Y using the weights w^* . The dark and light gray lines show the distribution of the 5,000 simulated regression lines; the dark gray lines denote attenuation, and the light gray lines denote augmentation

Figure 4 Augmentation effect of confounder measurement error on the estimate of $\zeta = \{\zeta_0, \zeta_1\}$. The solid line is the true regression line for the effect of A on Y . The dashed line is the mean simulation based, estimated effect of A on Y using the weights w^* . The dark and light gray lines show the distribution of the 5,000 simulated regression lines; the dark gray lines denote attenuation, and the light gray line denotes augmentation

Figure 5 Null effect of confounder measurement error on the estimate of $\zeta = \{\zeta_0, \zeta_1\}$. The solid line is the true regression line for the effect of A on Y . The dashed line is the mean simulation based, estimated effect of A on Y using the weights w^* . The dark and light gray lines show the distribution of the 5,000 simulated regression lines; the dark gray lines denote attenuation, and the light gray lines denote augmentation

Figure 6 Sign reversal effect of confounder measurement error on the estimate of $\zeta = \{\zeta_0, \zeta_1\}$. The solid line is the true regression line for the effect of A on Y . The dashed line is the mean simulation based, estimated effect of A on Y using the weights w^* . The dark and light gray lines show the distribution of the 5,000 simulated regression lines; the dark gray lines denote attenuation, and the light gray lines denote augmentation

Both attenuation and augmentation of the treatment effect were observed. Two unexpected effects were also noted. We call the first a *null effect*; the use of L_1^* or the set $\{L_1^*, L_2^*\}$ had no clear effect on the expected bias. The second, which we are calling *sign reversal* to distinguish it from Gustafson's definition of *reversal*, results in a different sign for the parameter of interest, $\text{sign}(\zeta) = -\text{sign}(\zeta^*)$, where ζ^* denotes the estimated coefficient using the mismeasured confounder(s).

4.1 Attenuation and augmentation

Attenuation and augmentation are known effects of measurement error [18, 21]. As expected with attenuation, there is an increase in the magnitude of the downward bias as confounder measurement error is introduced (scenarios S2 and S3 and S5 and S6 of Figure 3). As suggested by Gustafson [21], we observed greater attenuation as the number of mismeasured confounders increased (Table 3). Although bias is noted for scenarios S2 and S5, Table 4 suggests that the induced bias may not be meaningful with respect to the variability of the distribution of ζ_1 . The opposite is true for scenarios S3 and S6 (Table 4). We found that the standard error of the bias of ζ_1 was relatively small, suggesting a meaningful bias as well as the potential for extreme values to unduly influence the construction of simulation based confidence intervals when confounder measurement error occurs (Figure 3).

As with attenuation, the effect of augmentation was anticipated. The augmentation in Figure 4, scenarios S2 and S5, is slightly less pronounced than that of attenuation with much less variability of the empirical distribution of $\hat{\zeta}$ (Table 4). The augmentation effect is similar with and without censoring (Table 4). In scenarios S3 and S6 (Figure 4 and Table 4), the addition of an extra mismeasured confounder had little effect on the estimate of ζ and on the percentile interval. Table 4 shows that in all four scenarios, the intervals fail to capture the true value of ζ_1 . Tables 3 and 4 reveal that attenuation and augmentation of the causal parameter, ζ_1 , do not translate to the same measurement error effect on intercept, ζ_0 .

4.2 Null effect and sign reversal

Two unanticipated effects were observed: a null effect and sign reversal. A null effect occurs when the use of mismeasured confounders has little or no impact on the estimator (Figure 5). There are two distinguishing features: unbiased parameter estimation and small variation in the empirical distribution of $\hat{\zeta}_1$. The case shown in Figure 5 has little variation in the estimates across the 5,000 simulations (Table 3), which, from our set of simulations, does not appear to be atypical of the underlying distribution of $\hat{\zeta}$ for this effect. We observe no added attenuation or augmentation with more than one confounder measured with error.

The second is sign reversal (Figure 6). Although sign reversal (trend reversal) is discussed in the context of misclassification [3, 23, 24], it is not well documented for classical measurement error models and perhaps never observed due to the strong analytic and empirical results that support attenuation and augmentation [18, 19, 21, 25]. In scenarios S2 and S5 of Figure 6, we observe that $|\bar{\zeta}_1|$ is very close $|\zeta_1|$, but the sign is reversed, being positive rather than negative (Table 4). Scenarios S3 and S6 exhibit an increased effect of attenuation with the use of a second mismeasured confounder, suggesting that the anticipated effects of attenuation and augmentation are functioning in an intuitive manner despite the reversed sign. There is a tighter distribution of the empirical distribution of $\hat{\zeta}_1$ for scenarios S2, S3, S5, and S6, when compared to scenarios S1 and S3 (Table 3).

We used the 500 studies to estimate the occurrence of the four measurement error effects and defined the effects to be mutually exclusive. Null events are locations where the confidence interval covered the truth. The sign reversal events are defined as all those locations where the confidence interval did not capture the truth and there was a sign reversal. The augmentation and attenuation events are mutually exclusive by definition, but here they are defined conditional on being neither a null nor a sign reversal event.

We observe that augmentation events are the most frequent across all the scenarios and are followed closely by attenuation events (Table 5). When going from a model with only one mismeasured confounder to a model with two mismeasured confounders, there is a reduction in the proportion of null events and an increase in the other three event types. Attenuation events have the largest increase.

Table 5 Percent of observed measurement error effect over the 500 randomly selected locations in the parameter space for scenarios S2, S3, S5, and S6

Scenario	Null	Sign reversal	Augmentation	Attenuation
S2	16.0	7.8	42.8	33.4
S3	7.8	10.2	44.4	37.6
S5	21.2	7.2	40.4	31.2
S6	9.8	9.8	43.2	37.2

4.3 Secondary study

Using the sign reversal example, scenario S6, we investigate the observed effects of confounder measurement error on the bias of $\hat{\zeta}_1$ as a function of τ_1^2 , τ_2^2 , and ρ , $\rho = \{0, 0.5, 0.75, 0.9\}$ (Figure 7). To orient ourselves within the contour plots, scenario S6 corresponds to Figure 7 where $\rho = 0$ (top left), and $\tau_1 = \tau_2 = 0.5$; scenario S6 specified a location between the 0.6 and 0.8 contours. This location corresponds to region where we should expect to see sign reversal combined with attenuation. Furthermore, we make the general observation that as ρ increases, the distance widens between the contour lines.

We are able to demarcate four effect areas with three boundary lines. The first boundary occurs along the contour line where the bias is 0. To the left of this line is the area for which combinations of τ_1 and τ_2 , given the specified regression coefficients, will result in augmentation. To the right is the area that yields

Figure 7 Contour graph of the bias as a function of τ_1 and τ_2 when $\rho \in \{0, 0.5, 0.75, 0.9\}$ for the sign reversal example

attenuation. This area, in our example, is bounded on the right by the contour line where the bias is 0.45. To the right of this boundary line is the area where we would see attenuated sign reversal, that is the magnitude of the estimated regression parameter is less than that of the true parameter and the sign is opposite, $|\hat{\zeta}_1| < |\zeta_1|$ and $\text{sign}(\hat{\zeta}_1) = -\text{sign}(\zeta_1)$. This region is bounded on the right by the contour defined at 0.9. Here, we transition from attenuated sign reversal to augmented sign reversal, $|\hat{\zeta}_1| > |\zeta_1|$ and $\text{sign}(\hat{\zeta}_1) = -\text{sign}(\zeta_1)$, with equality holding on the boundary.

In our secondary study, changes in τ_1 appear to be more influential than changes in τ_2 , across all levels of ρ , for determining the type of effect seen. This may be related to the chosen parameterization of the censoring model. For the sign reversal example, we had $|\gamma_1| > |\gamma_2|$, thus changes in the imprecision of measurement for the more important variable have more of an impact on the type of effect observed. This is analogous to the effect of the measurement error for multiple mismeasured predictors using linear regression when the more important variable is measured with more error than others in the model [21, section 2.4].

The secondary study afforded the opportunity to explore the stabilized weight profile of each of the four types of measurement error effects within the sign reversal scenario (Figure 6). With this approach, we are able to fix (α, γ, β) and investigate how the stabilized weights change across the four types of measurement error. We obtained the 50th, 90th, 95th, and 99th percentile distribution of the weights using the 5,000 simulations; these were summarized using the median. The results for the 90th and 95th percentiles do not deviate much from those of the 50th percentile, so we only focus on the 50th and 99th.

The profile of the 50th percentile of the stabilized weights, $P_{50}(w^*)$, shows an ordering for the weight profile as a function of ρ (Figure 8). All four effects exhibit a positive trend as ρ increases. We observe an ordering of the minimum median stabilized weight: null, augmentation, attenuation, and sign reversal. This ordering is maintained for the maximum median stabilized weight. Despite these trends, the weights are similar in magnitude across all four effects.

The differences between the four measurement error effects become more pronounced when we consider the 99th percentile, the tail of the stabilized weight distributions (Figure 9). All four effects show that the stabilized weights trends are decreasing as ρ increases, which is opposite to the trend observed with the 50th percentile (Figure 8). We observe the reverse ordering of sign reversal to null effect with sign reversal having the smallest median $P_{99}(w^*)$ at each correlation. Furthermore, we observe a larger difference

Figure 8 Scatterplot of the 50th percentile of the stabilized weights, $M(P_{50}(w^*))$, for the four measurement error effects as a function of the confounder correlation, ρ , where M denotes the median and P_{50} denotes the 50th percentile

Figure 9 Scatterplot of the 99th percentile of the stabilized weights, $M(P_{99}(w^*))$, for the four measurement error effects as a function of the confounder correlation, ρ , where M denotes the median and P_{99} denotes the 99th percentile

in the magnitude of the median of the 99th percentile at each level of ρ . The unexpected result was that the sign reversal had the smallest median weights (Figure 9), which suggests that the profile of the tail of the weight distribution may be important in determining the observed measurement error effect. These empirical results suggest that the effect of confounder measurement error is a function of the model parameters, covariance structure of the confounders, and the noise of the measurement error models, $f(\alpha, \gamma, \zeta, \tau_1, \tau_2, \rho)$ which conceptually agrees with Gustafson [21, Result 2.3, eq. (2.8)] and with Buonaccorsi [18, eq. (6.38)].

5 Conclusions

We have observed counter-intuitive effects of confounder measurement error on the estimation of the causal parameter. The measurement error effects of attenuation and augmentation (*reversal*) are now joined by null effects and sign reversals. From our empirical study, we are able to identify some conceptual similarities between our results and those for multivariable linear and logistic regression. The secondary study suggests that effects are a function of the model parameters $\{\alpha, \gamma, \zeta, \tau_1, \tau_2, \rho\}$, thus we anticipate that future work may yield a relationship between the true and mismeasured parameters that is analogous to Gustafson [21, eq. (2.8) or (2.14)].

Given that IPTW uses functions of the propensity score to compute weights, it is natural to consider how models other than logistic regression would perform in this situation. Statistical learners (e.g. logistic regression, neural networks, regression trees, classification trees, and support vector machines) have been explored with respect to their ability to reduce bias, minimize standard errors, and balance covariates when model misspecification is of concern [26–29]. Empirical investigations suggest that boosting methods in conjunction with classification and regression trees (CART) [26] and random forests [26, 27] are promising approaches for balancing the bias-variance trade-off and handling covariate balance. More generally, boosting and bagging approaches are understood to provide better empirical results [26, 27, 29].

It is reasonable to translate these findings to IPTW models, such as MSMs, due to the similarities between IPTW and propensity scores, but the empirical propensity score investigations consider model misspecification of the systematic component and not measurement error. McCaffrey et al. [27] comment on this and clearly state that covariates may be mismeasured. It is suggested that propensity score estimation would not be affected if treatment allocation was a function of the observed, mismeasured covariate. On the other hand, if treatment allocation is dependent on the precisely measured, but unobserved covariates, then biases would result; this observation is relevant to our investigation. McCaffrey et al. [27] suggest that if there is good covariate balance for the mismeasured covariates, then propensity score methods may reduce the confounding effects in the error-free measures. McCaffrey et al. [30] propose an inverse probability weighting correction for error-prone covariates for propensity scores that builds on the work of Pearl [31], but does not consider the MSM context. As a counter point to these approaches, it has been shown that ideal performance of effect estimators is observed when treatment weights (propensity scores) exclude covariates that serve only to predict the treatment (exposure) but not the outcome [32–34], thus further investigations are warranted as such data relationships are typically not known. The propensity score literature provides direction for the use of statistical learners when the data are error-free, but there are unanswered questions about the ability of statistical learners to obtain unbiased estimates or bias-reduced estimates when confounders are mismeasured.

The simulation study is limited in that it estimates the prevalence of four effects, but does not analytically characterize the conditions leading to them as done by Gustafson [21], Carroll et al. [19], and Buonaccorsi [18]. Without analytic results, we are unable to identify the minimal set of influential parameters. Given the highly non-linear structure of this problem, a closed analytic solution may be elusive, as noted by Gustafson [21] in the context of logistic regression under additive measurement error. He provides sobering evidence, in the context of logistic regression, that we may be able to characterize the relationship between the parameters and the actual conditional distribution of the regression [21, section 2.6], but we should not expect to have all the nice features of previous measurement error results.

Our secondary study provided insights into the parameter settings that might lead to sign reversal effects, the most troubling of the effects of measurement error. Further work is necessary, however our observations suggest that translating prior understanding of how measurement error affects estimators to the effect of confounder measurement error is a plausible course of research for developing correction methodology. Another important direction will include the extension of this work to the longitudinal setting, where it is plausible that the impact of measurement error in time-varying covariates will compound the bias observed in the cross-sectional setting. With this perspective, future work will require a

combination of analytic characterization and numerical exploration in order to understand and develop the intuition necessary for both analytic and applied work with mismeasured confounders.

References

1. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
2. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000;11:561–70.
3. Hernán MÁ, Cole SR. Invited commentary: causal diagrams and measurement bias. *Am J Epidemiol* 2009;170:959–62.
4. Shahar E. The association of body mass index with health outcomes: causal, inconsistent or confounded? *Am J Epidemiol* 2009;170:957–8.
5. Shahar E. Shahar responds to causal diagrams and measurement bias. *Am J Epidemiol* 2009;170:963–4.
6. Cole SR, Jacobson LP, Tien PC, Kingsley L, Chmiel JS, Anastos K. Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *Am J Epidemiol* 2009;171:113–22.
7. Babanezhad M, Vansteelandt S, Goetghebuer E. Comparison of causal effect estimators under exposure misclassification. *J Stat Plann Inference* 2010;140:1306–19.
8. Ogburn EL, VanderWeele TJ. On the nondifferential misclassification of a binary confounder. *Epidemiology* 2012;23:433–439.
9. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980;112:564–9.
10. Bryan J, Yu Z, van der Laan MJ. Analysis of longitudinal marginal structural models. *Biostatistics* 2004;5:361–80.
11. Moodie EEM, Stephens DA. Marginal structural models: unbiased estimation for longitudinal studies. *Int J Public Health* 2011;56:117–19.
12. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry D, editors. *Statistical models in epidemiology: the environment and clinical trials, the IMA volumes in mathematics and its applications*, Vol. 116. New York: Springer-Verlag, 1999:95–134.
13. Robins JM. Marginal structural models. In 1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, 1998.
14. Robins JM. Association, causation, and marginal structural models. *Synthese* 1999;121:151–79.
15. Hernán MÁ. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;60:578–86.
16. Mortimer KM. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol* 2005;162:382–8.
17. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Stat Assoc* 2001;96:440–8.
18. Buonaccorsi JP. *Measurement error: models, methods, and applications*. Boca Raton, FL: Chapman & Hall and CRC, 2010.
19. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: a modern perspective*, 2nd ed. Boca Raton, FL: Chapman & Hall and CRC, 2006.
20. Pearl J. *Causality: models, reasoning, and inference*, 2nd ed. New York, NY: Cambridge University Press, 2009.
21. Gustafson P. *Measurement error and misclassification in statistics and epidemiology*. Boca Raton, FL: Chapman & Hall and CRC, 2004.
22. R Development Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. Available at: <http://www.R-project.org>, ISBN 3-900051-07-0.
23. Dosemeci M, Wacholder S, Lubin JH. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol* 1990;132:746–8.
24. Weinberg CR, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? *Am J Epidemiol* 1994;140:565–71.
25. Fuller WA. *Measurement error models*. Hoboken, NJ: John Wiley & Sons, 1987.
26. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2009;29:337–46.
27. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods* 2004;9:403–25.
28. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* 2008;17:546–55.
29. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol* 2010;63:826–33.
30. McCaffrey DF, Lockwood JR, Setodji CM. Inverse probability weighting with error-prone covariates. *Biometrika* 2013;100:671–80.

31. Pearl J. On measurement error in causal inference. In Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010), 2012.
32. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med* 2013;32:2837–49.
33. Lefebvre G, Delaney JA, Platt RW. Impact of mis-specification on the treatment model on estimates from a marginal structural model. *Stat Med* 2008;27:3629–42.
34. Moodie EEM. Risk factor adjustment in marginal structural model estimation of optimal treatment regimes. *Biom J* 2009;51:774–88.