11-18-2016

Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods

Kara E. Rudolph

School of Public Health, University of California, Berkeley, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, kara.rudophh@berkeley.edu

Elizabeth A. Stuart

Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Department of Health Policy, Johns Hopkins Bloomberg School of Public Health

Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods

[1], Kara E. Rudolph, PhD, MPH, MHS*[1,2], and Elizabeth A. Stuart, PhD[3,4,5]

[1]School of Public Health, University of California, Berkeley, California
[2]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
[3]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
[5]Department of Health Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

Abstract

Propensity score methods are a popular tool to control for confounding in observational data, but their bias-reduction properties are threatened by covariate measurement error. There are few easy-to-implement methods to correct for such bias. We describe and demonstrate how existing sensitivity analyses for unobserved confounding—propensity score calibration, Vanderweele and Arah's bias formulas, and Rosenbaum's sensitivity analysis—can be adapted to address this problem. In a simulation study, we examined the extent to which these sensitivity analyses can correct for several measurement error structures: classical, systematic differential, and heteroscedastic covariate measurement error. We then apply these approaches to address covariate measurement error in estimating the association between depression and weight gain in a cohort of adults in Baltimore City. We recommend the use of Vanderweele and Arah's bias formulas and propensity score calibration (assuming it is adapted appropriately for the measurement error structure), as both approaches perform well for a variety of propensity score estimators and measurement error structures.

Keywords: measurement error, propensity score, unobserved confounding

*Corresponding author:
13B University Hall, School of Public Health, Berkeley, CA 94720
kara.rudolph@berkeley.edu
tel. +15106431889
fax +15106437626

Propensity score methods are a popular tool in the analysis of observational data (1). However, the theory underlying and justifying their use assumes that the covariates included in the propensity score model are measured without error and in the same way across treatment groups. In reality, covariates are often measured with error and may be measured differently (e.g., by different instruments) or with differential measurement error across treatment groups. For example, disability measures from instruments such as the Instrumental Activities of Daily Living (IADL) scale and the physical component summary of the 36-Item Short Form Health Survey (SF-36-Phys) are mismeasured versions of the true, latent construct of disability. Such measures may not only be noisier versions of the truth, but may differ between treatment groups either because the intervention and control groups use two different instruments (e.g., IADL in the intervention group and SF-36 in the control group) or because the intervention leads to measurement differences in the SF-36. Ignoring these measurement issues may lead to incorrect effect estimates. In fact, using simulation studies based on real data, Steiner et al. (2011) showed that using one or more covariates measured with error (classical, nondifferential measurement error in which the mismeasured covariates were noisier versions of the true unobserved covariates) attenuated the bias-reducing properties of propensity score methods (including subclassification, weighting, and regression with the propensity score as covariate) (2).

Although covariate measurement error is likely in propensity score models—particularly in public health and the social sciences—there has been little research into methods that can correct for such error when using propensity score approaches (3) with a few exceptions (4–7). The number of strategies available could be increased by recognizing the link between covariate measurement error and unobserved confounding and adapting methods that assess sensitivity to an unobserved confounder to address covariate measurement error when using a propensity score approach.

In this paper we aim to: 1) explicate the link between covariate measurement error and unobserved confounding, 2) adapt existing, easy-to-implement sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods, 3) describe scenarios under which each approach may be appropriate, 4) evaluate their performance in a limited simulation study, and 5) apply these approaches to address covariate measurement error in estimating the association between depression and weight gain in a cohort of adults in Baltimore City.

Most discussions of measurement error in the literature have focused on classical measurement error that is nondifferential and homoscedastic. However, measurement error that is differential by treatment status may be especially pertinent in a propensity score context—e.g., when propensity scores are used to match control subjects from one study or population to treated subjects in an intervention study (8). Consequently, we will specify how the approaches considered herein can be applied to classical measurement error as well as to measurement error that is differential by treatment status in terms of both the location and scale parameters.

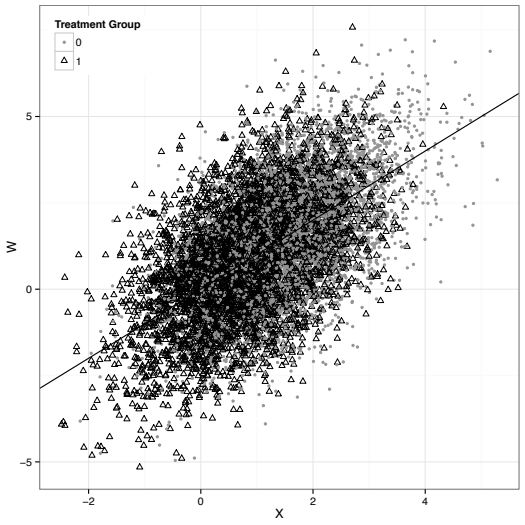
NOTATION, ESTIMANDS, AND ASSUMPTIONS

Let observed data $O = (Z, A, W, Y)$ and complete data $C = (Z, X, A, W, Y)$, where: Z

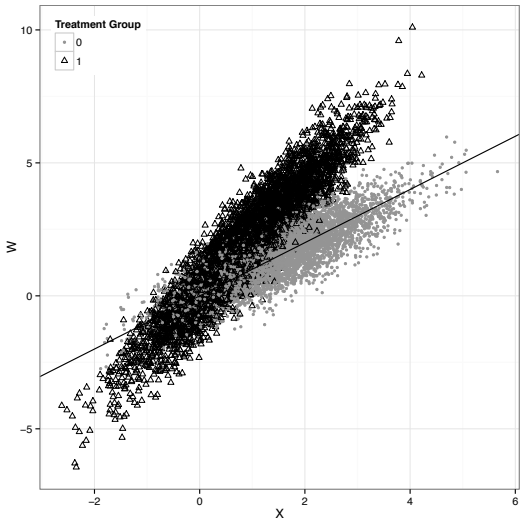
is an observed, continuous covariate, measured without error; X is an unobserved, continuous covariate, measured without error; A is an observed, binary (0/1) variable indicating treatment, measured without error; W is an observed, mismeasured version of X; and Y is an observed, continuous outcome, measured without error.

Measurement error scenarios.

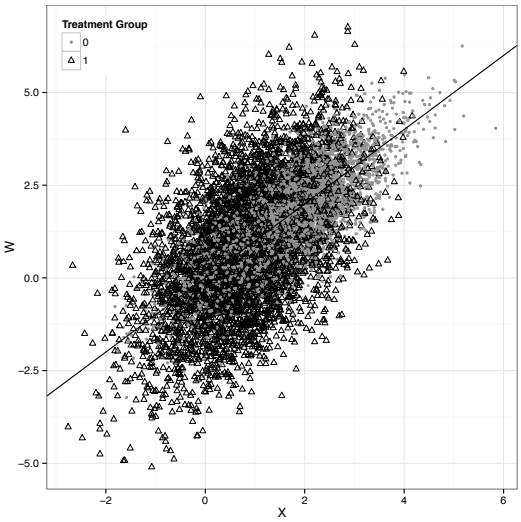
We consider additive, non-Berkson measurement error of the form: $W = X + U$, where $U \sim N(f(O), \sigma^2 f^2(O))$. We consider three measurement error scenarios. First, in classical measurement error, W is an unbiased, noisier version of X: $U \sim N(0, \sigma^2)$. An example of classical measurement error is blood pressure, where readings from an automatic blood pressure cuff are noisier versions of the true value (9). Second, we consider measurement error that is differential by treatment status in the location parameter: $U \sim N(f(A, X), \sigma^2)$, subsequently referred to as systematic differential measurement error. An example of this could be blood pressure measured by two different automatic cuffs, one for each of two treatment groups, that are not calibrated to each other. Third, we consider measurement error that is differential by treatment status in the scale parameter: $U \sim N(0, \sigma^2 f^2(A, X))$, subsequently referred to as heteroscedastic measurement error. An example of this could be blood pressure measured using a manual sphygmomanometer in the treatment group but using an automatic blood pressure cuff in the control group. In such a scenario, we may expect more variability in the control group versus the treatment group. Figure 1 depicts each of the three measurement error scenarios, and Figure 2 compares the naive and true propensity scores.



(a) Classical measurement error

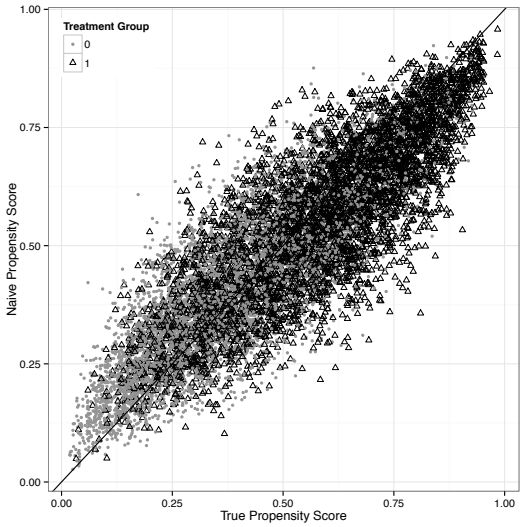


(b) Systematic differential measurement error

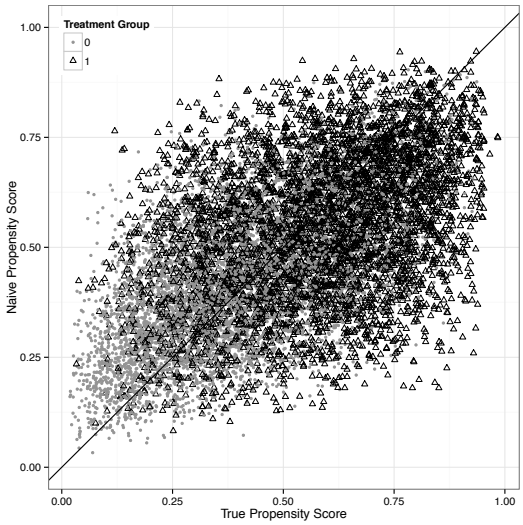


(c) Heteroscedastic measurement error

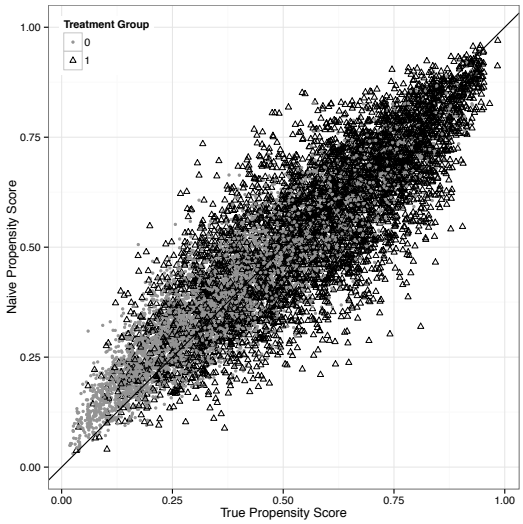
Figure 1: Measurement error scenarios. Scatter plots of the covariate measured without error, X, and the covariate measured with error, W, by treatment status.



(a) Classical measurement error



(b) Systematic differential measurement error



(c) Heteroscedastic measurement error

Figure 2: Measurement error scenarios. Scatter plots of the true propensity score (when X is used) and the naive propensity score (when W is used), by treatment status.

Estimands and assumptions

We consider two estimands of interest: the average treatment effect (ATE), $E(Y(1) - Y(0))$, and the conditional average treatment effect, $E(Y(1) - Y(0)|Z, X)$, where $Y(a)$ is the counterfactual outcome setting $A = a$ and the expectations are taken across all i individuals. If X is observed, identification of both estimands relies on the following assumptions. First, we assume strongly ignorable treatment assignment: for each $a \in \{0, 1\}$, we have $Y(a) \perp\!\!\!\perp A|X, Z$ under positivity: $0 < P(A = 1) < 1$. We also assume consistency: for each $a \in \{0, 1\}$, we have $Y(a) = Y$ on the event $A = a$. Finally, we assume the stable unit treatment value assumption (SUTVA): there is one version of each treatment condition and the treatment

assignment of individual i does not influence the potential outcome of another individual.

However, if we only observe a mismeasured version of X, W, then the estimands are not identifiable because the strongly ignorable treatment assignment assumption is not met. We describe how using W instead of X in an estimator fails to completely control for confounding in more detail in the following section.

COVARIATE MEASUREMENT ERROR AND UNOBSERVED CONFOUNDING

In curricula and the literature, measurement error and unobserved confounding (also called omitted variable bias)(10) are discussed as threats to valid causal inference, but typically as separate topics without consideration of their intersection in the case of covariate measurement error. Some notable exceptions include (2, 11, 12).

The equivalency between covariate measurement error and unobserved confounding can be seen through their impact on the assumption of ignorable treatment assignment. Let $\{X, Z\}$ be the set of confounding variables, consisting of the subset of observed confounding variables, $\{W, Z\}$, and unobserved confounding variables, Δ. When there is unobserved confounding,

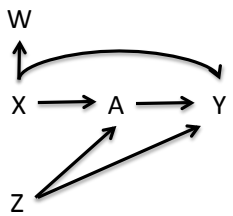
$$\begin{aligned} Y(a) \not\perp\!\!\!\perp A|W, Z \\ \perp\!\!\!\perp A|W, Z, \Delta (13). \end{aligned}$$

This can be recast in measurement error terms, using the notation defined previously:

$$\begin{aligned} Y(a) \not\perp\!\!\!\perp A|W, Z \\ \perp\!\!\!\perp A|W, Z, U \\ \perp\!\!\!\perp A|X, Z. \end{aligned}$$

The equivalency can also be seen through directed acyclic graphs (DAG), as described in Hernan and Robins (2016) (11). We can rewrite Hernan and Robins' Figure 9.8 as in Figure 3. This measurement error DAG is easily recast as an unobserved confounding DAG, because we see that X is a confounder and it is unobserved.

Figure 3: Directed acyclic graph representing measurement error and unobserved confounding.



SENSITIVITY ANALYSES

We now describe several easy-to-implement sensitivity analyses for unobserved confounding and describe how they can be adapted for measurement error. Sample code to implement each of these analyses is in the Web Appendix.

Propensity Score Calibration

Propensity score calibration (PSC) has been described previously as a method for reducing bias due to unobserved confounding (14). It uses a validation dataset that contains the treatment variable, A, and all covariates present in the propensity score model, including covariates measured without error, $\{X, Z\}$, and the subset of those variables that are also measured with error, W (see Table 1). PSC is similar to regression calibration (15) except that instead of modeling the mismeasured and correctly measured covariates, the naive (using mismeasured covariates) and true (using correctly measured covariates) propensity scores are modeled in the validation subset and then extrapolated to calibrate the naive propensity score in the main study dataset. The authors of the method state that the calibrated propensity score can be used in any propensity score approach, including matching, subclassification, and controlling for the propensity score in an outcome regression model (16). We found empirical support for this statement in our simulations. However, we found that using this method in an inverse probability of treatment-weighted estimator increased rather than decreased bias due to measurement error (results not shown but available upon request). Standard errors should be estimated by bootstrapping to propagate uncertainty from the calibration procedure.

Table 1: Variables needed in validation and main datasets for PSC.

	A	X	W	Z	Y
Validation dataset	✓	✓	✓	✓	
Main dataset	✓		✓	✓	✓

PSC makes the following assumptions: 1) that a validation dataset exists that contains A and all covariates, including versions measured with and without error, $\{X, W, Z\}$; 2) that the true propensity score is a function of the mismeasured propensity score, treatment, and any other covariates in the validation dataset; 3) that the PSC model in the validation dataset generalizes to the main dataset; and 4) that the naive propensity score, e_{naive}, is a surrogate for the true propensity score, e_{true}.The surrogacy assumption is violated if the naive propensity score contains additional information about the outcome that is not contained in the true propensity score. Perhaps contrary to initial intuition, this is a restrictive assumption (17). For measurement error scenarios, it is violated when the covariate measurement error is differential by treatment status or differential by a confounding variable. Consequently, this approach is theoretically only appropriate for classical measurement error (see Table 2). However in simulations (detailed in the Web Appendix and results shown in Table 3), we find that it has similar performance in the case of heteroscedastic measurement error. To relax the sensitivity of PSC to the surrogacy assumption in the case of systematic differential measurement error, we modified the algorithm to use weighted least squares (WLS) in modeling the relationship between the true propensity score as a function of the naive propensity score, A, and Z, allowing the variance to differ by strata of A (sample code

provided in the Web Appendix).

Vanderweele and Arah's sensitivity analysis

The Vanderweele and Arah sensitivity analysis for unobserved confounding has been described previously (18). Briefly, they provide formulas to calculate the bias caused by unobserved confounding in estimating conditional or marginal effects. These bias formulas involve setting values for various sensitivity parameters—for example, parameters relating the unobserved confounding variable to treatment and relating the unobserved confounding variable to the outcome—the number of which depends on the simplifying assumptions made. We may have little information from which to guess reasonable values for each parameter, but could explore a matrix of reasonable combinations, identifying those which result in a change in inference. The bias formulas are estimand-specific but can be used for any method of estimation. For example, the same bias formula for the ATE can be used regardless of whether propensity score matching or inverse probability of treatment weighting is used. The standard error of the bias-corrected ATE is the same as the standard error of the biased ATE if the parameters in the bias formula do not vary by strata of covariates. Otherwise, standard errors can be estimated by bootstrapping.

This approach can be used to correct for classical measurement error and systematic differential measurement error (see Table 2). Although it is not clear how to adapt the bias formulas for heteroscedastic measurement error, we find in simulations that bias formulas that ignore the differential measurement error in the scale parameter perform well (see Table 3). For classical measurement error, Vanderweele and Arah's simplified bias formula can be used, setting sensitivity parameters that describe the association between treatment, A, and the portion of X not captured by W, U, and between Y and U (18, p. 44). For systematic differential measurement error, the bias equation from Vanderweele and Arah's Theorem 1 can be used, setting the following sensitivity parameters: 1) the association between Y and U conditional on W and Z when $A = 1$ and 2) when $A = 0$, 3) the difference in the mean of U conditional on $A = 1$ and W and Z and the mean of U conditional on W and Z only, and 4) the difference in means detailed in 3) when $A = 0$. See the Web Appendix for sample code.

Rosenbaum's sensitivity analysis

Rosenbaum's approach has been described previously (19–21). It assumes that: 1) the data are in propensity score-matched pairs and 2) the treatment and control groups in the matched dataset are balanced on observed confounding variables . There are several versions of this sensitivity analysis: the original version, which assumes—adapted for measurement error—that the portion of the true, unobserved covariate, X, that is not captured in the observed, mismeasured covariate, W, U is a near perfect predictor of the outcome, Y; the dual version, which assumes that U is a near perfect predictor of treatment, A; and the simultaneous version, which sets sensitivity parameters for the association between U and A and for the association between U and Y, similar to Vanderweele and Arah's bias formulas. Because of the restrictive assumptions of the original and dual versions, we consider the simultaneous version here.

The two sensitivity parameters in the simultaneous sensitivity analysis are Γ and Δ. For simplicity, we consider a binary Y. The two sensitivity parameters are given by the following

equations:

$$\begin{aligned} log\frac{\pi_A}{1-\pi_A} &= \beta_0 + \beta_W W + log(\Gamma)U + \beta_Z Z \\ log\frac{\pi_Y}{1-\pi_Y} &= \alpha_0 + \alpha_A A + \alpha_W W + log(\Delta)U + \alpha_Z Z. \end{aligned}$$

Γ is the multiplier by which the portion of X that is not captured by W, U increases or decreases the odds of treatment assignment. If measurement error does not affect odds of treatment assignment, then $\Gamma = 1$. Similarly, Δ, is the multiplier by which U increases or decreases the odds of the outcome, Y.

We perform the sensitivity analysis by varying Γ and Δ simultaneously. If Y is a binary outcome variable, the two sensitivity parameters are used to set the upper or lower bound of McNemar's test statistic (20). The resulting p-value from this test is the p-value corrected for measurement error. If Y is a continuous outcome, then the two sensitivity parameters are used to set the upper or lower bound of the normalized Wilcoxon Signed Rank test statistic. When Y is continuous, Δ is interpreted as the conditional odds the subject with greater U also has greater Y for the pair with median rank. This clunky interpretation makes it difficult to choose reasonable values for this sensitivity analysis parameter.

Rosenbaum's approach can be used to correct for classical measurement error and systematic differential measurement error (see Table 2). As with the other two approaches we have considered, although it is not clear how to adapt the approach to address heteroscedastic measurement error, we find in simulations that the method performs similarly for the heteroscedastic measurement error scenario as it does for classical measurement error (see Table 3).

Table 2: Summary of sensitivity analyses for unobserved confounding applied to propensity score approaches with covariate measurement error: applicable estimands and measurement error structures.

Sensitivity Analysis	Estimand	Measurement Error Structure		
		Classical	Differential	
			Systematic	Heteroscedastic
PSC, modified approach	marginal ATE, conditional ATE	X	(X)	(X)
Rosenbaum's sensitivity analysis	test statistic, p-value	X		(X)
Vanderweele and Arah's bias formula	marginal ATE, conditional ATE	X	X	(X)

SIMULATION RESULTS

Table 3 presents the simulation results. Details of the simulation set-up are provided in the Web Appendix. For each method and for each measurement error scenario, we present

the performance of the naive approach that uses W instead of X, the no measurement error approach that uses X, and the approach that corrects for measurement error using one of the three sensitivity analyses.

For all three measurement error scenarios, the Vanderweele and Arah bias formula approach has the greatest potential for reducing bias due to measurement error, since it can reduce bias by 100% if the correct sensitivity parameters are used. In addition, the Vanderweele and Arah bias-corrected estimates have variances and MSEs that are similar to those of the true estimates. Ninety-five percent confidence interval coverage is also high.

We see that the PSC approach reduces but does not eliminate bias due to measurement error under all three scenarios. Performance of the unmodified, least squares PSC performs better for the classical and heteroscedastic measurement error scenarios than for the systematic differential measurement error scenario. For both the classical and heteroscedastic measurement error scenarios, using PSC to obtain a corrected estimate reduces bias by nearly 80% as compared to the naive approach and results in 95% CI coverage of more than 70%.

The surrogacy assumption is violated under both differential measurement error scenarios. However, performance is not affected in the heteroscedastic case, which corroborates previous results that demonstrated little practical impact of heteroscedastic error (22, page 81). As seen in Table 3, using the WLS modification improved performance of PSC in the systematic differential measurement error scenario. Using the standard PSC algorithm, the corrected estimates remained 34% biased, on average, while the WLS implementation resulted in 5% bias—similar to the PSC results of the other two measurement error scenarios. In addition, the WLS modification increased 95% CI coverage from 14.2% to 96%. However, these gains were at the expense of increased variance.

The Rosenbaum sensitivity analysis performed least well in terms of its ability to correct for covariate measurement error. This was true whether we performed matching with or without replacement. In addition, the corrected p-values were highly variable and spanned the range from zero to one over the 1,000 simulation iterations.

Table 3: Simulation results.

Sensitivity	Naive				True				Corrected			
Analysis	%Bias	Var	Cov	MSE	%Bias	Var	Cov	MSE	%Bias	Var	Cov	MSE
Classical Measurement Error												
Rosenbaum	-99.7	0.001		0.290	0.0	0.090		0.090	-54.5	0.060		0.146
Vanderweele and Arah	-43.5	0.004	0.0	1.710	0.0	0.004	92.8	0.004	-0.3	0.001	92.8	0.002
PSC, least squares	-43.2	0.003	0.0	1.684	0.0	0.001	94.5	0.001	-9.6	0.053	74.3	0.139
Systematic Differential Measurement Error												
Rosenbaum	-100.0	0.000		0.292	0.0	0.085		0.085	95.8	0.004		0.272
Vanderweele and Arah	-84.7	0.002	0.0	6.454	0.0	0.004	92.8	0.004	-1.1	0.001	83.3	0.002
PSC, least squares	-83.6	0.001	0.0	6.291	0.0	0.001	94.5	0.001	-33.9	0.077	14.2	1.154
PSC, WLS									-4.7	0.716	96.3	0.446
Heteroscedastic Differential Measurement Error												
Rosenbaum	-99.8	0.001		0.290	0.0	0.085		0.085	-64.3	0.046		0.169
Vanderweele and Arah	-36.2	0.004	0.0	1.181	0.0	0.004	92.8	0.004	-0.2	0.002	93.2	0.002
PSC, least squares	-35.8	0.003	0.0	1.159	0.0	0.001	94.5	0.001	-7.6	0.040	77.2	0.094
PSC, WLS									-10.8	0.126	80.9	0.255

APPLICATION

Overview and set-up

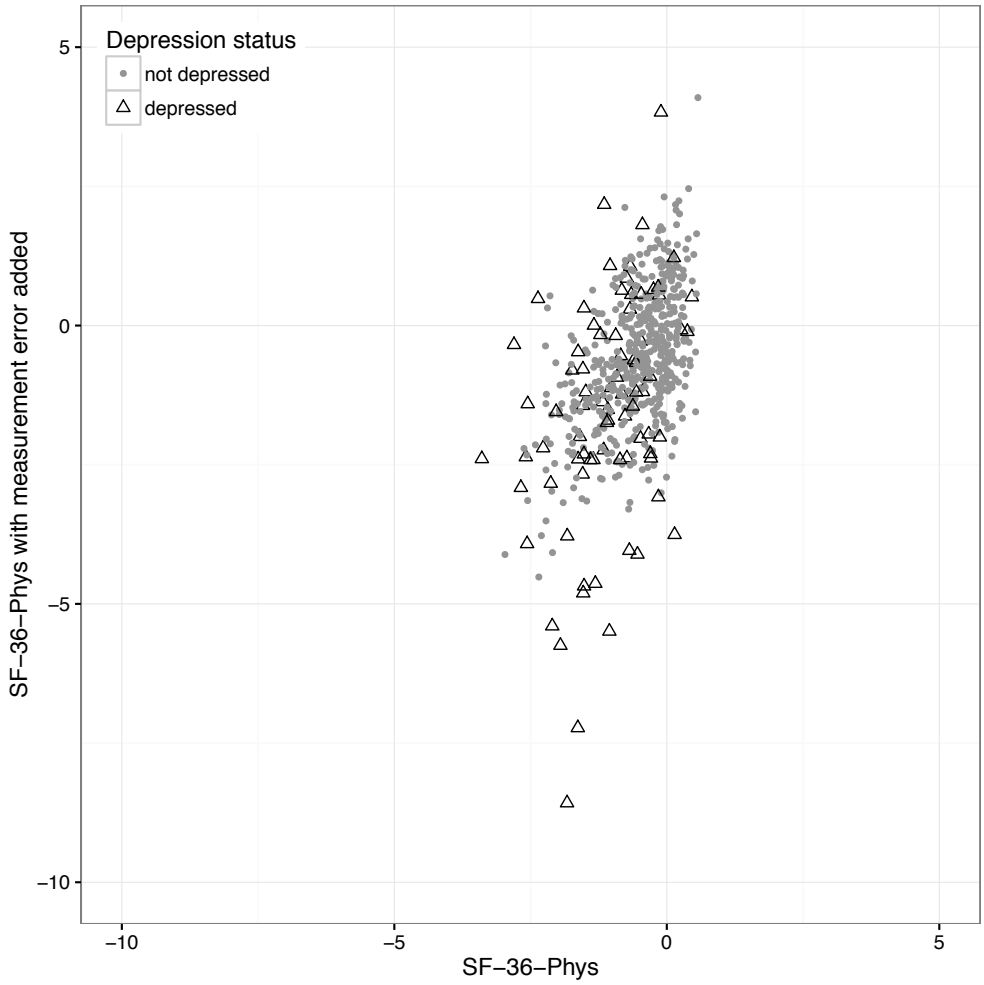
We now apply the above approaches to correct for covariate measurement error in estimating the association between baseline depression and subsequent change in body mass index (BMI) among middle- to older-aged women enrolled in the Baltimore Memory Study (BMS). The Baltimore Memory Study has been described previously (23). Participants gave informed consent and the Johns Hopkins University Institutional Review Board approved the study.

Depressive symptoms were assessed using the Center for Epidemiologic Studies Depression Scale (CES-D) (24). Participants were classified as depressed at baseline if their CES-D score was greater than or equal to 16 (24). Change in BMI was defined as the difference in BMI between visits 3 and 1. Potential confounding variables included baseline age, race (black versus white), marital status (married versus not), household wealth (23), level of education (23), retirement status, and smoking status. For the purposes of this simple example, we ignore measurement error in the exposure and each of the covariates listed above.

Disability status may be another confounding variable, as it has been shown to be associated with depression (25) and may be also associated with change in BMI. It is plausible that disability is measured with error, which may differ by depression status—e.g., those

with depression could have disability scores that are measured too low and with more noise than those without depression. For example, those with depression have more variable SF-36-Phys scores (variance of 0.625 versus 0.434) and scores that are lower on average (mean of -1.1 versus -0.6). However, we have no gold standard measurement of disability. For the purposes of illustration, we use the SF-36-Phys as the gold standard and simulate measurement error to add to obtain a mismeasured version: $W = X + N(0,1) + I(A = 1) * N(-0.5, 2)$, where A is the depression indicator, X is the SF-36-Phys, and W is the mismeasured version of the SF-36-Phys. W exhibits both systematic and heteroscedastic measurement error by exposure status, shown in Figure 4, in that those who are depressed score slightly lower and have more variability. The reliability of the mismeasured version among those who are depressed is 0.12; the reliability among those who are not depressed is 0.31.

Figure 4: Disability measure (SF-36-Phys) with and without added measurement error.



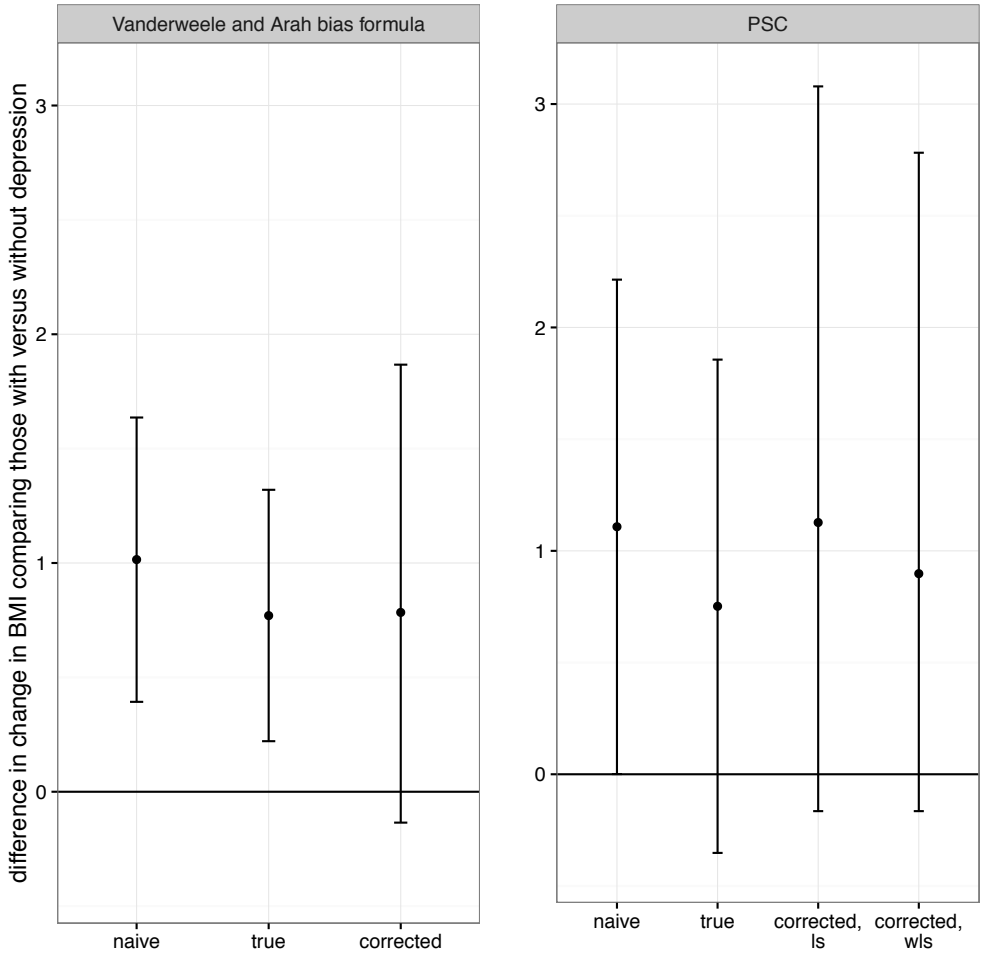
For this simple example, we include the N=597 women with complete data.We take a random one-third sample to create a validation subset (N=193). We then use this validation dataset to create the PSC model and to inform the sensitivity analysis parameters for the Vanderweele and Arah approach. We use stepwise selection to choose the best fitting

propensity score model from all possible second order terms. We control for confounding by conditioning on the linear propensity score for PSC and by inverse probability of treatment weighting for the Vanderweele and Arah approach. We estimate the average effect of depression on subsequent change in BMI, conditional on covariates, correcting for measurement error. The percentile method from 1,000 bootstrapped samples is used to estimate 95% CIs.

Results

We expect the WLS PSC approach and Vanderweele and Arah's bias formula to correct for our simulated systematic and heteroscedastic differential measurement error, as seen in Table 2. We include results using the original least squares PSC approach for comparison. Figure 5 shows the estimated effects comparing: 1) the "naive" estimate, using the version of SF-36-Phys with added measurement error; 2) the "true" estimate, using the SF-36-Phys without added error; 3) and the "corrected" estimates for each of the sensitivity analysis approaches.

Figure 5: Estimates and 95% CIs of the average effect of depression on subsequent change in BMI, conditional on covariates using the Vanderweele and Arah bias formula and PSC to correct for covariate measurement error.



We find that using the Vanderweele and Arah bias formula reduces bias by 94%, WLS PSC reduces bias by 59%, and the original least squares PSC slightly increases the bias. Although the bias is reduced using each of the two appropriate sensitivity analysis approaches, the confidence intervals widen.

These results should be interpreted with caution as we have made multiple simplifications. A more comprehensive analysis would account for missing data and informative drop out, time-varying confounding, and potential mediation.

DISCUSSION

Covariate measurement error and unobserved confounding are equivalent in terms of their potential to bias estimates. Few researchers undertake sensitivity analyses to estimate the potential impact of unobserved confounding and even fewer do so for measurement error. Moreover, when measurement error is considered, it is typically limited to classical measurement error even though more complex error structures may be present.

In this paper, we described and demonstrated how several easy-to-implement sensitivity analyses for unobserved confounding can be adapted to address classical, systematic differential, and heteroscedastic covariate measurement error in propensity score methods. In a limited simulation study, we provided optimal performance bounds for the extent to which these sensitivity analyses can correct for measurement error. To further lower barriers to implementation, we provide annotated R code in the Web Appendix that serves as a tutorial. We describe strengths and limitations of each sensitivity analysis below.

An advantage of PSC is that it can address multiple covariates measured with error simultaneously. It can be used with propensity score matching, subclassification, and regression adjustment of the propensity score, but not with weighting. However, the surrogacy assumption may be restrictive. Our adaptation of the method using WLS relaxes this assumption, allowing PSC to be used for systematic differential measurement error. This adaptation has the benefit of reducing bias and improving confidence interval coverage but at the expense of greater variance. Other limitations of PSC include that it tends to overadjust and break down when measurement error is large and/or the association between the naive and true propensity scores is weak (14). Most importantly, it reduces but may not eliminate bias due to measurement error. For example, in simulation studies in which all assumptions were met, Sturmer et al. found bias reductions between 32 and 106% (16). We found similar results (see Table 3). Similar to Sturmer et al., we found that bias reductions were closest to 100% in scenarios where the ATE=0 (results not shown but available upon request).

A significant advantage of the Vanderweele and Arah bias formulas is that if the correct sensitivity parameters are used and the assumptions are met, then the bias estimate is itself unbiased in expectation. Thus, this approach can fully correct for bias due to all three types of covariate measurement error considered. In addition, the approach can be used for any estimation method.

Rosenbaum's sensitivity analysis is perhaps the most familiar of all sensitivity analyses for unobserved confounding. However, it has several disadvantages in our context. The first is that it can only be used with propensity score matched data. The second is that it can be used to obtain a corrected test statistic or p-value, but it is not clear how it would provide an adjusted estimate for the ATE. Another disadvantage is that the interpretation of Δ is not

straightforward when Y is continuous, which makes it difficult to posit or provide sensitivity analysis values. Finally, we find that the method may reduce bias but that this reduction is far from complete (see Table 3).

In addition to demonstrating method performance with a simulation study, we applied these approaches to a data example estimating the association between depression and subsequent change in BMI among women, using propensity score methods to control for confounding. For the purposes of illustration, we added simulated differential measurement error to the disability covariate. We found that both recommended approaches, the Vanderweele and Arah bias formula and WLS PSC, reduce bias due to the covariate measurement error, thus demonstrating the practical utility of such sensitivity analyses.

This paper was limited in scope. Our goals were 1) to show the connection between bias caused by unobserved confounding and that caused by covariate measurement error and 2) to demonstrate how several simple approaches for addressing unobserved confounding could also be used to address covariate measurement error. There exist numerous other approaches for addressing covariate measurement error that may perform better and rely on fewer assumptions (4–6, 12, 26–34). However, with few exceptions (12), many of these approaches are not as easy to implement for nonstatisticians. Comparing performance among these other approaches and lowering barriers to implementation are areas for future work.

In conclusion, we recommend the use of Vanderweele and Arah's bias formulas and PSC (assuming it is adapted appropriately for the measurement error structure) to assess sensitivity of results to covariate measurement error. Both approaches are appropriate for a variety of propensity score estimators and measurement error structures. Real-world data are messy. Concerns about bias due to unobserved confounding and/or measurement error should be addressed rather than ignored. We hope that methods such as the ones examined in this paper will be more widely utilized in addressing such concerns.

Sources of financial support: EAS's time was supported 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) and the Robert Wood Johnson Foundation Health & Society Scholars program.
Acknowledgements: We thank Drs. Brian Schwartz and Thomas Glass for support in providing the Baltimore Memory Study data.

References

1. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci.* 2010;25(1):1.

2. Steiner PM, Cook TD, Shadish WR. On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*. 2011;36(2):213–236.

3. Millimet DL. The elephant in the corner: a cautionary tale about measurement error in treatment effects models. *Advances in Econometrics: Missing-Data Methods A*. 2011; 27:1–39.

4. Hong H, Rudolph KE, Stuart EA. Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika*. 2016: In press.

5. McCaffrey DF, Lockwood J, Setodji CM. Inverse probability weighting with error-prone covariates. *Biometrika*. 2013;100(3):671-80.

6. Lockwood J, McCaffrey DF. Matching and weighting with functions of error-prone covariates for causal inference. *J Am Stat Assoc*. 2015; DOI: 10.1080/01621459.2015.1122601

7. Webb-Vargas Y, Rudolph KE, Lenis D, et al. An imputation-based solution to using mismeasured covariates in propensity score analysis. *Stat Methods Med Research*. 2015; DOI: 10.1177/0962280215588771

8. Dehejia RH, Wahba S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J Am Stat Assoc*. 1999;94(448):1053–62.

9. Higgins JR, de Swiet M. Blood-pressure measurement and classification in pregnancy. *Lancet*. 2001;357(9250):131–35.

10. Heckman JJ, Singer B. *Longitudinal analysis of labor market data*. Cambridge, UK: Cambridge University Press, Publishers; 2008.

11. Hernan M, Robins J. *Causal inference*. Boca Raton, FL: Chapman & Hall/CRC, Publishers; forthcoming.

12. Blackwell M, Honaker J, King G. A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods and Research*. 2015;1–39.

13. Pearl J. *Causality*. Cambridge, UK: Cambridge University Press, Publishers; 2009.

14. Sturmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol*. 2005;162(3):279–89.

15. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr*. 1997;65(4):1179S–86S.

16. Sturmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration—a simulation study. *Am J Epidemiol*. 2007;165(10):1110–8.

17. Lunt M, Glynn RJ, Rothman KJ, et al. Propensity score calibration in the absence of surrogacy. *Am J Epidemiol*. 2012;175(12):1294–1302.

18. VanderWeele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*. 2011;22(1):42–52.

19. Rosenbaum P. *Design of observational studies.* New York, NY: Springer series in statistics, Publishers; 2010.

20. Keele LJ. Package 'rbounds'. 2014.

21. Gastwirth JL, Krieger AM, Rosenbaum PR. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*. 1998;85(4):907–20.

22. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement error in nonlinear models: a modern perspective*. Boca Raton, FL: Chapman & Hall/CRC, Publishers; 2012.

23. Schwartz BS, Glass TA, Bolla KI, et al. Disparities in cognitive functioning by race/ethnicity in the Baltimore Memory Study. *Environ Health Perspect.* 2004;112(3):314-20.

24. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Applied psychological measurement*. 1977;1(3):385–401.

25. Turner RJ, Noh S. Physical disability and depression: A longitudinal analysis. *J Health Soc Behav*. 1988;29(1):23–37.

26. McCandless LC, Gustafson P, Levy A. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat Med*. 2007;26(11):2331–47.

27. Gustafson P, McCandless L, Levy A, et al. Simplified bayesian sensitivity analysis for mismeasured and unobserved confounders. *Biometrics.* 2010;66(4):1129–37.

28. Lin HW, Chen YH. Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *Am J Epidemiol*. 2014;180(3):308-17.

29. Carroll R, Gail M, Lubin J. Case-control studies with errors in covariates. *J Am Stat Assoc*. 1993;88(421):185–99.

30. Ghosh-Dastidar B, Schafer JL. Multiple edit/multiple imputation for multivariate continuous data. *J Am Stat Assoc*. 2003;98(464):807–17.

31. Imai K, Yamamoto T. Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis. *American Journal of Political Science*. 2010;54(2):543–60.

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

33. Hossain S, Gustafson P Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Stat Med*. 2009;28(11):1580–1600.

34. Gustafson P. Measurement error modelling with an approximate instrumental variable. *J R Stat Soc Series B Stat Methodol*. 2007; 69(5):797–815.

Supplementary Web Appendix for: Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods

1 Simulation overview and set-up

We demonstrate how to implement each of the sensitivity analysis approaches discussed in the main text with a limited simulation study. Using PSC, we estimate the conditional ATE; using Vanderweele and Arah's bias formulas, we estimate the marginal ATE (which is the same as the conditional ATE under our data generating mechanisms); and using Rosenbaum sensitivity analysis, we estimate the p-value. For each simulation, we consider a sample of size $N = 10,000$ and run 1,000 simulation iterations. We evaluate performance in terms of mean percent bias, mean variance, 95% confidence interval coverage, and mean squared error (MSE).

In practice, we would generate multiple corrected estimates based on a variety of sensitivity parameters or validation datasets. For the purposes of this limited simulation study, however, we assume correct knowledge of sensitivity parameters and the presence of a generalizable validation dataset. Thus, our simulation study represents an optimal performance bound—how well each method could perform under the given data-generating mechanism.

Table 1 gives the 1) classical, 2) systematic differential, and 3) heteroscedastic measurement error data-generating mechanisms, following the same notation as in the main text. As seen in Table 1, we use the same data-generating mechanisms for both the PSC and Vanderweele and Arah bias formula approaches. We use different data-generating mechanisms for the Rosenbaum approach for two reasons: first, to ensure an inferential difference between the naive and true estimate, and second, to use a binary Y variable to improve interpretation of the Δ sensitivity parameter.

Web Table 1: Simulation data generating mechanisms. For the Rosenbaum sensitivity analysis, only changes to the data generating mechanisms are given.

Classical	Systematic Differential	Heteroscedastic Differential
Propensity score calibration & Vanderweele and Arah's bias formulas		
$Z \sim N(1,1)$ $X \sim N(1+0.2Z,1)$ $A \sim Ber(Logit^{-1}(2log(1.2) - log(1.2)X - log(1.2)Z))$		
$W \sim N(X, \sqrt{2})$	$W \sim N(X+1.2AX, 0.5)$	$W \sim N(X, \sqrt{0.5(1+A)^2})$
$Y \sim 3A + 3X + 2Z$		
Rosenbaum's sensitivity analysis		
$A \sim Ber(Logit^{-1}(-4log(2) + log(2)X + log(2)Z))$		
	$W \sim N(X - AX, 0.5)$	
$Y \sim 2X + 2Z$		

In the next section, we provide annotated R code to implement each measurement error sensitivity analysis.

2 R code

2.1 Sensitivity analyses functions

```
require(zoo)
require(nlme)
require(MatchIt)

# the simulation set-up code in the following section makes a data frame with the following columns
# t = treatment variable, 0/1
# w = observed covariate measured with error, continuous
# x = unobserved covariate measured without error, continuous
# z = observed covariate measured without error, continuous
# u = w-x
# y = outcome variable, continuous
# validation = indicator of whether observation is in validation subset, 0/1

expit<-function(p){
  exp(p)/(1+exp(p))
}
logit<-function(p){
  log(p/(1-p))
}
naiveiptw<- function(dat){
    sampled.data<-dat

    sampled.data$ps<-predict(glm(t ~ w + z, data=sampled.data, family="binomial"), type="response")
```

```
24      sampled.data$iptw<-ifelse(sampled.data$t==1, 1/sampled.data$ps, 1/(1-
            sampled.data$ps))
25      return(summary(glm(y ~ t , data=sampled.data, family="gaussian",
            weights=sampled.data$iptw))$coef[2])
26          }
27
28  #PSC, LS (classical measurement error)
29  psc <- function(dat){
30    sampled.data<-dat
31
32    validationdat<-sampled.data[sampled.data$validation==1,]
33    studydat<-sampled.data[sampled.data$validation==0,]
34
35    #fit EP ps model in validation set
36    validationdat$epps<-predict(glm(t ~ w*z, data=validationdat, family="
          binomial"), type="response")
37    #fit GS ps model in validation set
38    validationdat$gsps<-predict(glm(t ~ x + z, data=validationdat, family="
          binomial"), type="response")
39
40    model<-glm(gsps ~ t+logit(epps), data=validationdat, family="binomial")
41
42    #get EP ps in study dataset
43    studydat$epps<-predict(glm(t ~ w*z, data=studydat, family="binomial"),
          type="response")
44    studydat$predgsps<-predict(model, newdata=studydat, type="response")
45
46    return(summary(glm(y ~ t + logit(predgsps), data=studydat, family="
          gaussian"))$coef[2])
47      }
48
49  #PSC, WLS (systematic differential measurement error)
50  psc.wls <- function(dat){
51    sampled.data<-dat
52
53    validationdat<-sampled.data[sampled.data$validation==1,]
54    studydat<-sampled.data[sampled.data$validation==0,]
55
56    #fit EP ps model in validation set
57    validationdat$epps<-predict(glm(t ~ w+z, data=validationdat, family="
          binomial"), type="response")
58    #fit GS ps model in validation set
59    validationdat$gsps<-predict(glm(t ~ x + z, data=validationdat, family="
          binomial"), type="response")
60
61    glsfit<-gls(qlogis(gsps) ~t*qlogis(epps)+z, data=validationdat, weights=
          varIdent(form = ~ 1 | t))
62
63    #get EP ps in study dataset
64    studydat$epps<-predict(glm(t ~ w+z, data=studydat, family="binomial"),
          type="response")
65    studydat$predgsps<-predict(glsfit, newdata=studydat)
66
```

```
67    return(summary(glm(y ~ t + predgsps, data=studydat, family="gaussian"))$
          coef[2])
68      }
69
70  #Vanderweele and Arah, classical measurement error
71  vasimp<- function(dat){
72      sampled.data<-dat
73
74      atew<-naiveiptw(sampled.data)
75
76      correst<-atew - (summary(lm(y ~ t + z + w + u, data=sampled.data))$
            coefficient[5] * summary(lm(u ~ t + z + w , data=sampled.data))$
            coefficient[2])
77
78      return(correst)
79  }
80
81  #Vanderweele and Arah, systematic differential measurement error
82  vanonconst<- function(dat){
83      sampled.data<-dat
84
85      predut1<-predict(lm(u ~ w + z, data=sampled.data[sampled.data$t==1,]),
            newdata=sampled.data)
86      predut0<-predict(lm(u ~ w + z, data=sampled.data[sampled.data$t==0,]),
            newdata=sampled.data)
87      predu<-predict(lm(u ~ w + z, data=sampled.data), newdata=sampled.data)
88
89      bias<-(summary(lm(y~ z + w+ u, data=sampled.data[sampled.data$t==1,]))$
            coefficient[4]*mean(predut1-predu)) - (summary(lm(y~ z + w+ u, data
            =sampled.data[sampled.data$t==0,]))$coefficient[4]*mean(predut0-
            predu))
90      atew<-naiveiptw(sampled.data)
91
92      correst<-atew - bias
93      return(correst)
94  }
95
96
97  #Rosenbaum sensitivity analysis
98  p.upper<- function(theta, pi) {
99        (theta*pi+(1-theta)*(1-pi))
100       }
101
102 rosenbaum<-function(dat){
103   sampled.data<-dat
104
105   delta<-exp(summary(glm(y ~ t + z + w + u, data=data, family="binomial"))$
          coefficients[5])
106   gamma<-exp(summary(glm(t ~ z + w + u, data=data, family="binomial"))$
          coefficients[4])
107
108   theta <- delta/(1+delta)
109   pi <- gamma/(1+gamma)
```

```
110
111   m.out<-matchit(t ~ w + z, data = data, method = "nearest", replace=FALSE)
112    pairs<-data.frame(treated=as.numeric(row.names(m.out$match.matrix)),
          control=as.numeric(m.out$match.matrix))
113
114    ys<-matrix(rep(NA, 10*nrow(pairs)), ncol=10)
115    for(j in 1:nrow(ys)){
116         ys[j,1]<-c(data$y[row.names(data)==pairs[j,1]])
117         ys[j,2]<-c(data$y[row.names(data)==pairs[j,2]])
118       }
119
120    ysdat<-data.frame(treated=ys[,1], control=ys[,2])
121    ysdat$x<-ifelse(ysdat$treated==1 & ysdat$control==0, 1, 0)
122    ysdat$n<-ifelse(ysdat$treated!=ysdat$control, 1, 0)
123
124    #corrected
125    delta<-exp(summary(glm(y ~ t + z + w + u, data=data, family="binomial"))$
          coefficients[5])
126    gamma<-exp(summary(glm(t ~ z + w + u, data=data, family="binomial"))$
          coefficients[4])
127
128    theta <- delta/(1+delta)
129    pi <- gamma/(1+gamma)
130
131    return(binom.test(sum(ysdat$x), sum(ysdat$n), p=p.upper(theta, pi),
          alternative="greater")$p.value)
132 }
```

SAFunctions.R

2.2 Sample code for simulation

```
1  ####################################
2  ## PSC and Vanderweele and Arah sensitivity analysis
3  ####################################
4  expit<-function(p){
5    exp(p)/(1+exp(p))
6  }
7
8  ## Classical measurment error data generating mechanism
9
10 nsims <- 1000
11 set.seed(132)
12 n<-10000
13 truth<-3
14
15 # z is a coefficient measured without error
16 z<-rnorm(n, 1, 1)
17 # x is a coefficient measured with error
18 x<-rnorm(n, 1+ (.2*z), 1)
19
20 # the treatment depends on both x and z
```

```
beta0<- 2*log(2)
beta1<- -log(2)
beta2<- -log(2)
t<-rbinom(n, 1, prob=expit(beta0 + beta1*z + beta2*x))

# y is the outcome variable
psi0<-0
psi1<-3
psi2<-3
psi3<-2
meany<-psi0 + psi1*t + psi2*x + psi3*z
y<- rnorm(n, meany, 1)

# w is the mismeasured x.
# classical measurement error
gamma0<-0
gamma1<-1
gamma2<- 0
gamma3<- 0
meanw<-gamma0 + gamma1*x + gamma2*t + gamma3*t*x
w<-rnorm(n, meanw, sqrt(2))

#systemtatic differential measurement error
#gamma3<- 1.2
#meanw<-gamma0 + gamma1*x + gamma2*t + gamma3*t*x
#w<-rnorm(n, meanw, .5)

#heteroscedastic
#delta0<-0.5 #measurement error in control group
#delta1<-1 #extra measurement error in tx group
#mevar<-delta0*(1+(delta1*t))^2
#w<-rnorm(n, x, sqrt(mevar))

u<-w-x

validation<-rbinom(n,1, prob=.1)

data<-data.frame(z,x,t,w,y, u, validation)


#####################################
## Rosenbaum sensitivity analysis
#####################################

## Classical measurment error data generating mechanism
set.seed(132)
n<-10000

# z is a coefficient measured without error
z<-rnorm(n, 1, 1)
# x is a coefficient measured with error
x<-rnorm(n, 1+ (.2*z), 1)

```

```
# the treatment depends on both x and z
beta0<- -4*log(2)
beta1<-log(2)
beta2<-log(2)
t<-rbinom(n, 1, prob=expit(beta0 + beta1*z + beta2*x))
probt<-expit(beta0 + beta1*z + beta2*x)

# w is the mismeasured x.
#classical measurement error
gamma0<-0
gamma1<-1
gamma2<- 0
gamma3<- 0
meanw<-gamma0 + gamma1*x + gamma2*t + gamma3*t*x
w<-rnorm(n, meanw, sqrt(2))

#systematic differential measurement error
#gamma3<- -1
#meanw<-gamma0 + gamma1*x + gamma2*t + gamma3*t*x
#w<-rnorm(n, meanw, .5)

#heteroscedastic measurement error
#delta0<-0.5 #measurement error in control group
#delta1<-1 #extra measurement error in tx group
#mevar<-delta0*(1+(delta1*t))^2
#w<-rnorm(n, x, sqrt(mevar))

u<-w-x

# y is the outcome variable
psi0<-0
psi1<-0
psi2<-2
psi3<-2

y<-rbinom(n, 1, prob=expit(psi0 + psi1*t + psi2*x+ psi3*z))

data<-data.frame(x=x, w=w, t=t, y=y, z=z, u=u)
```

sim.R