Causal Selection of Covariates in Regression Calibration for Mismeasured Continuous Exposure

Wenze Tang, Donna Spiegelman, Xiaomei Liao, and Molin Wang

Abstract: Regression calibration as developed by Rosner, Spiegelman, and Willett is used to adjust the bias in effect estimates due to measurement error in continuous exposures. The method involves two models: a measurement error model relating the mismeasured exposure to the true (or gold-standard) exposure and an outcome model relating the mismeasured exposure to the outcome. However, no comprehensive guidance exists for determining which covariates should be included in each model. In this article, we investigate the selection of the minimal and most efficient covariate adjustment sets under a causal inference framework. We show that to address the measurement error, researchers must adjust for, in both measurement error and outcome models, any common causes (1) of true exposure and the outcome (2) of measurement error and the outcome. We also show that adjusting for so-called prognostic variables that are independent of true exposure and measurement error in the outcome model, may increase efficiency, while adjusting for any covariates that are associated only with true exposure generally results in efficiency loss in realistic settings. We apply the proposed covariate selection approach to the Health Professional Follow-up Study dataset to study the effect of fiber intake on cardiovascular disease. Finally, we extend the originally proposed estimators to a nonparametric setting where effect modification by covariates is allowed.

Keywords: Causal inference; Continuous exposure; Covariate selection; Measurement error; Regression calibration

Regression calibration as a popular tool to address the bias due to nondifferential measurement errors in continuous exposures. There are two versions of regression calibration methods: one developed by Rosner, Spiegelman, and Willett and the other developed by Carroll, Ruppert, and Stefanski. Both versions involve two models: a measurement error model (also known as calibration model) linking the mismeasured exposure to the true or gold-standard exposure and an outcome model linking either the mismeasured exposure or imputed true or gold-standard exposure to the error-free outcome. For both methods, a set of pretreatment covariates is typically adjusted for in the measurement error model and the outcome model, but what should be included in this set of covariates has not been systematically discussed.

For example, in assessing the effect of fiber intake on the risk of cardiovascular disease (CVD), age and smoking are two important confounding variables for the exposure-outcome relationship, with age potentially also influencing measurement error by affecting memory and cognitive function. Other risk factors for CVD such as marital status, in contrast, may not be associated with either fiber intake or measurement error.

In this article, we aim to answer the optimal covariate selection question in the Rosner, Spiegelman, and Willett regression calibration method under a counterfactual framework. With the aid of a directed acyclic graph (DAG), we investigate (1) what is the minimal covariate adjustment set that would assure validity and (2) whether adjusting for additional covariates outside of the minimal covariate adjustment set in the measurement error model and/or outcome model increases efficiency, including the question of whether different sets of covariates can be adjusted for in the two models.

In the Review section, we define the causal contrast of interest, the statistical models, and the resulting estimators, followed by an introduction to the relevant DAGs and assumptions. In the section of Validity and Efficiency of Regression Calibration Estimators, we provide the identification formula for our causal estimand. We then present the results regarding the validity and efficiency of Rosner, Spiegelman, and Willett estimators under each DAG structure. In Simulation, we present results from a series of Monte Carlo simulations evaluating the finite sample performance of each estimator. In Real-data Example, we apply the proposed covariate selection strategy to the study of the effect of fiber intake on the risk of CVD.
within the Health Professional Follow-up Study (HPFS). We conclude this article with a summary and further discussions.

**REVIEW**

**Causal Estimand of Interest**

We denote $X$, $Z$, $V$, $Y$ respectively as true or gold-standard exposure, surrogate/mismeasured exposure, covariate(s), and outcome, where $X$, $Z$ are assumed to be scalar and $V$ can be either a scalar or a set of covariates. We focus on the main study/external validation study design; that is, independent samples $(Z, V, Y)$ and $(X, Z, V)$ are available from the main study and validation study, respectively. We assume that the parameters in the measurement error model estimated using the validation study population would be the same as those estimated from the main study population had we been able to run the measurement error model in both populations (i.e., transportability condition). Following the standard causal inference literature, we define the potential outcome $Y^x$ as the value of $Y$ that each participant would have experienced had their exposure value been fixed at $X = x$.

We are primarily interested in the conditional average treatment effect (hereafter conditional ATE) on the additive scale denoted by $E[Y^x - Y^{x'} | V]$, that is the mean difference in the potential outcome had everyone been assigned exposure level $x$ versus exposure level $x'$ within levels of the covariates $V$. Parametric regression models typically estimate conditional ATEs. We may also be interested in the ATE within levels of the covariates $V$.

**Generalized Rosner, Spiegelman, and Willett Regression Calibration Estimators**

In the standard Rosner, Spiegelman, and Willett method, for a given covariate $V$, analysts typically choose one of the two linear measurement error models (adjusted model (1) vs. unadjusted model (2)) and one of the two linear outcome models (adjusted model (3) vs. unadjusted model (4)):

$$E[X|Z, V] = \alpha_0 + \alpha_1Z + \alpha_2V,$$

$$E[X|V] = \alpha_0^* + \alpha_1^*Z,$$

$$E[Y|Z, V] = \gamma_0 + \gamma_1Z + \gamma_2V,$$

$$E[Y|Z] = \gamma_0^* + \gamma_1^*Z,$$

where (1) and (3) imply (2) and (4), respectively.

We denote by subscript an estimator as $\hat{\beta}_{(OM)}$ if it includes a given $V$ in both the measurement error and outcome models and as $\hat{\beta}_{(-)}$ if the covariate is included in neither model:

$$\hat{\beta}_{(OM)} = \frac{\hat{\gamma}_1}{\hat{\alpha}_1},$$

(6)

In model (5) for example, $\alpha_1$ is also known as the attenuation factor, and by dividing $\gamma_1$, the biased effect estimate subject to measurement error, by $\alpha_1$ we can recover the exposure effect of interest. In this article, we also investigate the following extended estimators:

$$\hat{\beta}_{(-M)} = \frac{\hat{\gamma}_1^*}{\hat{\alpha}_1^*},$$

(7)

$$\hat{\beta}_{(-O)} = \frac{\hat{\gamma}_1}{\hat{\alpha}_1},$$

(8)

Thus, these estimators respectively represent covariate adjustment strategies: (1) $\hat{\beta}_{(OM)}$: $V$ is included in both models, (2) $\hat{\beta}_{(-M)}$: $V$ is included in neither model, (3) $\hat{\beta}_{(-O)}$: $V$ is only included in the measurement error model, and (4) $\hat{\beta}_{(O-)}$: $V$ is only included in the outcome model.

**Directed Acyclic Graphs for Measurement Error Structures**

In this article, we focus on eight variations of the nondifferential exposure measurement error structure, represented in the eight causal DAGs in Figure 1. The nondifferentiality assumption requires the independence between $Z$ and $Y$ conditional on $X$ in the absence of covariate.

We partition the covariate set $V$ into eight mutually exclusive subsets of the covariate(s) $V_j$ where $j = 1, \ldots, 8$, with each DAG reflecting how $V_j$ relates to $X$, $Z$ and $Y$ (Figure 1). For example, DAG 4 describes the scenario where the covariate $V_{4(ZXY)}$ is both a confounder and directly contributes to measurement error. To facilitate reading, we write the additional subscript $(XZY)$ to emphasize the fact that the covariate has an arrow to $X$, $Z$, and $Y$ under DAG 4. DAG 5 to 8 are replicates of DAG 1 to 4 except the covariate $V_j$ under consideration is no longer a risk factor for the outcome.

We can establish the following nondifferential measurement error conditions (i.e., surrogacy assumptions): (1) $Y \perp Z|X$ under DAG 1, 3, 5, 6, 7, and 8 and (2) $Y \perp Z|X, V_j$ under all DAGs, where $j = 1, \ldots, 8$. We can also establish no confounding assumptions (i.e., exchangeability assumptions) (1) $Y \perp X$ for DAG 1, 2, 5, 6, 7, and 8 and (2) $Y \perp X|V_j$ for all DAGs, where $j = 1, \ldots, 8$.

For simplicity but with no loss of generality, we consider covariates of each type in the absence of the other types. In realistic settings, all eight types would likely appear on the same DAG. We demonstrate how the proof for validity can be generalized to such complex settings via one particular example in Section 1.4 of the eAppendix; http://links.lww.com/EDE/C120, where we assume that both $(V_{3(X-Y)}, V_{4(X-Y)})$ are present.

© 2024 Wolters Kluwer Health, Inc. All rights reserved.
VALIDITY AND EFFICIENCY OF REGRESSION CALIBRATION ESTIMATORS

Identification of Causal Effect of Interest

We show in Section 1.2 of the eAppendix; http://links.lww.com/EDE/C120 that under the following sufficient conditions:

1. Exchangeability: \( Y^* \perp X | V \) (i.e., no confounding for the effect of \( X \) on \( Y \) given \( V \)),
2. Nondifferential measurement error (surrogacy): \( Y \perp Z | (X, V) \),
3. Consistency: \( Y^*, Z^* \) take observed values \( Y, Z \) when \( X = x \), and
4. Linearity: conditional ATE follows the form \( E[Y^* - Y^*'] = \beta(x - x') \), where \( \beta \) is the constant effect associated with one unit increase in \( X \),

we can nonparametrically identify \( \beta(V) \) and thus conditional ATE defined in the Review section for all DAGs 1 to 8 as:

\[
\beta(V) = \frac{E[Y|Z = z, V] - E[Y|Z = z', V]}{E[X|Z = z, V] - E[X|Z = z', V]}.
\] (9)

Furthermore, if we additionally make the following modeling assumption:

5. The models for \( E[X|Z = z, V = v] \) and \( E[Y|Z = z, V = v] \) as in (1) and (3) are correctly specified and that \( E[V|Z = z] \) is a linear function of \( Z \),

\( \beta(V) \) reduces to \( \beta \), where \( \beta \) is a constant effect associated with one unit increase in \( X \), and is identified as:

\[
\beta = \frac{\gamma_1}{\alpha_1},
\] (10)

which has the same form as the estimator (5).

Similarly, we show in the same Section of eAppendix; http://links.lww.com/EDE/C120 that under the following sufficient conditions:

1. Exchangeability: \( Y^* \perp X \) (i.e., no confounding for \( X \)'s effect on \( Y \)),
2. Nondifferential measurement error (surrogacy): \( Y \perp Z | X \),
3. Consistency: \( Y^*, Z^* \) take observed value \( Y, Z \) when \( X = x \), and
4. Linearity: ATE follows the form \( E[Y^* - Y^*'] = \beta(x - x') \), where \( \beta \) is the constant effect associated with one unit increase in \( X \),

we can nonparametrically identify \( \beta \) and thus ATE defined in the Review section for DAGs 1, 5, 6, 7, and 8 as:

\[
\beta = \frac{E[Y|Z = z] - E[Y|Z = z']}{E[X|Z = z] - E[X|Z = z']},
\] (11)

Under further modeling assumptions:

5. The models for \( E[X|Z = z] \) and \( E[Y|Z = z] \) as in (2) and (4) are correct and that \( E[V|Z = z] \) is a linear function of \( Z \), \( \beta \) is identified as:

\[
\beta = \frac{\gamma_1}{\alpha_1},
\] (12)

which has the same form as the estimator (6).

Validity and Efficiency of Regression Calibration Estimators

We use results from the previous subsection and the Review section to evaluate whether each estimator, \( \hat{\beta}_{(OM)}, \hat{\beta}_{(-O)}, \hat{\beta}_{(-M)} \), and \( \hat{\beta}_{(-OM)} \), correctly estimates the causal effect. We also analytically evaluate the asymptotic efficiency of valid estimators for a given DAG for a continuous outcome under linear models (1) to (4). See Sections 1 and 2 of the eAppendix; http://links.lww.com/EDE/C120 for proofs.

Summary of Analytical Results for Rosner, Spiegelman, and Willett Estimators

Table 1 summarizes the validity of different estimators under DAGs 1 to 8, with the most efficient estimators...
TABLE 1. Validity and Efficiency of Rosner, Spiegelman, and Willett Estimators for Continuous Outcomes With Linear Models

| DAG | $V_j$ as in$^*$ | $\hat{\beta}(OM)$ | $\hat{\beta}_{(-\gamma)}$ | $\hat{\beta}_{(-M)}$ | $\hat{\beta}_{(O-)}$ |
|-----|-----------------|---------------------|--------------------------|----------------------|----------------------|
| DAG 1, $V_1(-\gamma)$ | Valid (efficient) | Valid | Valid$^b$ | Valid$^b$ (efficient) |
| DAG 2, $V_2(-Z\gamma)$ | Valid | Biased | Biased | Biased |
| DAG 3, $V_3(-X\gamma)$ | Valid | Biased | Biased | Biased |
| DAG 4, $V_4(X\gamma Z)$ | Valid | Biased | Biased | Biased |
| DAG 5, $V_5(-\gamma)$ | Valid | Biased | Biased | Biased |
| DAG 6, $V_6(-Z\gamma)$ | Valid (efficient) | Valid | Valid$^d$ | Valid$^d$ |
| DAG 7, $V_7(-X\gamma)$ | Valid | Valid (efficient) | Valid$^d$ | Valid$^d$ |
| DAG 8, $V_8(X\gamma Z)$ | Valid$^e$ | Valid$^e$ | Valid$^e$ | Valid$^e$ |

$^a$The subscript such as $(\gamma \gamma)$ emphasizes how the given covariate relate to $X$, $Z$, and $Y$. For example, DAG 2 describes a situation where covariate $V_2(-Z\gamma)$ systematically affects measurement error and is a risk factor for the outcome.

$^b$The estimator is valid due to independence $Y \perp Z | V$ implied by the DAG.

$^c$The estimator is valid due to independence $X \perp Z | V$ implied by the DAG.

$^d$Relative efficiency depends on strength and direction of $\rho_{\epsilon_X \epsilon_Z}, \rho_{\epsilon_Z \epsilon_Y}$, and $\rho_{\epsilon_Y \epsilon_X}$, and the sample size of main study and validation study.

indicated. Taking DAG 7 as an example, our result suggests that while it would be valid to adjust for $V_7(X,-\gamma)$ in either both the outcome model and measurement error model (i.e., $\hat{\beta}_{(-\gamma)}$) or neither of the two models (i.e., $\hat{\beta}_{(O-)}$), which is, in this case, the most efficient option, it is invalid to adjust for this type of variable in either measurement error model alone (i.e., $\hat{\beta}_{(-M)}$) or outcome model alone (i.e., $\hat{\beta}_{(O-)}$).

Our result suggests that $V_2(-Z\gamma), V_3(X,Y), V_4(X\gamma Z)$ are the minimal covariate adjustment set and need to be included in both measurement error and outcome models. With the exception of $V_1(-Y)$ and trivially $V_8(X\gamma Z)$, including any other covariates in either the measurement error model alone or outcome model alone would result in bias. Including $V_1(-Y)$ in either outcome model alone or both the measurement error model and outcome model would increase efficiency, so does including $V_6(-Z\gamma)$ in both measurement error and outcome models. Under DAG 8, the relative efficiency of estimators $\hat{\beta}_{(OM)}$ and $\hat{\beta}_{(-\gamma)}$ depends on the strength and direction of the correlation of $V - X(\rho_{\epsilon_X})$, of $X - Z$ conditional on $V(X_{Z\gamma})$, of $V$ and the measurement error (i.e., $\rho_{\epsilon_X X_{Z\gamma}}$) and of $X - Y$ conditional on $V(\rho_{X_{Z\gamma}Y})$. To evaluate to what extent the relative efficiency is determined by $\rho_{\epsilon_X X_{Z\gamma}}$ and $\rho_{\epsilon_X X_{Z\gamma}}$, we plot the analytical asymptotic relative efficiency (ARE) $\hat{\beta}_{(-\gamma)} = \frac{\text{Var}(\beta_{(-\gamma)})}{\text{Var}(\beta_{(OM)})}$ with varying values of the three (conditional) correlations while fixing exposure effect and main study and validation study sample sizes (eFigure 1 of the eAppendix; http://links.lww.com/EDE/C120). We find that to gain efficiency by adjusting $V_8(X\gamma Z)$ in both measurement error and outcome models, $V_8(X\gamma Z)$ needs to be very weakly associated with true/gold-standard exposure and strongly associated with measurement error, a rare scenario in realistic settings. In eFigure 2 of the eAppendix; http://links.lww.com/EDE/C120, we also plot the ARE under DAG 1 where we vary the value of $V - Y$ correlation condition on $X$ (i.e., $\rho_{\epsilon_X X_{Z\gamma}}$) and of $\rho_{\epsilon_X X_{Z\gamma}}$ while fixing the sample sizes. The result suggests that we gain more efficiency by adjusting for $V_1(-Y)$ as $\rho_{\epsilon_X X_{Z\gamma}}$ and $\rho_{\epsilon_X X_{Z\gamma}}$ become greater.

Approximation in Generalized Linear Models With Logistic Link

For binary outcomes modeled by logistic regression, the estimators will approximate the log odds ratio free of exposure measurement error with little bias (e.g., percent bias less than 12%) if one of the following conditions is met: (1) $\text{Var}(X|Z, V)\hat{\beta}_{(-\gamma)}$ is small (e.g., less than 0.5), where $\text{Var}(X|Z, V)$ can be estimated as the mean squared error from measurement error model and $\beta$ is the causal parameter of interest$^{26,27}$ or (2) the disease is rare (e.g., less than 5%) and the measurement error in the measurement error model is homoskedastic.$^8$

We note that the analytical efficiency results for continuous outcomes under linear models do not necessarily hold for binary outcomes modeled by logistic regression. Specifically, under DAG 1, Neuhaus et al.$^{35}$ showed that the coefficient estimate for $Z$ (i.e., $\hat{\gamma}_1$) when regressing a binary $Y$ on $Z$ with logistic link not conditioning on $V_1(-Y)$ is always closer to null than the $V_1(-Y)$-adjusted coefficient estimate for $Z$ (i.e., $\hat{\gamma}_1$), a phenomenon known as noncollapsibility of the odds ratio.$^{39,40}$ Thus, under DAG 1, the efficiency of Rosner, Spiegelman, and Willett estimators, which depends on $\hat{\gamma}_1$ and $\hat{\gamma}_1$, do not necessarily follow the linear regression results.

**SIMULATION**

We performed a series of Monte Carlo simulation experiments to empirically evaluate the finite sample validity and efficiency of the estimators for both continuous and binary outcomes, with varying model parameterization and sample sizes under each DAG.

**Simulation Study Design**

All data were generated using the following data-distributing process with 1,000 samples:

1. $X = \eta_{\gamma} V + \epsilon_{\epsilon_x} \sim N(0,1)$
2. $Z = \theta_{\gamma} X + \epsilon_{\gamma} \sim N(0,0.5)$
3. $Y = \beta_{\gamma} X + \beta_{V} V + \epsilon_{\gamma} \sim N(0,1)$ for continuous outcomes $Y$, and
4. $Y - \text{Bern}(p)$, where $\log(p / (1 - p)) = -5 + \beta_{x} X + \beta_{v} V$ for binary outcome $Y$. 

© 2024 Wolters Kluwer Health, Inc. All rights reserved.
For the scenarios with binary outcomes, all simulations satisfy the condition that \( \text{Var}(X|Z, V) \beta_S^2 < 0.5^{26} \). As the base case for the eight simulation experiments corresponding to the eight DAGs, we generated samples of size \( n=5,000 \) for continuous outcome \( Y \) and 10,000 for binary \( Y \), where a subset of size \( n_{VS} = 400 \) is randomly sampled as external validation data with \( (V, X, Z) \) retained. The remaining \( n_{MS} = 4,600 \) or 9,600 was kept as main study data where only \( (V, Z) \) is observed. Smaller sample sizes were not used (e.g., 2,000 for binary outcome) because modern epidemiologic cohorts that use the regression calibration method often have a sample size greater than 10,000\(^{31–33} \) and the asymptotic statistical properties may not be evident when rare diseases are studied under smaller sample sizes. Covariate \( V \) is distributed as \( N(0,1) \) and the causal effect of \( X \) on \( Y \) is set to be \( \beta_s = 0.5 \), additive scale for continuous outcome and logit scale for binary outcome (equivalent to an odds ratio of approximately 1.65). For other parameters, we used all possible combinations of the following parameter values: \( \eta_k \in (0.4,0), \theta_k \in (0.5, \theta_k \in (0.1,0), \beta_k \in (0.8,0) \), which corresponds to each of the scenarios in the \( k^2 = 8 \) DAGs in Figure 1, with zero value representing the removal of a specific arrow on the DAG. Details for the simulation scenarios are available in Section 4 and eTable 1 of the eAppendix; http://links.lww.com/EDE/C120.

### Results

The simulation results under the base case for the point estimates are reported in Table 2 and for empirical variances of the valid estimators are reported in Table 3. For the point estimates, we present percent biases. For efficiency comparison of the valid estimators, we reported both analytical and empirical relative efficiency (AREN and ERE, respectively) under continuous outcome, calculated as the analytical or empirical variance

| TABLE 2. Percent Bias for Point Estimates in the Simulation Study under Base Case |
|---------------------------------------------------------------|
| \( V_j \) as in \( \hat{\beta} \) & \( \hat{\beta}_{OM} \) & \( \hat{\beta}_{(-)} \) & \( \hat{\beta}_{(-M)} \) & \( \hat{\beta}_{(O-)} \) |
| Continuous Outcome & | & | |
| \( V_1,(-\cdot\cdot\cdot) \) & 0 & 0 & 0 & 0 & 1 & 1 & 0 |
| \( V_2,(-\cdot\cdot\cdot) \) & 0 & 32 & 30 & 2 & 0 & 31 & 28 & 2 |
| \( V_3(X,-\cdot\cdot\cdot) \) & 0 & 55 & 67 & -7 & 0 & 52 & 64 & -7 |
| \( V_4(XZ,-\cdot\cdot\cdot) \) & 0 & 78 & 87 & -5 & 0 & 75 & 84 & -5 |
| \( V_5,(-\cdot\cdot\cdot) \) & 0 & 0 & 0 & 1 & 1 & 1 & 1 |
| \( V_6(Z,\cdot\cdot\cdot) \) & 0 & 0 & -2 & 2 & 1 & 1 & 1 & 3 |
| \( V_7(X,-\cdot\cdot\cdot) \) & 0 & 0 & 5 & -5 & 1 & 2 & 7 & -4 |
| \( V_8(XZ,-\cdot\cdot\cdot) \) & 0 & 0 & 5 & -5 & 1 & 2 & 7 & -4 |

*\( ^{c} \)Satisfies small measurement error condition \( \text{Var}(X|Z, V) \beta_S^2 < 0.5 \).

| TABLE 3. Empirical Asymptotic Relative Efficiency (ERE) and Variance in the Base Case Simulation Study When More Than One Estimator is Valid |
|---------------------------------------------------------------|
| \( V_j \) as in \( \hat{\beta} \) & \( \hat{\beta}_{OM} \) & \( \hat{\beta}_{(-)} \) & \( \hat{\beta}_{(-M)} \) & \( \hat{\beta}_{(O-)} \) |
| Outcome Type & | & | |
| Continuous & \( V_1,(-\cdot\cdot\cdot) \) & 1 (1.14) & 0.79 (1.44) & 0.79 (1.44) & 1.00 (1.14) |
| & \( V_2,(-\cdot\cdot\cdot) \) & 1 (1.14) & 1.00 (1.14) & 1.00 (1.14) & 1.00 (1.14) |
| & \( V_3(-Z,-\cdot\cdot\cdot) \) & 1 (1.14) & 0.97 (1.18) & 1.00 (1.14) & 1.00 (1.14) |
| & \( V_4(X,-\cdot\cdot\cdot) \) & 1 (1.14) & 1.19 (0.96) & 1.00 (1.14) & 1.00 (1.14) |
| & \( V_5(XZ,-\cdot\cdot\cdot) \) & 1 (1.14) & 1.28 (0.89) & 1.00 (1.14) & 1.00 (1.14) |
| Binary & \( V_1,(-\cdot\cdot\cdot) \) & 1 (2.11) & 1.01 (2.09) & 1.01 (2.09) & 1 (2.11) |
| & \( V_2,(-\cdot\cdot\cdot) \) & 1 (2.90) & 1.00 (2.91) & 1.00 (2.91) & 1 (2.91) |
| & \( V_3(-Z,-\cdot\cdot\cdot) \) & 1 (2.90) & 0.97 (2.99) & 1.00 (2.91) & 1.00 (2.91) |
| & \( V_4(X,-\cdot\cdot\cdot) \) & 1 (2.84) & 1.25 (2.28) & 1.00 (2.91) & 1.00 (2.91) |
| & \( V_5(XZ,-\cdot\cdot\cdot) \) & 1 (2.84) & 1.31 (2.17) & 1.00 (2.91) & 1.00 (2.91) |

*\( ^{c} \)Calculated as the empirical variance of \( \hat{\beta}_{OM} \) over the empirical variance of each estimator \( \hat{\beta}_j \) under consideration, that is ERE(\( \hat{\beta}_j \)) = \( \frac{\text{Var}(\hat{\beta}_{OM})}{\text{Var}(\hat{\beta}_j)} \), where \( \hat{\beta}_j \) could be any of the four estimators \( \hat{\beta}_{OM}, \hat{\beta}_{(-)}, \hat{\beta}_{(-M)} \) or \( \hat{\beta}_{(O-)} \).

*\( ^{d} \)Calculated as the variance of the point estimates \( \hat{\beta}_j \) over simulation replicates for each estimator. \( \times 10^{-3} \) for continuous outcome and \( \times 10^{-3} \) for binary outcome.

*\( ^{e} \)Satisfies small measurement error condition \( \text{Var}(X|Z, V) \beta_S^2 < 0.5 \).

324 | www.epidem.com © 2024 Wolters Kluwer Health, Inc. All rights reserved.
of $\hat{\beta}_{(OM)}$ over the analytical or empirical variance of other valid estimators, and analytical and empirical variance for the valid point estimates obtained over simulation replicates. For binary outcomes, only empirical results are presented. Analytical and simulation results for all scenarios can be found in eTables 2–5 of the eAppendix; http://links.lww.com/EDE/C120, which is briefly summarized in the remainder of this section.

For continuous outcomes, all simulation results are consistent with analytical expectations. For binary outcomes with the logistic outcome model and the linear measurement error model, the percent bias is less than 6% across all scenarios except one (where the percent bias is 12%). However, we noticed that adjusting for $V_{1(-Y)}$ in either outcome model only or in both models resulted in similar variances as when adjusting for $V_{1(-Y)}$ in either $\beta_{(-M)}$ or $\beta_{(--)}$ (eTable 5 of the eAppendix; http://links.lww.com/EDE/C120). This could be due to that the odds ratios are noncollapsible and that in finite samples, the efficiency gain from including the covariate cannot compensate for the loss of efficiency due to the estimation of the additional parameters.

For covariates $V_{2(-ZY)}$, and $V_{4(-ZY)}$, adjusting for them in the outcome model only results in much less biased estimates (e.g., percent bias < 10% in most cases) than would be obtained if one does not adjust for them in any model (e.g., percent bias ≥ 30% in most cases) provided that the correlation between these covariate(s) and measurement error is small (e.g., $\rho_{(x|z)} \leq 0.2$).

**REAL-DATA EXAMPLE**

**Methods**

We applied the proposed covariate selection framework to the study of the effect of total daily fiber intake on the risk of cardiovascular event, defined as either myocardial infarction or angina, among the HPFS. HPFS is an ongoing prospective study of 51,529 US male health professionals 40 to 75 years of age at enrollment in 1986.33,34 The study protocol was approved by the Institutional Review Boards of the Brigham and Women’s Hospital and the Harvard T.H. Chan School of Public Health. The surrogate exposure, fiber intake (g/day), was measured using a food frequency questionnaire every 4 years since 1986.33 A validation study of size 651 using dietary records was conducted in 2012 among the cohort participants. The participants were followed up from 1990 to 2016.17,35 For this example, we created a cohort consisting of participants who completed the sleep duration question in 1987 and the sunscreen use question in 1992. We excluded participants who died, reported a CVD event, or had any cancer diagnosis other than melanoma before 1990, when follow-up started. The final analysis includes 22,379 participants, among whom 409 participants were in the validation study and the remaining (n = 21,970) were the main study participants.

We included the following covariates in our analysis, with their corresponding covariate set indicated in parentheses: family history of CVD ($V_{1(-Y)}$), energy intake, baseline hypertension, diabetes and hypercholesterolemia, smoking, marital status ($V_{3(X-Z)}$), age, physical exercise/activity,14,35 body mass index (BMI), use of multivitamin supplements,35,36 alcohol intake,37 depression status,15,38 sleep duration16,39($V_{4(X|Z)}$), and frequency of sunscreen use ($V_{5(X-Z)}$). We encoded these assumptions into the DAG in Figure 2. We allow correlation between covariate sets through the unmeasured common cause $U$. For example, physical exercise and fiber intake could be correlated due to an unmeasured common cause related to a healthy lifestyle.

**FIGURE 2.** Proposed directed acyclic graph for the study of effect of fiber intake on cardiovascular disease incidence in Health Professional Follow-up Study. BMI indicates body mass index; CVD, cardiovascular disease; DR, dietary records; FFQ, food frequency questionnaire.
Note that sunscreen use cannot reasonably be assumed to directly cause changes in fiber intake. Therefore, we show the backdoor path through the unmeasured \( U \) (e.g., healthy lifestyle), connecting \( V_3(x, y) \) and \( X \). In this way, sunscreen use has the essential feature that it is not a direct cause of cardiovascular incidence and that it is conditionally independent of measurement error. See Section 7 of the eAppendix; http://links.lww.com/EDE/C120 for more details.

In accordance with the theoretical results, we adjusted for covariate sets \( V_3(x, y) \) and \( V_4(x, y) \), the minimal covariate adjustment set in the absence of \( V_2(z, y) \) in both the outcome and measurement error models (\( \beta_{OM} \)). We also adjusted for \( V_1(\cdot, \cdot) \) in the outcome model only (\( \beta_{OM} \)) and \( V_7(x, \cdot) \) in both outcome and measurement error models (\( \beta_{OM} \)) in addition to including \( V_3(x, y) \), \( V_4(x, y) \) in both models. We additionally evaluated the bias induced by omitting \( V_3(x, y) \) or \( V_4(x, y) \) from the measurement error model alone or from both models.

Results

We summarize the effect of 10 g in daily fiber intake on incidence of CVD in Table 4. Over the 16 years of follow-up, 2,090 (9%) cardiovascular events were reported. The standard analysis regressing CVD event on total fiber adjusting for \( V_3(x, y) \) and \( V_4(x, y) \) produces an attenuated odds ratio estimate of 0.93 associated with 10 g increase in daily total fiber intake. Adjusting for this set of covariates in both measurement error and outcome models gives an adjusted odds ratio of 0.83, so do all other valid estimators. Consistent with theory, we lose efficiency by additionally adjusting for \( V_7(x, \cdot) \) in both outcome and measurement error models while the point estimate remained similar, increasing the standard error of the parameter of interest from \( 9.46 \times 10^{-2} \) when adjusting only for \( V_3(x, y), V_4(x, y) \) to \( 9.54 \times 10^{-2} \). Omitting \( V_4(x, y) \) from both models produced the most biased point estimate of 1.03. As long as \( V_3(x, y) \) or \( V_4(x, y) \) are adjusted for in the outcome model, the bias resulting from omitting these variables from the measurement error model is small. This result is consistent with those from our simulation as the empirical partial correlation between each of \( V_3(x, y), V_4(x, y) \) and \( Z \) (conditional on \( X \) and other covariates) are all less than 0.17.

DISCUSSION

Summary of Main Results

First and foremost, we showed that \( V = (V_2(z, y), V_3(x, y), V_4(x, y)) \), that is, any common causes of (1) true/gold-standard exposure and outcome and (2) measurement error and outcome, are the minimal covariate set and should be adjusted for in both outcome and measurement error models. When designing a study that includes exposure measurement error adjustment, investigators should ensure the collection of \( V \) in both the main and validation studies. One may find it counter-intuitive that confounders not contributing to measurement error (i.e., \( V_3(x, y) \)) must be included in the measurement error model, with the rationale that under DAG \( \text{OM} \), \( Z \) and \( V_1(\cdot, \cdot) \) or \( V_7(x, \cdot) \) in both models, we induce an association between \( X \) and \( V_2(z, y) \), making \( V_2(z, y) \) de facto confounders for the \( X - Y \) relationship.

### TABLE 4. Effect Estimates of 10 g/day Increase in Fiber Intake on CVD Risk in Health Professional Follow-up Study

| Adjusted Sets (Models) | Unadjusted Analysis | Adjusted Analysis |
|------------------------|---------------------|------------------|
|                        | Unadjusted OR | 95% CI | Adjusted OR | 95% CI | SE (10^-2) |
| Theoretically valid estimators | | | | | |
| \( V_3(x, y), V_4(x, y) \) (OM) | 0.93 | (0.86, 1.00) | 0.83 | (0.69, 1.00) | 9.46 |
| \( V_1(\cdot, \cdot), V_3(x, y), V_4(x, y) \) (OM) | 0.93 | (0.86, 1.00) | 0.83 | (0.69, 1.00) | 9.45 |
| \( V_1(\cdot, \cdot), V_7(x, \cdot) \) (OM) | 0.93 | (0.86, 1.00) | 0.83 | (0.69, 1.00) | 9.44 |
| \( V_1(\cdot, \cdot), V_3(x, y), V_4(x, y) \) (OM) | 0.93 | (0.87, 1.00) | 0.83 | (0.69, 1.01) | 9.54 |
| Theoretically biased estimators | | | | | |
| \( V_3(x, y) \) (OM) | 1.01 | (0.86, 1.00) | 1.03 | (0.89, 1.20) | 7.68 |
| \( V_4(x, y) \) (OM) | 0.92 | (0.86, 1.00) | 0.81 | (0.67, 0.97) | 9.40 |
| \( V_4(x, y) \) (OM) | 0.93 | (0.86, 1.00) | 0.85 | (0.71, 1.00) | 8.54 |
| \( V_4(x, y) \) (OM) | 0.93 | (0.86, 1.00) | 0.83 | (0.69, 1.00) | 9.38 |

\( \text{OM} \), \( \text{OM} \), and \( \text{OM} \) indicate whether the covariates are adjusted in both outcome and measurement error models, measurement error model only or outcome model only. For example, \( V_1(\cdot, \cdot) \) (OM) indicates that \( V_1(\cdot, \cdot) \) is adjusted in MEM only and \( V_3(x, y), V_4(x, y) \) are adjusted in both MEM and outcome model.

CI indicates confidence interval; CVD, cardiovascular disease; OR, odds ratio; SE, standard error.

© 2024 Wolters Kluwer Health, Inc. All rights reserved.
Second, our results suggest that the Rosner, Spiegelman, and Willett regression calibration method can be relaxed to allow for the adjustment of risk factors that are conditionally independent of both true/gold-standard exposure and measurement error in the outcome model only. This provides an opportunity to gain statistical efficiency when such variables are only available in the main study. Another opportunity for increasing efficiency would be to adjust for non-risk factors that are strong determinants of measurement error but have no association with true/gold-standard exposure (e.g., \( V_{6(-z-)} \)). In short, to optimize efficiency with validity assured in expectation, one should adjust for variables of type \( V_{1(-y-)} \) through \( V_{6(-z-)} \) in both measurement error and outcome models.

Last, we recommend that researchers leave out any non-risk factors that are even weakly associated with the true/gold-standard exposure, as adjustment for these covariates in both outcome and measurement error models almost always results in efficiency loss.

**Main Study/Internal Validation Study Design**

For internal validation studies, the true/gold-standard exposure \( X \) is partially observed in the main study. Spiegelman et al.\(^40\) proposed an inverse-variance-weighted estimator that combines the adjusted estimate using an external validation study with the additional regression coefficient estimate obtained from regressing outcome on the exposure in the main study, with the inverse-variance-weighted estimator being asymptotically nearly unbiased and efficient as the optimal maximum-likelihood estimator.

**Effect Modification by Covariate \( V \)**

The identification formula (9) allows for effect modification by \( V \) and can be estimated nonparametrically or parametrically. However, parametric estimation of \( \beta(V) \) requires special attention. For example, we demonstrate in Section 8.1 of the eAppendix; http://links.lww.com/EDE/C120 that under an additive-scale \( X-V \) interaction, one would need to add a \( Z-V \) product term and a \( V^2 \) term in the outcome model, that is, \( E[Y|Z, V] = \beta_0 + \beta_1 Z + \beta_2 V + \beta_3 ZV + \beta_4 V^2 \). Together with MEM in (1), this gives a conditional ATE estimator \( \hat{\beta}(V) = \frac{\hat{\beta}_3 + \hat{\beta}_4 V}{\hat{\alpha}_1} \).

More generally, we can extend the parametric estimators in the Review section to the following nonparametric estimators:

\[
\hat{\beta}_{(OM)}(V) = \frac{\hat{E}[Y|Z = z, V] - \hat{E}[Y|Z = z', V]}{\hat{E}[X|Z = z, V] - \hat{E}[X|Z = z', V]} \tag{13}
\]

\[
\hat{\beta}_{(-M)}(V) = \frac{\hat{E}[Y|Z = z] - \hat{E}[Y|Z = z']}{\hat{E}[X|Z = z, V] - \hat{E}[X|Z = z', V]} \tag{14}
\]

\[
\hat{\beta}_{(O-)}(V) = \frac{\hat{E}[Y|Z = z, V] - \hat{E}[Y|Z = z', V]}{\hat{E}[X|Z = z] - \hat{E}[X|Z = z']} \tag{15}
\]

We also show in Section 8.2 of the eAppendix; http://links.lww.com/EDE/C120 that the validity result in Table 1 still apply in all cases except that \( \hat{\beta}_{(-M)}(V) \) can no longer estimate conditional ATE when \( V_{1(-y-)} \) is an effect modifier. The evaluation of (relative) efficiency for the above estimators is beyond the scope of this study.

**Other Measurement Error Structures**

We are aware that the measurement error structures in Figure 1 do not cover all research contexts. For example, instead of being a confounder, a variable such as a biomarker or BMI may be a mediator between dietary intake and outcome, with BMI potentially inducing exposure measurement error, as depicted in DAGs I and II Figure 3A, B, respectively. Under these scenarios, it is unclear what meaningful causal estimation, if any, the Rosner, Spiegelman, and Willett estimators correspond to. We will discuss these measurement error structures in new papers.

Finally, we have devoted a separate manuscript to the covariate selection issue using Carroll, Rupert, and Stefanski regression calibration method, as it does not always require the same covariate adjustment rules as the Rosner, Spiegelman, and Willett regression calibration method.

**ACKNOWLEDGMENTS**

We thank Jessica Young, James Robins, and Christopher Boyer (Harvard University) for helpful comments on different versions of this article and its supplemental material.
REFERENCES

1. Rimm EB, Giovannucci EL, Stampfer MJ, Colditz GA, Litin LB, Willett WC. Reproducibility and validity of an expanded self-administered semi-quantitative food frequency questionnaire amongst male health professionals. Am J Epidemiol. 1992;135:1114–26; discussion 1127.

2. Giovannucci E, Liu Y, Rimm EB, et al. Prospective study of predictors of vitamin D status and cancer incidence and mortality in men. J Natl Cancer Inst. 2006;98:451–459.

3. Tanasescu M, Leitzmann MF, Rimm EB, Willett WC, Stampfer MJ, Hu FB. Exercise type and intensity in relation to coronary heart disease in men. JAMA. 2002;288:1994–2000.

4. Kaestner R, Kausal N. Effect of immigrant nurses on labor market outcomes of us nurses. J Urban Econ. 2012;71:219–229.

5. Bash LD, Erlinger TP, Coresh J, Marsh-Manzi J, Folsom AR, Astor BC. Inflammation, hemostasis, and the risk of kidney function decline in the atherosclerosis risk in communities (ARIC) study. Am J Kidney Dis. 2009;53:596–605.

6. Célinea V, Philippat C, Agier L. An empirical validation of the within-subject biospecimens pooling approach to minimize exposure misclassification in biomarker-based studies. Epidemiology. 2019;30:756–767.

7. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med. 1989;8:1051–69; discussion 1071.

8. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol. 1990;132:734–745.

9. Liao X, Zucker DM, Yi Li, Spiegelman D. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. Biometrics. 2011;67:50–58.

10. White IR. Concomitant-marginal: dealing with measurement error; multiple imputation or regression calibration?. Int J Epidemiol. 2006;35:1081–1082.

11. Stefanski L, Carroll R, Ruppert D, Crainiceanu C. Measurement Error in Nonlinear Models: A Modern Perspective. 2nd ed. Chapman and Hall/CRC; 2020.

12. Hardin JW, Schmiediche H, Carroll RJ. The regression-calibration method for fitting generalized linear models with additive measurement error. Stata J. 2003;3:361–372.

13. Ambrose JA, Barua RS. The pathophysiology of cigarette smoking and cardiovascular disease: an update. J Am Coll Cardiol. 2004;43:1731–1737.

14. Mandolesi L, Polverino A, Montuori S, et al. Effects of physical exercise on cognitive functioning and wellbeing: biological and psychological benefits. Front Psychol. 2018;9:509.

15. McDermott LM, Ebmeier KP. A meta-analysis of depression severity and cognitive function. J Affect Disord. 2009;119:1–8.

16. Ma Y, Liang L, Zheng F, Shi L, Zhong B, Xie W. Association between sleep duration and cognitive decline. JAMA Netw Open. 2020;3:e201357 3–e201357.

17. Cahill LE, Chiue SE, Mekary RA, et al. Prospective study of breakfast eating and incident coronary heart disease in a cohort of male us health professionals. Circulation. 2013;128:337–343.

18. Logan R, Spiegelman D. Sas: The sas %blinplus macro. 2012. Available at: https://yssp. yale. edu/cmips/research/software/blinplus/. Accessed 30 September 2022.

19. VänderWeele TJ, Hernán MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. Am J Epidemiol. 2012;175:1303–1310.

20. Weinberg CA, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend?. Am J Epidemiol. 1994;140:565–571.

21. Hernán MA, Cole SR. Invited commentary: causal diagrams and measurement bias. Am J Epidemiol. 2009;170:959–62; discussion 963.

22. Weiskopf MG, Webster TF. Trade-offs of personal vs. more proxy exposure measures in environmental epidemiology. Epidemiology. 2017;28:635–643.

23. Zivich PN, Shook-Sa BE, Edwards JK, Westreich D, Cole SR. On the use of covariate supersets for identification conditions. Epidemiology. 2022;33:559–562.

24. Hernán M, Robins J. Causal Inference: What If?. 1st ed. Chapman and Hall/CRC; 2020.

25. Richardson TS, Robins J. Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. University of Washington; 2013.

26. Kalu J. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. Stat Med. 1994;13:1135–1148.

27. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. J Am Stat Assoc. 2000;95:51–61.

28. Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika. 1993;80:807–815.

29. Sjölander A, Dahlqwist E, Zetterqvist J. A note on the noncollapsibility of rate differences and rate ratios. Epidemiology. 2016;27:356–359.

30. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev/Revue Internationale de Statistique. 1991;59:227–240.

31. Bao Y, Bertoia ML, Lenart EB, et al. Origin, methods, and evolution of the three nurses’ health studies. Am J Public Health. 2016;106:1573–1581.

32. Design of the women’s health initiative clinical trial and observational study. The Women’s Health Initiative Study Group. Control Clin Trials. 1998;19:61–109.

33. Gu X, Wang DD, Fung TT, et al. Dietary quality and risk of heart failure in men. Am J Clin Nutr. 2022;116:378–385.

34. Kim HJ, Giovannucci E, Rosner B, Willett WC, Cho E. Longitudinal and secular trends in dietary supplement use: nurses’ health study and health professionals follow-up study, 1986-2006. J Acad Nutr Diet. 2014;114:436–443.

35. Liu S, Buring JE, Sesso HD, Rimm EB, Willett WC, Manson JE. A prospective study of dietary fiber intake and risk of cardiovascular disease among women. Am Coll Cardiol. 2002;39:49–56.

36. Grima NA, Pase MP, Macpherson H, Pipangas A. The effects of multivitamins on cognitive performance: a systematic review and meta-analysis. J Alzheimers Dis. 2012;29:561–569.

37. Brennan SE, McDonald S, Page MJ, et al. Long-term effects of alcohol consumption on cognitive function: a systematic review and dose-response analysis of evidence published between 2007 and 2018. Syst Rev. 2020;9:1–39.

38. Van der Kooy K, Van Hout H, Marwijk H, Marten H, Stohouwer C, Beekman A. Depression and the risk for cardiovascular diseases: systematic review and meta-analysis. Int J Geriatr Psychiatry. 2007;22:613–626.

39. Hoevenaar-Blom MP, Spijkerman AMW, Kromhout D, van den Berg FB. Exercise type and intensity in relation to coronary heart disease in men. Am J Epidemiol. 2002;155:1994–2000.

40. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Stat Med. 2001;20:139–160.