Reducing Bias Due to Exposure Measurement Error Using Disease Risk Scores

David B. Richardson*, **Alexander P. Keil, Stephen R. Cole, and Jessie K. Edwards**

* Correspondence to Dr. David B. Richardson, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (e-mail: david.richardson@unc.edu).

Initially submitted January 27, 2020; accepted for publication September 23, 2020.

Suppose that an investigator wants to estimate an association between a continuous exposure variable and an outcome, adjusting for a set of confounders. If the exposure variable suffers classical measurement error, in which the measured exposures are distributed with independent error around the true exposure, then an estimate of the covariate-adjusted exposure-outcome association may be biased. We propose an approach to estimate a marginal exposure-outcome association in the setting of classical exposure measurement error using a disease score–based approach to standardization to the exposed sample. First, we show that the proposed marginal estimate of the exposure-outcome association will suffer less bias due to classical measurement error than the covariate-conditional estimate of association when the covariates are predictors of exposure. Second, we show that if an exposure validation study is available with which to assess exposure measurement error, then the proposed marginal estimate of the exposure-outcome association can be corrected for measurement error more efficiently than the covariate-conditional estimate of association. We illustrate both of these points using simulations and an empirical example using data from the Orinda Longitudinal Study of Myopia (California, 1989–2001).

bias; cohort studies; epidemiologic methods; regression analysis

In observational epidemiologic studies, the exposure of interest is not under control of the investigator. Two challenges that typically arise in observational studies due to lack of control over exposure assignment are error in estimates of exposure and potential confounding of the exposure-outcome association. The first follows because we must estimate exposure rather than assign it, while the second follows because, unlike in an experimental design, the investigator cannot rely upon randomization to lead in expectation to balance between exposure groups in other factors that affect disease.

Interestingly, the way that we address the challenge of confounding can exacerbate, or alleviate, the consequences of exposure measurement error in epidemiologic analyses of an association between a continuous exposure variable and outcome. Settings where exposure measurement errors occur include, but are not limited to, measurements of environmental factors, such as ambient levels of an air pollutant; occupational factors, such as personal dosimetry measures of ionizing radiation; and nutritional factors, such as serum measures of Vitamin D (1–3). In fact, exposure measurement error is likely to be an important limitation of a large proportion of epidemiologic studies (4).

It is well known that if the exposure variable suffers classical measurement error, in which the measured exposures are distributed with independent error around the true exposure, then an estimate of the exposure-outcome association obtained in a linear regression may be attenuated (5). Perhaps less well known is that the degree of attenuation due to classical exposure measurement error tends to increase as confounding variables, which are by definition predictors of the exposure, are introduced into the linear regression model (1, p. 52; 6). Therefore, while covariate adjustment can potentially reduce bias due to confounding in an estimate of the exposure-outcome association, it may tend to increase attenuation due to classical exposure measurement error.

We propose a simple method to avoid the bias inflation that can occur as confounding variables are included in a multivariable regression model to obtain covariate-adjusted estimates of association between the error-prone exposure and outcome of interest. Assuming one can identify an unexposed reference group, our approach achieves this by shifting the covariates to a model for a disease score that is subsequently used to obtain a marginal estimate of association using a model-based standardization, which is a generalization of direct standardization to the exposed sample

(7). The proposed approach offers a useful method to reduce bias due to exposure measurement error in observational epidemiologic analyses. In addition, we show that if a validation study is available to assess the measurement error structure then the proposed marginal estimate of the exposure-outcome association can be corrected for measurement error more efficiently than the covariate-conditional estimate of association.

METHODS

We focus on the setting of an epidemiologic study with a continuous exposure variable, a set of well-measured confounders, and a continuous outcome of primary interest. We first provide theory and then provide simulations to demonstrate the proposed method. We also comment on extension to a regression model for a binary outcome variable. The motivating setting is one in which the exposure variable of primary interest is measured on a scale with origin at zero, and right-skewed, such that exposures are often near the origin. Let subscript i index subject, Y_i, denote the outcome of interest, T_i denote a nonnegative continuous exposure, and $\mathbf{Z}_i = \{Z_{i1}, \ldots, Z_{ik}\}$ denote the k covariates that are potential confounders of the associations between T_i and Y_i. Suppose that the expectation (denoted $E[\bullet]$) of the outcome of interest, Y_i, follows a linear model of the form,

$$E\left[Y_i | Z_i, T_i\right] = \varphi_0 + \sum\nolimits_{j=1}^{k} \varphi_j Z_{ij} + \alpha T_i. \qquad (1)$$

Typically, in an epidemiologic study, the true exposure, T_i, is not observed. What we observe is a surrogate exposure variable, X_i, that provides an imperfect measure of T_i. A simple case is a classical measurement error model in which the measured surrogate exposures are distributed with independent error around the true exposure, of the form

$$X_i = T_i + \eta_i, \text{where } \eta_i \sim N\left(0, \sigma^2_{X|T}\right).$$

Given covariates that are potential confounders of the association of interest, an investigator may fit a regression model that includes the surrogate exposure measure, X, as an explanatory variable, along with the measured covariates, $\mathbf{Z}$, of the form

$$E\left[Y_i | Z_i, X_i\right] = \phi_0 + \sum\nolimits_{j=1}^{k} \phi_j Z_{ij} + \beta X_i. \qquad (2)$$

Alternatively, standardization affords us a way to compare the mean of Y between groups defined by X, controlling for confounding by $\mathbf{Z}$. Specifically, assume that the investigator wants to estimate the exposure effect by comparing the exposed group mean of the outcome variable Y with the expected mean of counterfactual outcomes in a group with the same $\mathbf{Z}$ distribution as the exposed (8). This comparison may be summarized as $E(Y|X = x) - E(Y_0|\mathbf{Z} = z, X = x)$, where the potential outcome under the absence of exposure is denoted Y_0. Our proposed approach makes use of a disease score, F($\mathbf{Z}$), a function of $\mathbf{Z}$ that confers conditional independence between the potential outcome under the absence of exposure and $\mathbf{Z}$ (7, 8). Under our proposed approach, the score is estimated by fitting a regression model to empirical data for an unexposed reference group. We discuss several options for identifying an unexposed reference group in the Discussion. For the case of a single binary regressor variable, $Z = z$, the disease score could be estimated by fitting a linear regression model to the data for the unexposed reference group, $E\left[Y_i | Z_i, X_i = 0\right] = \theta_0 + \theta_1 Z_i$, and then setting $\hat{F}\left(z_i\right) = \hat{\theta}_0 + \hat{\theta}_1 z_i$, for all individuals i in the study sample. A standardized measure of the change in the expectation of Y given a unit change in X is quantified in the study sample as the difference between the mean Y conditional on exposure and the expected mean of the potential outcomes in the absence of exposure (8–13). This estimate of the standardized exposure-outcome association can be obtained by fitting a regression model to the study sample data (nota bene, not the reference group) that includes the estimated score, $\hat{F}\left(\mathbf{Z}\right)$, as an offset,

$$E\left[Y_i | \hat{F}\left(\mathbf{Z}_i\right), X_i\right] = \psi_0 + \hat{F}\left(\mathbf{Z} = z_i\right) + \gamma X_i \qquad (3)$$

Nonparametric bootstrap standard errors for the estimated coefficients can be obtained (Web Appendices 1 and 2, available at https://doi.org/10.1093/aje/kwaa208).

Note that under equation 1 the association between T and Y does not vary across levels of $\mathbf{Z}$, and an estimate of the parameter describing the effect of T on Y, $\hat{\alpha}$, is collapsible (i.e., in the absence of confounding by $\mathbf{Z}$, a crude estimate of α equals the $\mathbf{Z}$-conditional estimate of association). Given these conditions, we may readily compare the standardized estimate of association obtained by fitting model equation 3, $\hat{\gamma}$, with the covariate-conditional estimate of association that would be obtained by fitting model equation 1, $\hat{\alpha}$, and to the conditional estimate of association that would be obtained by fitting model equation 2, $\hat{\beta}$. Thus, we can establish the performance of each estimator relative to a known true exposure-response, if available, and to exposure-response estimates that can feasibly be estimated in a given data set.

Bias due to exposure measurement error

Given classical measurement error, we may expect attenuation bias in our estimate of the true exposure-outcome trend due to performing a regression on X, an error-prone version of T. The expected covariate-conditional estimate of association derived under equation 2 follows the expression

$$E\left(\beta\right) = \alpha \frac{\sigma^2_{T|\mathbf{Z}}}{\sigma^2_{X|\mathbf{Z}}}$$

where $\sigma^2_{T|\mathbf{Z}}$ denotes the $\mathbf{Z}$-conditional variance of T, and $\sigma^2_{X|\mathbf{Z}}$ denotes the $\mathbf{Z}$-conditional variance of X. Given a correctly estimated disease score, the expected covariate-standardized estimate of association derived under equation 3 follows the expression

$$E\left(\gamma\right) = \alpha\,\frac{\sigma_T^2}{\sigma_X^2}.$$

Therefore, the expected value for the estimated covariate-adjusted association between the surrogate measure and outcome, β, will tend to differ from the expected value for the covariate-standardized estimate of association, γ, and our sample estimate of $\hat{\gamma}$ will tend to be closer than $\hat{\beta}$ to the true exposure-disease estimate of association of interest, α. If X_i is a perfect proxy for T_i, such that $\frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{X|\mathbf{Z}}^2} = 1$ then the estimates, $\hat{\beta}$ and $\hat{\gamma}$, obtained using the surrogate measure X suffer no attenuation bias; if Z is independent of T, such that $\frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{X|\mathbf{Z}}^2}$, then the estimates $\hat{\beta}$ and $\hat{\gamma}$ will suffer the same degree of attenuation bias due to classical measurement error, and if $\alpha = 0$ then the estimates will tend to be unbiased, $\hat{\beta} = \hat{\gamma} = 0$. However, in the setting of interest, where the surrogate exposure variable, X_i, provides an imperfect measure of T_i, $\alpha \neq 0$, and covariates $\mathbf{Z}_i$ are confounders, attenuation bias will occur and it will be greater for the covariate-conditional estimate than the standardized estimate of association because $\frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{X|\mathbf{Z}}^2} < \frac{\sigma_T^2}{\sigma_X^2}$ when $\mathbf{Z}_i$ predict T_i (noting that under these conditions $\frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{X|\mathbf{Z}}^2} = \frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{T|\mathbf{Z}}^2+\sigma_{X|T}^2} < \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_{X|T}^2}$).

The arguments above are developed in the context of linear regression for a continuous outcome variable. Previous work suggests that similar qualitative conclusions regarding the impact of classical measurement error on estimated parameters can be drawn for logistic and log linear regression models for binary outcomes (14), and an approach for incorporation of a disease risk score in regression analyses that involve a binary outcome variable has been described previously (15). While we focus on a model of additive measurement errors of the form, $X_i = T_i + \eta_i$, where $\eta_i \sim N\left(0, \sigma_{X|T}^2\right)$, another measurement error model often discussed in environmental and occupational settings is a multiplicative measurement error model of the form $X_i = T_i \exp(\eta_i)$, or equivalently $\ln(X_i) = \ln(T_i) + \eta_i$. Notably, if the model for the true exposure-outcome association is linear in $\ln(T_i)$, and the investigator fits a regression model for $\ln(X_{\mathrm{i}})$, then the description we provide for attenuation holds because the errors are simply additive on a log scale (14).

Correction for measurement error

Suppose that we have a validation study in which we have measured X and T, along with covariates $\mathbf{Z}$. We may correct for attenuation bias in our estimate of the exposure-outcome using information from the validation study on the association between X and T. The measurement error–corrected association of the covariate-conditional estimate of association, $\hat{\beta}$, follows, $\hat{\beta}^{corr} = \hat{\beta}\,\frac{\hat{\sigma}_{X|\mathbf{Z}}^2}{\hat{\sigma}_{T|\mathbf{Z}}^2}$, where $\frac{\hat{\sigma}_{X|\mathbf{Z}}^2}{\hat{\sigma}_{T|\mathbf{Z}}^2}$ denotes the ratio of $\mathbf{Z}$-conditional variances of T and X. The measurement error–corrected standardized regression estimate of the association of interest $\hat{\gamma}$, follows the expression $\hat{\gamma}^{corr} = \hat{\gamma}\frac{\hat{\sigma}_X^2}{\hat{\sigma}_T^2}$. Using a validation study, we can derive a corrected estimate of the association of interest. Both measurement error–corrected estimates, $\hat{\beta}^{corr}$and $\hat{\gamma}^{corr}$, will tend to be unbiased estimates of the true exposure-disease estimate of association of interest; however, the approaches differ in statistical efficiency. Moreover, given a validation sample of fixed size, reliable estimates of the covariate-conditional variances, $\sigma_{X|Z}^2$ and $\sigma_{T|Z}^2$, are more difficult to obtain than reliable estimates of the marginal variances, σ_X^2 and σ_T^2, and these challenges increase as the dimensionality of the covariate vector $\mathbf{Z}$ increases.

The variance of the measurement error–corrected estimate of the covariate-adjusted association between the surrogate measure and outcome, $\hat{\beta}^{corr}$, will tend to be greater than the variance of the measurement error–corrected covariate-standardized estimate of association, $\hat{\gamma}^{corr}$. The relative advantage of the proposed standardized approach when compared with correction of bias in a covariate-conditional model will tend to increase as the covariate vector increases. Web Appendix 3 includes an expression for approximate variance estimation of the proposed measurement error–corrected covariate-standardized estimate.

Simulation example

We use simulated data to illustrate the performance of the proposed method. We simulated data for 1,000 studies with 4,500 people in the study sample and 500 people in an unexposed reference sample to illustrate a setting with moderate sample sizes, and we simulated 1,000 studies with 5,000 people in the study sample and 5,000 people in an unexposed reference sample to illustrate a setting with a large sample sizes. Each person was randomly assigned a covariate value Z by sampling from a normal distribution, $Z = N(0,1)$. In the study sample, true exposure, T, was generated by sampling from the distribution, $T = \log(1 + \exp(\omega_{T|Z}Z + N(0,1)))$, and a surrogate exposure, X, was assigned under a model of classical additive error, of the form $X = T+N\left(0, \sigma_{X|T}^2\right)$, such that surrogate measures could take values less than zero due to measurement error. In the reference sample, T and X were set to 0. The outcome variable Y was assigned by sampling from normal distribution under a model of the form $Y = 1 + 2Z + 1T + N(0,1)$. Simulations were conducted for scenarios where $\omega_{T|Z} = 0$, 1, 2, and for scenarios where $\sigma_{X|T}^2 = 0.2$, 0.5, 1, similar to the ranges of measurement errors that have been posited in simulations in a range of epidemiologic substantive areas (16–18).

For each simulated data set, we calculated $\frac{\sigma_T^2}{\sigma_X^2}$ and $\frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{X|\mathbf{Z}}^2}$ in the study sample. We fitted a regression model for Y conditional on T and Z to summarize the association under the data-generating model, and we fitted a model for Y

conditional on X and Z to summarize the biased estimate of association when fitting a model using the error-prone variable X rather than T. We then estimated the proposed standardized regression model, first estimating the disease score among those in the reference sample, and then incorporating the estimated score as a regression model offset in a model for Y as a function of X fitted to the study sample (as in equation 3), with a robust variance estimate. We summarized results from the simulated studies by computing the mean of the estimated X-Y association, the estimated standard deviation of the estimates (the empirical standard error, ESE), and the average estimated standard error of the estimate (ASE).

In addition, we simulated a small validation study of 100 people in which we observed X and T, along with covariates $\mathbf{Z}$. Using this validation study sample, we calculated a measurement error–corrected standardized estimate of association as $\hat{\gamma}\ \frac{\hat{\sigma}_X^2}{\hat{\sigma}_T^2}$. We also calculated $\frac{\sigma_{T|\mathbf{Z}}^2}{\sigma_{X|\mathbf{Z}}^2}$ and calculated a measurement error–corrected covariate-conditional estimate of association as $\hat{\beta}\frac{\hat{\sigma}_{X|\mathbf{Z}}^2}{\hat{\sigma}_{T|\mathbf{Z}}^2}$. We summarized corrected estimates from the simulated studies by computing the mean of the corrected estimates of the X-Y association and the estimated standard deviations of the mean estimates.

To illustrate the impact of classical measurement error on estimated parameters in regression models in which the outcome variable was a binary variable, in each simulated data set we also generated a random binary outcome, D, that took a value of 1 with probability, $p = \exp(1T) \times [\exp(-3.5+2Z)/(1+\exp(-3.5+2Z))]$. We fitted a general relative risk regression model for D conditional on X and Z to summarize the association when fitting a model using the error-prone variable X rather than T. We fitted a logistic model for D conditional on Z to estimate a disease risk score among those in the reference sample, and we estimated a standardized regression estimate by fitting a regression model for D as a function of X to the study sample with log link, including the natural log of the disease risk score as an offset (15).

Empirical example

We illustrate the proposed method in empirical data that were derived from the Orinda Longitudinal Study of Myopia (California, 1989–2001), a cohort study of ocular component development and risk factors for nearsightedness among children, including family history of myopia and the amount and type of visual activity that a child performed (19). The exposure of primary interest was self-reported hours per week reading for pleasure (READHR, in units of hours); and the outcome of interest was spherical equivalent refraction (SPHEQ, in units of diopter, a measure of the eye's effective focusing power). Covariates included age at study entry (AGE, in years), year of study entry (STUDYYEAR, in years), sex (GENDER, 1 = female, else 0), maternal history of myopia (MOMMY, 1 = yes, else 0), and paternal history of myopia (DADMY, 1 = yes, else 0). Here, we consider those who reported a complete absence of reading for pleasure (0 hours) as an accurate indication of the absence of exposure. Among those who reported reading for pleasure, we assumed that exposure estimates suffer error proportional to the true value, $\ln(X_i) = \ln(T_i) + \eta_i$. Using data for those 180 children who reported 0 hours per week reading for pleasure, we fitted a regression model for SPHEQ as a function of AGE, STUDYYEAR, GENDER, MOMMY, and DADMY, and we derived an estimated disease score as the predicted value of SPHEQ given the fitted model and observed covariates. Using data for those 438 children who reported 1 or more hours per week reading for pleasure, we estimated the diopter change per log-unit increase in hourly reading by fitting a regression model for SPHEQ as a function of ln(READHR), with the estimated disease risk score included as an offset term. We compared results estimated using the proposed approach with those estimated using a covariate-conditional regression model for SPHEQ as a function of ln(READHR), AGE, STUDYYEAR, GENDER, MOMMY, and DADMY in the study sample of those children with 1 or more hours per week reported reading for pleasure, and to a crude regression model for SPHEQ as a function of ln(READHR).

RESULTS

Simulation

Table 1 reports the simulation results for varying degrees of measurement error, $\sigma_{X|T}^2$ (0.2, 0.5, or 1.0), and for varying magnitudes of the association between covariate Z and T, $\omega_{T|Z}$. In all simulations, the average estimated parameter, $\hat{\alpha}$, was equal to 1. In all simulations, the average estimated parameter $\hat{\beta}$ corresponded to $\hat{\alpha}\left(\frac{\hat{\sigma}_{T|\mathbf{Z}}^2}{\hat{\sigma}_{X|\mathbf{Z}}^2}\right.$, such that the Z-conditional estimate of association obtained using the error-prone proxy X was attenuated relative to the simulation setup for the underlying association between T and Y conditional on Z. As the degree of exposure measurement error, $\sigma_{X|T}^2$, increased, the attenuation relative to the simulation setup increased in $\hat{\beta}$. In general, the ratio $\hat{\sigma}_{T|Z}^2/\hat{\sigma}_{X|Z}^2$ increased toward 1 as the variance of the measurement error, $\sigma_{X|T}^2$, decreased toward 0, and as the Z-conditional variance of T increased toward positive infinity. In all simulation scenarios, the empirical standard error of $\hat{\alpha}$ conformed to the average of the estimated standard errors, and the empirical standard error of $\hat{\beta}$ conformed to the average of the estimated standard errors (Web Table 1).

Proposed approach: continuous outcome variable

Under the simulation conditions examined, the proposed marginal estimate of the exposure-outcome association (model 3) suffered less bias due to classical measurement error than the covariate-conditional estimate of association (model 2) when the covariate Z was associated with true exposure T (i.e., $\omega_{T|Z} > 0$). When exposure was independent of covariates (i.e., $\omega_{T|Z} = 0$) the standardized estimate of the X-Y association tended to equal the Z-conditional estimate of the X-Y association ($\hat{\gamma} = \hat{\beta}$). The standardized and

Table 1. Results of Simulations of Associations Between a Continuous Covariate, Continuous Exposure, Mismeasured Continuous Surrogate Exposure, and Continuous Outcome

Simulation Setup			**Simulated Data Characteristics**			**Estimates Obtained From Fitted Models**		
$\sigma^2_{X\|T}$	$\omega_{T\|Z}$	$\hat{\sigma}^2_T$	$\hat{\sigma}^2_T/\hat{\sigma}^2_X$	$\hat{\sigma}^2_{T\|\mathbf{Z}}/\hat{\sigma}^2_{X\|\mathbf{Z}}$	**Model**	**Parameter**[a]	**Estimate**	**SE**[b]
0.2	2	1.51	0.88	0.72				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.02
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.72	0.02
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.88	0.01
	1	0.57	0.74	0.61				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.03
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.61	0.02
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.74	0.02
	0	0.27	0.58	0.58				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.03
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.58	0.02
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.58	0.02
0.5	2	1.51	0.75	0.50				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.02
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.50	0.02
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.75	0.01
	1	0.57	0.53	0.39				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.03
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.39	0.02
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.53	0.02
	0	0.27	0.35	0.35				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.03
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.35	0.02
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.35	0.02
1.0	2	1.51	0.60	0.34				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.02
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.34	0.01
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.60	0.01
	1	0.57	0.36	0.24				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.03
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.24	0.01
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.36	0.01
	0	0.27	0.21	0.21				
					$E[Y_i\|Z_i, T_i]$	$\hat{\alpha}$	1.00	0.03
					$E[Y_i\|Z_i, X_i]$	$\hat{\beta}$	0.21	0.01
					$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}$	0.21	0.01

Abbreviations: SE, standard error; *T*, continuous exposure; *X*, mismeasured continuous surrogate exposure; *Y*, continuous outcome; *Z*, continuous covariate.

[a] The estimated parameter $\hat{\alpha}$ quantifies the association between *T* and *Y*, adjusted for *Z*; the estimated parameter $\hat{\beta}$ quantifies the association between *X* and *Y*, adjusted for *Z*; and the estimated parameter $\hat{\gamma}$ quantifies the association between *X* and *Y* standardized to the *Z*-distribution among the exposed.

[b] Empirical standard error.

Table 2. Correction for Classical Exposure Measurement Error Using a Validation Subsample, Showing Results of Simulations of Associations Between a Continuous Covariate, Continuous Exposure, Mismeasured Continuous Surrogate Exposure, and Continuous Outcome

Simulation Setup		Model	Correction[a]	Estimate	SE[b]
$\sigma^2_{X\|T}$	$\omega_{T\|Z}$				
0.2	2	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.02	0.18
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.01	0.07
	1	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.03	0.21
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.02	0.12
	0	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.03	0.21
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.02	0.15
0.5	2	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.04	0.25
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.02	0.11
	1	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.06	0.29
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.05	0.20
	0	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.06	0.31
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.08	0.30
1.0	2	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.06	0.30
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.03	0.18
	1	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.07	0.34
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.08	0.34
	0	$E[Y_i\|Z_i, X_i]$	$\hat{\beta}(\sigma^2_{X\|Z}/\sigma^2_{T\|Z})$	1.05	0.32
		$E[Y_i\|\hat{F}(Z_i), X_i]$	$\hat{\gamma}(\sigma^2_{X}/\sigma^2_{T})$	1.10	0.46

Abbreviations: SE, standard error; *T*, continuous exposure; *X*, mismeasured continuous surrogate exposure; *Y*, continuous outcome; *Z*, continuous covariate.

[a] The estimated parameter $\hat{\beta}$ quantifies the association between *X* and *Y*, adjusted for *Z*, and the estimated parameter $\hat{\gamma}$ quantifies the association between *X* and *Y* standardized to the *Z*-distribution among the exposed.

[b] Empirical standard error.

covariate conditional estimates of association had similar precision (i.e., the empirical standard error of $\hat{\gamma}$ was equal to the empirical standard error of $\hat{\beta}$). Under the simulation conditions, the average of the estimated robust standard errors was similar to the empirical standard error of $\hat{\gamma}$ (Web Table 1), whereas in simulation scenarios with a smaller unexposed reference group, robust standard errors were not always conservative and suggest bootstrap-based confidence intervals may be preferred for $\hat{\gamma}$ (Web Table 1).

Measurement error correction using a validation study

The simulation results in Table 2 illustrate that when an exposure validation study is available with which to assess exposure measurement error, the proposed marginal estimate of the continuous exposure variable–continuous outcome variable association can be corrected for measurement error more efficiently than the covariate-conditional estimate of association. The proposed approach led to corrected estimates of association that tended to be very close to the true association specified under the simulation setup, $\hat{\alpha} = 1$. When the covariate *Z* is a confounder, and therefore associated with true exposure *T* (i.e., $\omega_{T|Z} > 0$), the empirical standard errors for measurement error–corrected measures of association derived using the proposed approach tended to be smaller than the empirical standard errors for measurement–error corrected covariate-conditional estimates of association.

Proposed approach: binary outcome variable

Web Table 2 reports results of simulations in which the outcome is a binary variable. The average covariate-adjusted estimated parameter $\hat{\beta}$ obtained using the error-prone proxy *X* is attenuated relative to the simulation setup for the underlying association. The average covariate-standardized estimated parameter, $\hat{\gamma}$, tended to suffer less attenuation bias than the covariate-adjusted association, $\hat{\beta}$. Only, when *T* and *Z* are uncorrelated (i.e., $\omega_{T|Z} = 0$), was the standardized estimate equal to the *Z*-conditional estimate of the exposure-outcome association ($\hat{\gamma} = \hat{\beta}$).

Empirical results

The proposed standardized estimate of the association between reading for pleasure and spherical equivalent refraction was −0.0704 (95% confidence interval: −0.1392, −0.0015) diopter change per log-unit increase in hourly reading for pleasure each week. The covariate conditional estimate of the association between reading for pleasure and spherical equivalent refraction was −0.0277 (95% confidence interval: −0.1062, 0.0508) diopter change per log-unit increase in hourly reading for pleasure each week. The crude estimate of the association (−0.0630; 95% confidence interval: −0.1412, −0.0151) is fairly close to the standardized estimate (and fairly far from the covariate conditional estimate), which would suggest that there is not much net confounding bias being addressed by conditioning on the set of covariates but quite a bit of inflation of attenuation bias due to measurement error occurring upon conditioning on these covariates.

DISCUSSION

This paper discusses regression analysis of an exposure-response association with an error-prone exposure variable. First, we reviewed relevant theory as well as providing simulations to illustrate that a multivariable regression analysis that adjusts for confounders may suffer greater attenuation bias due to classical exposure measurement error than a covariate-standardized linear regression analysis (Table 1). Second, we illustrated that when a gold-standard assessment is available and a validation study has been conducted, measurement error correction using the marginal estimates of (true and surrogate) exposures derived in a validation study may be substantially more statistically efficient than an approach to correction for measurement error that requires covariate-conditional estimates of exposures (Table 2).

In the classical exposure measurement error setting, a covariate-conditional linear regression estimate of association suffers bias. Attenuation bias due to classical measurement tends to increase as covariates that are associated with exposure are introduced into the regression model. Interestingly, this poses a notable challenge to the logic of any "change in estimate" approach to variable selection for regression modeling (20, 21) because, in settings where there is classical exposure measurement error, an investigator cannot distinguish between a change-in-estimate of the exposure-outcome association that results from control for confounding bias upon inclusion of a covariate in the regression model and a change-in-estimate of an exposure-outcome association that occurs due to increasing attenuation bias from classical exposure measurement error upon inclusion of a covariate in the model. Our proposed approach reduces attenuation bias to the ratio of marginal variances, σ_T^2/σ_X^2, and the degree of attenuation bias will not depend upon the adjustment set of covariates.

In simulations we focused on a simple setting with a single covariate assumed to be measured without error. The simple setting allowed us to illustrate that simple expectations from theory concur with simulation results obtained when applying the proposed model. The approach readily extends to multivariable regressions, as illustrated in our empirical example. In this study we did not address the setting where the confounder also is measured with error. Prior work suggests that when classical measurement error affects confounders, as well as the exposure of interest, there will be residual confounding in a multivariable regression estimate of the exposure-outcome of primary interest coupled with the inflation of attenuation bias due to classical measurement error in the exposure variable of interest that occurs upon conditioning for covariates that are predictors of exposure (6). The proposed marginal estimator also is expected to be susceptible to residual confounding when covariates are measured with error; however, as in the setting of interest in the present study (where only the exposure of interest is measured with error), the proposed marginal estimator is not susceptible to the bias inflation that can occur as confounding variables are included in a multivariable regression model. In simulations, while the true exposure is nonnegative, we allowed that the explanatory variable included in the regression model could be negative due to classical measurement error. In occupational settings, for example, negative values may occur when background exposures must be subtracted from personal exposure readings, and in the context of multiplicative errors, if the investigator fits a regression model for $\ln(X_i)$, that explanatory variable may be negative (although T and X are nonnegative).

The proposed approach relies upon a disease score that is estimated in an unexposed reference group. An investigator may have multiple options for defining the unexposed reference group to be used when estimating a disease score. If the exposure of interest is newly emerging, a well-defined unexposed group may be defined using historical information. For example, in a study of the safety of a newly introduced agent, one could use historical data to model the disease score before introduction of the new agent. Such settings arise in occupational and environmental studies when a novel environmental contaminant or a change in industrial process occurs; one could use historical data to model the disease score prior to the introduction of the environmental or occupational hazard. Historical reference groups, and external reference groups, are often used in the literature on disease risk score models as basis for estimation of a disease risk score (22, 23). Alternatively, in some settings an investigator will undertake a qualitative analysis to identify the presence or absence of exposure prior to a quantitative analysis to determine the magnitude or concentration of exposure. Given an accurate determination of the absence of exposure by the initial test, there may be no need to perform a quantitative assessment of exposure on subjects for whom there were negative reports obtained by the initial test. Other settings arise in environmental studies where knowledge of process or exposure transport is used to define an unexposed group, such that quantitative estimates of the magnitude or concentration of exposure are derived only for those presumed to have exposure potential. In some studies, the agent of concern is ubiquitous, and an "unexposed" reference group is constituted by a group with a minimal background level of exposure or constituted by those for whom measured values were below the minimal limit of detection. We then proceed by using this group for estimation of a disease score; observations below the limit of detection contribute to covariate standardization while the slope of the exposure-outcome trend is estimated based on the observations above the limit of detection.

The proposed approach targets estimation of the exposure effect among the exposed, such that different groups defined by levels of exposure are not necessarily mutually standardized to a common target population. Lack of mutual standardization is inconsequential when the exposure effect measures are homogeneous over levels of covariates. This was the setting described by the data-generating model we presented in equation 1 and conforms to the conditions we specified in our simulations. We have recently described a model that facilitates assessment of when conditions hold for examination of trends using a disease risk score (24). A key condition for the proposed approach to hold is that the estimated coefficients for the covariates, $\mathbf{Z}$, derived in the disease score model corresponds to the values for those coefficients that we would derive if we fit the fully conditional model in equation 2 (i.e., in a model that included T, the true exposure). One suggested approach to assess whether the model for estimation of a disease risk score is appropriately specified is a "dry-run" analysis in which the referent group is partitioned into pseudo-exposed and pseudo-unexposed groups so that differences in the observed covariates resemble differences between the actual exposed and unexposed populations, but the adjusted (pseudo-)exposure-outcome association is expected to be null. Therefore, the disease risk score model is evaluated by its ability to retrieve an unconfounded null estimate after adjustment in this "dry-run" analysis (25).

Exposure measurement error and confounding are ubiquitous in observational studies. The proposed approach provides a useful approach to reduce bias due to classical exposure measurement error in settings where there is potential confounding that may be controlled through standardization by a set of measured covariates. Moreover, when a validation study is available, the proposed approach may provide an efficient way to correct for bias due to exposure measurement error. While in the text we focus on a setting of an underlying disease model where the outcome follows a linear regression for a continuous outcome, this is largely to focus on a framework where the biases under each approach can be readily calculated based on prior methods in the literature. As illustrated in the simulations with binary outcome variables, the approach has applicability to nonlinear regression settings, including risk regressions where similar patterns of attenuation are expected (6, 14).

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States (David B. Richardson, Alexander P. Keil, Stephen R. Cole, and Jessie K. Edwards).

D.B.R. received funding support from the US Department of Health and Human Services, Centers for Disease Control and Prevention, and National Institute for Occupational Safety and Health (grant R01OH011409).

Conflict of interest: none declared.

REFERENCES

1. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. London, UK: Chapman & Hall; 2006.
2. Zeger SL, Thomas D, Dominici F, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect*. 2000; 108(5):419–426.
3. Gilbert ES, Fix JJ. *Laboratory Measurement Error in External Dose Estimates and Its Effects on Dose-Response Analyses of Hanford Worker Mortality Data*. Richland, WA: Hanford Site; 1996.
4. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int J Epidemiol*. 2020;49(1): 338–347.
5. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. *Annu Rev Public Health*. 1993;14:69–93.
6. Armstrong BG. The effects of measurement errors on relative risk regressions. *Am J Epidemiol*. 1990;132(6): 1176–1184.
7. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95(2):481–488.
8. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
9. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163(3):262–270.
10. Brookhart MA, Wyss R, Layton JB, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6(5): 604–611.
11. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
12. Tadrous M, Gagne JJ, Stürmer T, et al. Disease risk score as a confounder summary method: systematic review and recommendations. *Pharmacoepidemiol Drug Saf*. 2013; 22(2):122–129.
13. Stürmer T, Schneeweiss S, Brookhart MA, et al. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol*. 2005;161(9):891–898.
14. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med*. 1998;55(10):651–656.
15. Richardson DB, Keil AP, Kinlaw AC, et al. Marginal structural models for risk or prevalence ratios for a point exposure using a disease risk score. *Am J Epidemiol*. 2019; 188(5):960–966.

16. Tapsoba JD, Chao EC, Wang CY. Simulation extrapolation method for Cox regression model with a mixture of Berkson and classical errors in the covariates using calibration data. *Int J Biostat*. 2019;15(2):20180028.
17. Alexeeff SE, Carroll RJ, Coull B. Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics*. 2016;17(2):377–389.
18. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Stat Med*. 2008;27(30):6332–6350.
19. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley; 2013.
20. Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health*. 1989;79(3):340–349.
21. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*. 2008;167(5):523–529.
22. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res*. 2009;18(1):67–80.
23. Wyss R, Glynn RJ, Gagne JJ. A review of disease risk scores and their application in pharmacoepidemiology. *Curr Epidemiol Rep*. 2016;3:277–284.
24. Richardson DB, Keil AP, Cole SR, et al. Assessing exposure-response trends using the disease risk score. *Epidemiology*. 2020;31(2):e15–e16.
25. Wyss R, Hansen BB, Ellis AR, et al. The "dry-run" analysis: a method for evaluating risk scores for confounding control. *Am J Epidemiol*. 2017;185(9):842–852.