Supporting Information for "Addressing Confounding and Exposure Measurement Error Using Conditional Score Functions" by Bryan S. Blette, Peter B. Gilbert, and Michael G. Hudgens

1 Web Appendix A: Large Sample Properties

In Web Appendix A, the large sample properties of the proposed estimators discussed in section 3 of the paper are proven.

1.1 G-formula CSM estimator

1.1.1 Large sample properties

Consistency and asymptotic normality of the g-formula estimator are proven using standard estimating equation theory (Stefanski and Boos, 2002). The original CSM estimator is an M-estimator and the g-formula can be written in the form of an unbiased estimating equation. Thus, the proposed g-formula CSM estimator can be written as

$\sum_{i=1}^{n} \psi_{GF-CSM}(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Sigma_{me}, \Theta_{GF}) = 0$, where $\Theta_{GF} = (\beta_0, \beta_a^T, \beta_l^T, vec(\beta_{al}), \phi, E\{Y(\boldsymbol{a})\})$

and:

$$\psi_{GF-CSM}(Y, \boldsymbol{L}, \boldsymbol{A}^*, \Sigma_{me}, \Theta_{GF}) = \begin{bmatrix} \{Y - \mathrm{E}(Y|\boldsymbol{L}, \boldsymbol{\Delta})\}(1, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{L} \otimes \boldsymbol{\Delta})^T \\ \phi - \{Y - \mathrm{E}(Y|\boldsymbol{L}, \boldsymbol{\Delta})\}^2/\{\mathrm{Var}(Y|\boldsymbol{L}, \boldsymbol{\Delta})/\phi\} \\ g^{-1}(\beta_0 + \boldsymbol{a}\beta_a + \boldsymbol{L}\beta_l + \boldsymbol{a}\beta_{al}\boldsymbol{L}^T) - E\{Y(\boldsymbol{a})\} \end{bmatrix}$$

The parameter of interest $E\{Y(\boldsymbol{a})\}$ is in the last estimating equation of the stack, and the proof relies on (i) the usual g-formula proof based on the causal assumptions made in Section 2 of the paper and (ii) that the CSM estimator $\hat{E}(Y|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{L})$ was previously shown to be consistent (Carroll et al., 2006):

$$\begin{aligned} E[g^{-1}(\beta_0 + \boldsymbol{a}\beta_a + \boldsymbol{L}\beta_l + \boldsymbol{a}\beta_{al}\boldsymbol{L}^T) - E\{Y(\boldsymbol{a})\}] &= E\{g^{-1}(\beta_0 + \boldsymbol{a}\beta_a + \boldsymbol{L}\beta_l + \boldsymbol{a}\beta_{al}\boldsymbol{L}^T)\} - E\{Y(\boldsymbol{a})\} \\ &= E\{\hat{E}(Y|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{L})\} - E[E\{Y(\boldsymbol{a})|\boldsymbol{L}\}] \\ &= E\{E(Y|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{L})\} - E[E\{Y(\boldsymbol{a})|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{L}\}] \\ &= E\{E(Y|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{L})\} - E\{E(Y|\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{L})\} \\ &= 0 \end{aligned}$$

So the estimating function for the parameter $E\{Y(\boldsymbol{a})\}$ is unbiased. Denote $\hat{\Theta}_{GF}$ as the solution to $\sum_{i=1}^{n} \psi_{GF-CSM}(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Sigma_{me}, \hat{\Theta}_{GF}) = 0$. Then by the proof above $\sqrt{n}(\hat{\Theta}_{GF} - \Theta_{GF}) \sim N(\boldsymbol{0}, A^{-1}B(A^{-1})^T)$ where A and B are consistently estimated by

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\Theta_{GF}^T} \psi_{GF-CSM}(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Sigma_{me}, \hat{\Theta}_{GF})$$

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} \psi_{GF-CSM}(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Sigma_{me}, \hat{\Theta}_{GF}) \psi_{GF-CSM}^T(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Sigma_{me}, \hat{\Theta}_{GF})$$

In the R code implementing the methods, this sandwich variance estimation is accomplished using the R package geex (Saul and Hudgens, 2020).

1.1.2 Relationship to classical causal estimators

It has been noted (see Carroll et al. (2006)) that the CSM estimating equations reduce to the score equations for a GLM when the measurement error covariance matrix $\Sigma_{me} = 0_{m\times m}$. Thus under no measurement error, the procedure described above reduces to a stack of estimating equations corresponding to the common practice of performing the g-formula while specifying a GLM for the outcome regression, making it a special case of the proposed estimator.

1.2 IPW CSM estimator

1.2.1 Large sample properties

Consistency and asymptotic normality of the IPW CSM estimator are proven as above, using M-estimator theory. The partial M-estimator (Stefanski and Boos, 2002) corresponding to the parameters of interest is $\sum_{i=1}^{n} \psi(Y_i, \boldsymbol{Z}_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Theta_{IPW}) = \sum_{i=1}^{n} SW_i(Y_i - E[Y_i|\boldsymbol{\Delta}_i])\boldsymbol{\Delta}_i^T = 0$. It suffices to show that the expectation of the estimating function $\psi(Y, \boldsymbol{Z}, \boldsymbol{L}, \boldsymbol{A}^*, \Theta_{IPW})$ is equal to 0. Although $\boldsymbol{\Delta} = (\Delta_1, ..., \Delta_m)$ is a vector of length m, the estimator form is the same for each row of the estimating equation stack and without loss of generality, the estimating function is proven to be unbiased for $SW(Y - E[Y|\boldsymbol{\Delta}])\Delta_k$ for an arbitrary $1 \leq k \leq m$.

First consider a similar but infeasible estimator, with the same form but where it is weighted by the true propensity weights. Let $SW = \frac{h(\boldsymbol{A})}{f(\boldsymbol{A}|\boldsymbol{L})}$ such that the numerator is any function of $\boldsymbol{A}$ and the denominator is a conditional density of exposures given confounders. Furthermore, suppose that the denominator density equals the true conditional density, denoted $f(\boldsymbol{A}|\boldsymbol{L}) = f_0(\boldsymbol{A}|\boldsymbol{L})$. Let E_0 notation refer to taking the expectation under the true causal parameter vector from the MSM, which is nested within $E(Y|\boldsymbol{\Delta})$. In a slight abuse of notation, let Δ_k be the random variable corresponding to the k^{th} element of the random vector $\boldsymbol{\Delta}$, rather than the vector $\boldsymbol{\Delta}$ for individual k which is instead notated as bold $\boldsymbol{\Delta}_k$ in the manuscript.

$$
\begin{aligned}
E_0\left[\frac{h(\boldsymbol{A})}{f_0(\boldsymbol{A}|\boldsymbol{L})}\{Y - E(Y|\boldsymbol{\Delta})\}\Delta_k\right] &= E_0\left(E\left[\frac{h(\boldsymbol{A})}{f_0(\boldsymbol{A}|\boldsymbol{L})}\{Y^{\boldsymbol{A}} - E(Y^{\boldsymbol{A}}|\boldsymbol{\Delta})\}\Delta_k|\boldsymbol{L}\right]\right) \\
&= E_0\left[\int_{\boldsymbol{a}} \frac{h(\boldsymbol{a})}{f_0(\boldsymbol{a}|\boldsymbol{L})}\{Y^{\boldsymbol{a}} - E(Y^{\boldsymbol{a}}|\boldsymbol{\Delta})\}\Delta_k f_0(\boldsymbol{a}|\boldsymbol{L})d\mu(\boldsymbol{a})\right] \\
&= E_0\left\{\int_{\boldsymbol{a}} h(\boldsymbol{a})(Y^{\boldsymbol{a}} - E[Y^{\boldsymbol{a}}|\boldsymbol{\Delta}])\Delta_k d\mu(\boldsymbol{a})\right\} \\
&= \int_{\boldsymbol{a}} E_0\left\{h(\boldsymbol{a})(Y^{\boldsymbol{a}} - E[Y^{\boldsymbol{a}}|\boldsymbol{\Delta}])\Delta_k\right\} d\mu(\boldsymbol{a}) \\
&= \int_{\boldsymbol{a}} E_0\left\{E\left[h(\boldsymbol{a})(Y^{\boldsymbol{a}} - E[Y^{\boldsymbol{a}}|\boldsymbol{\Delta}])\Delta_k|\boldsymbol{\Delta}\right]\right\} d\mu(\boldsymbol{a}) \\
&= \int_{\boldsymbol{a}} E_0\left\{h(\boldsymbol{a})\Delta_k E\left[(Y^{\boldsymbol{a}} - E[Y^{\boldsymbol{a}}|\boldsymbol{\Delta}])|\boldsymbol{\Delta}\right]\right\} d\mu(\boldsymbol{a}) \\
&= \int_{\boldsymbol{a}} E_0\left\{h(\boldsymbol{a})\Delta_k(E[Y^{\boldsymbol{a}}|\boldsymbol{\Delta}] - E[Y^{\boldsymbol{a}}|\boldsymbol{\Delta}])\right\} d\mu(\boldsymbol{a}) \\
&= 0
\end{aligned}
$$

where $d\mu(\boldsymbol{a})$ is defined as the Lebesgue measure. The first equality uses causal consistency, the second equality uses conditional exchangeability, and positivity is needed for the integral to be well-defined. Thus the infeasible estimator is consistent and asymptotically normal by standard M-estimator theory. The asymptotic variance is given by the usual sandwich estimator as described in the previous section.

From here there are two jumps to the corresponding estimator in the paper where weights are estimated. The first is that one needs to estimate the treatment weights from some kind of model, even if no treatments were mismeasured. This substitution is well known to result in a consistent estimator as long as the propensity score model is correctly specified, because then this estimator will equal the estimator described above plus an $o_p(1)$ term. The second jump was alluded to in Section 3.3 of the paper, that one can use weights estimated from a propensity model that is fit using the mismeasured exposures. This will not necessarily affect the consistency of the estimator. For example, suppose the exposures are independent given $\boldsymbol{L}$ and that each exposure has a linear relationship with the confounders, i.e., with

simplified scalar notation, $A = \boldsymbol{L}\alpha + \epsilon_{ps}$. Then under additive measurement error, each mismeasured observed exposure also has a linear relationship with the confounders given by: $A^* = \boldsymbol{L}\alpha + \epsilon_{ps} + \epsilon_{me}$. So if linear propensity models are fit using the mismeasured exposures (noting that $\boldsymbol{A}^*$ is a collider on the only path connecting ϵ_{me} and $\boldsymbol{L}$), one would still get consistent estimates of the propensity model parameters α (and subsequently the weights), albeit with more variability. Therefore the proposed estimator would still be consistent. When exposures have non-linear, complex relationships with confounders, consistency may not be guaranteed, but previous explorations of this topic suggest that the measurement error will likely only introduce mild issues (Carroll et al., 2006).

1.2.2 Relationship to classical causal estimators

Note that when the measurement error covariance matrix $\Sigma_{me} = \mathbf{0}_{m\times m}$, the sufficient statistic $\boldsymbol{\Delta}$ reduces to the observed exposure vector. Then the IPW estimator reduces to the form $\sum_{i=1}^{n} \psi(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Theta_{IPW}) = 0$ where:

$$\psi(Y, \boldsymbol{L}, \boldsymbol{A}^*, \Theta_{IPW}) = \begin{bmatrix} SW\{Y - E(Y|\boldsymbol{A}^*)\}(1, \boldsymbol{A}^*)^T \\ SW\left[\phi - \frac{\{Y-E(Y|\boldsymbol{A}^*)\}^2}{Var(Y|\boldsymbol{A}^*)/\phi}\right] \end{bmatrix}$$

This is exactly the score function vector for a GLM weighted by SW. Thus, an IPW estimator fit using a weighted GLM for outcome Y is a special case of the proposed IPW CSM estimator where there is no measurement error present.

1.3 DR CSM Estimator

1.3.1 Large Sample Properties

Once again, consistency and asymptotic normality of the proposed estimator is proven using M-estimator theory. This estimator is a solution to the estimating equation

$\sum_{i=1}^{n} \psi_{DR-CSM}(Y_i, \boldsymbol{L}_i, \boldsymbol{A}_i^*, \Sigma_{me}, \Theta_{DR}) = 0$, where $\Theta_{DR} = \Theta_{GF}$ and

$$\psi_{DR-CSM}(Y, \boldsymbol{L}, \boldsymbol{A}^*, \Sigma_{me}, \Theta_{DR}) = \begin{bmatrix} SW\{Y - \mathrm{E}(Y|\boldsymbol{L}, \boldsymbol{\Delta})\}(1, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{L} \otimes \boldsymbol{\Delta})^T \\ SW[\phi - \{Y - \mathrm{E}(Y|\boldsymbol{L}, \boldsymbol{\Delta})\}^2 / \{\mathrm{Var}(Y|\boldsymbol{L}, \boldsymbol{\Delta})/\phi\}] \\ g^{-1}(\beta_0 + \boldsymbol{a}^*\beta_a + \boldsymbol{l}\beta_l + \boldsymbol{a}^*\beta_{al}\boldsymbol{l}^T) - E\{Y(\boldsymbol{a})\} \end{bmatrix}$$

First, suppose that the propensity score models are correctly specified. This effectively means that we are fitting the outcome regression in a pseudo-population where the effect of treatment on the outcome is no longer confounded. This means that integrating out $\boldsymbol{L}$ in the standardization step will have no effect, and the DR CSM estimator will be asymptotically equivalent to the IPW estimator, which was previously shown to be consistent and asymptotically normal when the propensity models are correctly specified.

Now, suppose that the outcome regression is correctly specified. Then the proof follows the same form as the g-formula CSM proof since the outcome predictions from the IP-weighted CSM model converge in expectation to outcome predictions from the non-weighted CSM model regardless of whether the propensity model is correctly specified. However, the $\hat{\beta}$ estimated by the DR CSM estimator when the propensity model is mis-specified may be a particularly inefficient estimator of the true β outcome model parameters. Robins et al. (2007) considered this a notable weakness of this type of DR estimator, but also concluded that this DR estimator form is stronger than other described DR forms when weights are highly variable (which is extremely common in the continuous exposure setting Naimi et al. (2014)).

1.3.2 Relationship to classical causal estimators

Once again, under no measurement error the CSM equations reduce to the score equations of a GLM. Thus the proposed DR CSM estimator will reduce to the DR estimator described in Hirano and Imbens (2001) where the specified outcome regression is a GLM.

1.4 Uniqueness of EE solutions

Each of the proofs above relies on there being a unique solution to each set of estimating equations. It has been noted in prior work (Stefanski and Carroll, 1987) that similar conditional score equations do not always have unique solutions, but that multiple solutions are very rare in practice. In the various simulations of this paper, multiple solutions or estimator divergence were encountered with similar rarity, at most 1 or 2 times per 2000 simulations, unless considering extreme data generating mechanisms. Thus, the estimators should have good behavior in general, but practitioners should be aware of rare instances of multiple solutions, unusual estimates, and/or root-solving algorithm divergence errors.

2 Web Appendix B: Accounting for Two-Phase Sampling

Many studies (including the HVTN 505 trial) use a two-phase sampling design. Such a design is particularly useful when the primary exposure(s) and outcome are easy to measure, but exposures and covariates of secondary interest are expensive or difficult to measure. Because each of the proposed methods above belongs to the estimating equation framework, it is straightforward to incorporate previously described methods for causal inference from studies with two-phase sampling. In this section, one such approach is demonstrated using a simulation study. In particular, for this simulation and the application section analysis, the simple IPW method described in Wang et al. (2009) is implemented, but the DR approaches from the same paper or from Rose and van der Laan (2011) could also be explored to account for two-phase sampling within the proposed methods.

The simple IPW method is implemented by weighting each of the proposed estimating equations by the inverse probability of selection for the second-phase of the study (multiplying treatment weights by sampling weights for the IPW-CSM and DR-CSM estimators) and restricting the analysis to those selected. This works well for the subset of the HVTN

505 trial that is the focus of Section 5 of the paper, but may be inadequate when analyzing exposures measured in a sub-sample in conjunction with exposures measured in the full sample.

2.1 Two-phase sampling simulations

The structure of the first simulation study described in Section 4 of the paper is replicated, but under a two-phase sampling design. In particular, a case-cohort design is used where the exposure is measured for a random sub-cohort as well as for every case. This is done for a sample size of $n = 2000$ under three scenarios, with sub-cohorts of size 5%, 10% and 25%. The results of 2000 simulation runs are presented in Web Figure 1 and Web Table 1.

The methods seem to perform well as they did in the full-sampling simulation provided in the paper, although there is some bias and under-coverage when the sub-cohorts are smaller, likely due to a low effective sample size. In addition, the estimators failed to converge in some of the small sub-cohort settings. However, the DR CSM estimator with sampling weights converged in all analyses presented in Section 5 of the paper.

3 Web Appendix C: Additional Simulations

In this section, the methods are studied under two assumption violations: (i) when positivity doesn't hold and (ii) when measurement error doesn't follow a classical additive model.

3.1 Under positivity violation

To evaluate the proposed g-formula CSM method under positivity violation, the general structure of the first simulation study from Section 4 of the paper is replicated almost exactly. A moderate positivity violation is created by changing how the treatment A_1 is generated from $\mathcal{N}(2 + 0.3L_1 - 0.5L_2, 0.6)$ to $\mathcal{N}(2 + 0.3L_1 - 0.5L_2, 0.35)$. This breaks the phenomenon of mostly overlapping treatment values experienced by simulated subjects with different

covariate values, although in a technical sense is not a structural violation of positivity since the distributions would have the same support given infinite sample size. The results of the simulation study are presented in Web Figure 2 and Web Table 2, following the same format as Table 1 and Figure 2 in the paper.

The results overall look similar to that in Table 1 of the paper. There is some bias and undercoverage for the proposed g-formula CSM methods, but the proposed method still generally performs better than the comparator methods in this scenario. Performance is weaker at the extremes of the exposure support, which is expected given that there is very little data at the extremes with the new data generating mechanism and not fully due to the positivity violation. A reasonable range to evaluate the methods would be from 0.5 to 3.5 in these simulations.

Positivity violations become more likely with more treatment variables and with treatment variables that are continuous or take on many values. In these settings positivity should receive just as much scrutiny as the conditional exchangeability assumption. If positivity is implausible, it may be possible to define an estimator in our setting similar to that described in Neugebauer and van der Laan (2005) which was robust to their analogous "experimental treatment assumption".

3.2 Under non-additive measurement error

Next the proposed methods are evaluated when treatment measurement error does not follow the classical additive model. In particular, the second simulation study from Section 4 of the paper is replicated, but the simulation of mismeasured treatment A_3^* is changed such that it follows a multiplicative error model simulated as $A_3^* = A_3\epsilon_{me_1}$ where $\epsilon_{me_1} \sim \mathcal{N}(1, 0.1)$. The methods are still performed assuming additive measurement error with known measurement error covariance as specified in section 4 of the paper, and the A_3^* distribution under this additive ME assumption is similar to the distribution under the true multiplicative ME generative model. The results are presented in Web Table 3. The proposed IPW CSM

method continued to perform well for treatments A_1 and A_2 for which the assumptions hold, but exhibited strong bias for the A_3 effect. Practitioners of the proposed methods should be cautious that if classical additive measurement error models do not hold for their exposures, they may get worse results than even standard regression models.

4 Web Appendix D: More complex model specifications for the IPW CSM estimator

The proposed IPW CSM estimator assumes a linear marginal structural model form. While this is helpful to match the conditional score framework described in Section 3.1, it is too restrictive for some potential applications. To this end we note that transformations of elements of $\boldsymbol{A}$ and interactions thereof can be included in the MSM specification as long as they are either assumed to be correctly measured or assumed to follow a classical additive measurement error model. For example, if a transformation of an exposure is assumed to follow a multiplicative measurement error model then that variable cannot be included in the MSM. However, if the variable is strictly positive, then its log transform would follow an additive measurement error model and can be included in the model. In general, transformations of correctly measured exposures can be included in the MSM specification without restriction.

Finally, while conditional score functions are somewhat limited in scope in terms of model specification, the related method of corrected score functions has been extended to problems of additive but non-normal measurement error (Buzas and Stefanski, 1996) and to non-additive measurement models in certain cases (Nakamura, 1990; Li, Palta, and Shao, 2004). Describing how to use such corrected score functions to estimate causal parameters could be the focus of future work in this area.

Web Table 1. *Simulation study for case-cohort design. Bias: 100 times the average bias across simulated data sets for each method; ASE: 100 times the average of estimated standard errors; ESE: 100 times the standard deviation of parameter estimates; Cov: Empirical coverage of 95% confidence intervals for each method, rounded to the nearest integer. % FTC: Percent of simulations which failed to converge, rounded to the nearest integer.*

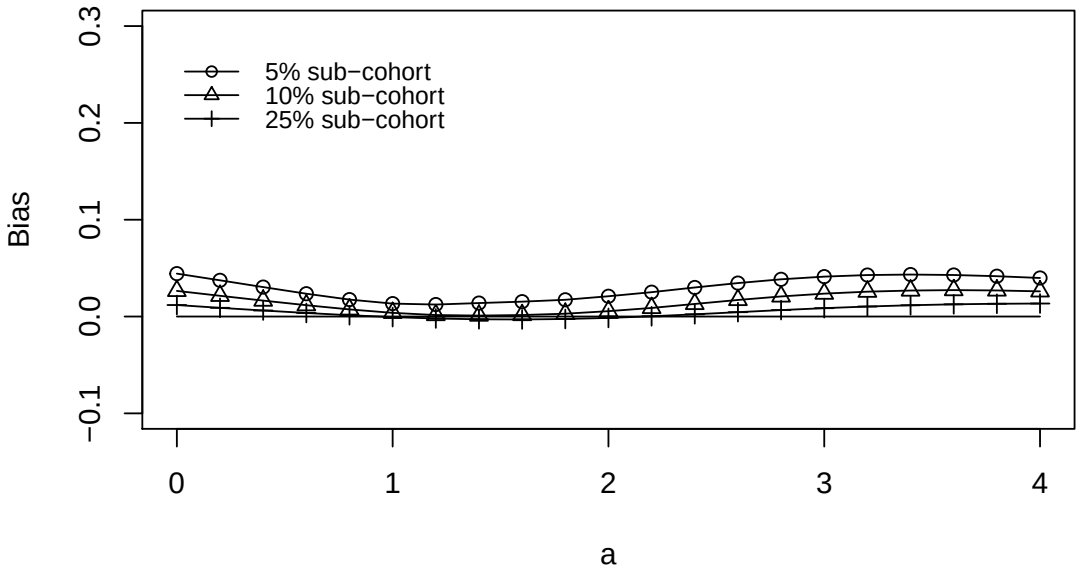
Sub-cohort Size	Bias	ASE	ESE	Cov	% FTC
5%	4.1	8.7	7.1	84%	6%
10%	2.4	6.3	5.6	90%	2%
25%	0.9	4.1	3.9	94%	0%

Web Table 2. *Simulation study under positivity violation. Bias, ASE, ESE, and Cov defined as in Web Table 1.*

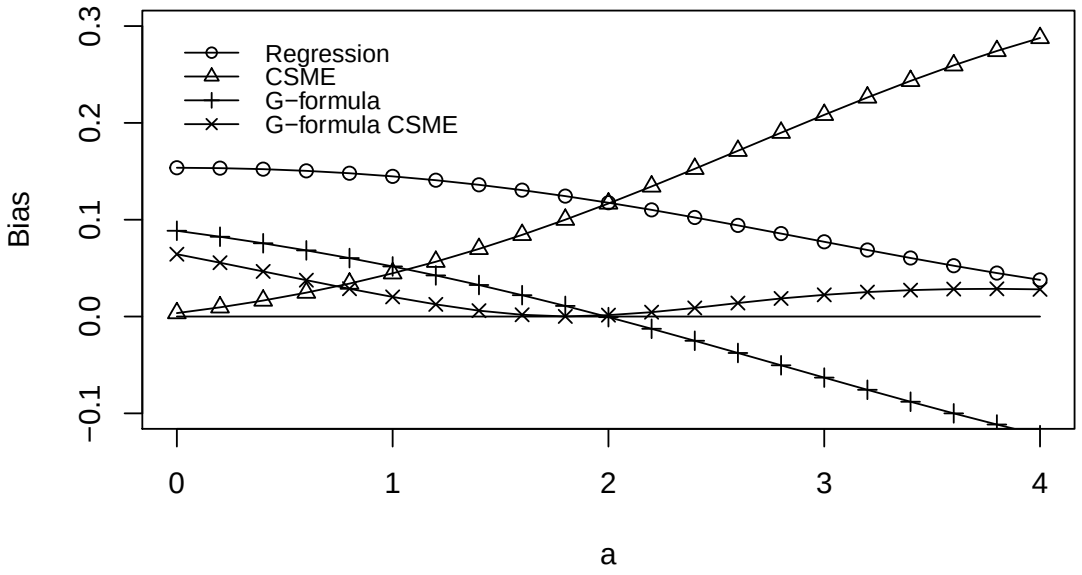
Estimator	Bias	ASE	ESE	Cov
Regression	-69.8	21.5	21.8	11%
CSM	-17.0	80.3	68.5	95%
G-formula	-6.3	3.0	3.0	44%
G-formula CSM	2.2	9.3	9.1	92%

Web Table 3. *Simulation study under non-additive measurement error. Bias, ASE, ESE, and Cov defined as in Web Table 1.*

	ψ_1				ψ_2				ψ_3			
Estimator	Bias	ASE	ESE	Cov	Bias	ASE	ESE	Cov	Bias	ASE	ESE	Cov
Regression	4.9	14.0	13.3	93%	10.3	28.0	27.7	93%	1.9	15.8	15.4	94%
CSM	21.7	20.0	19.3	82%	8.7	29.4	28.4	93%	-35.0	33.5	32.7	86%
IPW	-9.9	9.1	9.0	79%	0.0	20.0	19.7	94%	4.7	15.9	15.5	93%
IPW CSM	0.6	13.0	12.7	95%	-0.9	20.4	20.1	95%	-29.0	33.0	32.0	88%



Web Figure 1: Estimated dose-response curve bias for the DR CSM method compared across three sub-cohort sizes. Bias refers to the average bias across 2,000 simulated data sets for each method evaluated at each point on the horizontal axis $a = (0, 0.2, 0.4, ..., 4)$.



Web Figure 2: Estimated dose-response curve bias for each of the four methods under positivity violation. Bias refers to the average bias across 2,000 simulated data sets for each method evaluated at each point on the horizontal axis $a = (0, 0.2, 0.4, ..., 4)$.

References

Buzas, J. and Stefanski, L. (1996). A note on corrected-score estimation. *Statistics & Probability Letters* **28,** 1–8.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* **2,** 259–278.

Li, L., Palta, M., and Shao, J. (2004). A measurement error model with a poisson distributed surrogate. *Statistics in Medicine* **23,** 2527–2536.

Naimi, A. I., Moodie, E. E., Auger, N., and Kaufman, J. S. (2014). Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology* **25,** 292–299.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77,** 127–137.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* **129,** 405–426.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science* **22,** 544–559.

Rose, S. and van der Laan, M. J. (2011). A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics* **7,**.

Saul, B. and Hudgens, M. (2020). The calculus of M-Estimation in R with geex. *Journal of Statistical Software, Articles* **92,** 1–15.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician* **56,** 29–38.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74,** 703–716.

Wang, W., Scharfstein, D., Tan, Z., and MacKenzie, E. J. (2009). Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71,** 947–969.