A Bespoke Instrumental Variable Approach to Correction for Exposure Measurement Error

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

* Correspondence to Dr. David Richardson, Department of Environmental and Occupational Health, Program in Public Health, Susan and Henry Samueli College of Health Sciences, University of California, Irvine, 100 Theory, Suite 100, Irvine, CA 92697 (e-mail: david.richardson@uci.edu).

Initially submitted September 27, 2021; accepted for publication July 22, 2022.

A covariate-adjusted estimate of an exposure-outcome association may be biased if the exposure variable suffers measurement error. We propose an approach to correct for exposure measurement error in a covariate-adjusted estimate of the association between a continuous exposure variable and outcome of interest. Our proposed approach requires data for a reference population in which the exposure was a priori set to some known level (e.g., 0, and is therefore unexposed); however, our approach does not require an exposure validation study or replicate measures of exposure, which are typically needed when addressing bias due to exposure measurement error. A key condition for this method, which we refer to as "partial population exchangeability," requires that the association between a measured covariate and outcome in the reference population equals the association between that covariate and outcome in the target population in the absence of exposure. We illustrate the approach using simulations and an example.

bias; cohort studies; epidemiologic methods; regression analysis

Abbreviations: CI, confidence interval; IV, instrumental variable.

Exposures are often difficult to accurately quantify. Therefore, exposure measurement error is a limitation of most epidemiologic studies. In the context of linear (1) and generalized linear (2) regression models, it is well known that if a continuous exposure suffers classical measurement error (i.e., its measured values are distributed around the true exposure with independent error), then an estimate of the exposure-outcome association may be attenuated (3, 4), although attenuation may not occur in other settings (5). Despite the prevalence of such problems, it remains rare for epidemiologic analyses to employ quantitative approaches to address exposure measurement error (5). One reason may be that many approaches require prior knowledge about the structure and magnitude of measurement error, data from an exposure validation study, or replicate exposure measurements.

In a prior paper, it was shown that a marginal estimator of an exposure-outcome association (i.e., one that handles confounders through a model-based approach to direct standardization) may suffer less bias due to classical measurement error than a covariate-conditional estimator of association (6, 7); however, to fully correct for bias due to exposure measurement error, the approach required information derived from replicate measurements or an exposure validation study with a gold standard (6). Here, we propose a method to correct for bias due to exposure measurement error in a covariate-adjusted estimate of an exposure-outcome association, and the proposed approach does not require such validation data. Instead, our proposed approach requires data for an external reference population that satisfies certain partial exchangeability conditions.

We describe the proposed method in the context of both linear and log-linear outcome models where an exposure variable of primary interest suffers measurement error. The proposed method, which builds upon our prior work using an instrument variable-like analysis (8), may be used to correct for classical measurement error as well as some types of nonclassical measurement error. The method is illustrated with simulations as well as an example.

METHODS

We focused on the setting of an epidemiologic study with an exposure variable of primary interest, A, an outcome, Y, and confounders measured with negligible error, Z (Figure 1). Our interest is in the effect of A on Y adjusted for Z. We assumed that in the target population, A is not deterministically assigned (e.g., as a function of Z), such that, prior to being exposed, each person had a non-negligible opportunity to receive any value of A (including 0).

Suppose that the outcome of interest, Y, follows either a linear additive model of the form,

$$E[Y|A,Z] = \alpha_0 + \alpha_1 A + \alpha_2 Z, \tag{1}$$

or a log-linear model of the form,

$$E[Y|A,Z] = \exp\left(\beta_0 + \beta_1 A + \beta_2 Z\right). \tag{2}$$

Unfortunately, we do not observe A. Rather, we observe a mismeasured version of A, denoted A^*, prone to measurement error of the form,

$$A^* = A + U, \tag{3}$$

where $E(U|A,Z) = 0$. This measurement error model accommodates classical measurement error (i.e., where $U \sim N\left(0, \sigma_U^2\right)$) but also allows us to relax some of the assumptions of that model—for example, allowing that the variance of U increases proportional to the true exposure, A, or varies depending upon the outcome variable, Y (see Web Appendix 1, available at https://doi.org/10.1093/aje/kwac133). Our measurement error model (equation 3) may apply to binary exposure variables; however, for binary A and A^* we note that the condition $E(U|A,Z) = 0$ holds under only limited circumstances (see Web Appendix 1).

If we fitted a regression model for Y on A^*, adjusted for Z, similar to equation 1 or 2 but with the mismeasured surrogate A^* rather than A, we would obtain an estimate that suffered bias due to exposure measurement error.

Proposed method

Reference population. The proposed method requires availability of a reference population in which exposure or treatment is set, as a result of a possibly hypothetical external intervention, to a constant value for the whole population. Here we will focus on the setting in which the reference population lacked the opportunity for exposure to A, and therefore A is set to 0 for the entire reference population.

An unexposed reference population may be based on observations prior to the exposure becoming available, in which case the external intervention preventing the opportunity for exposure is calendar time. For example, in environmental epidemiology, if the exposure of interest has emerged recently (e.g., a new flame retardant) then information prior to introduction of a new exposure may serve this purpose. In other settings, an unexposed reference population may arise from spatial or physical considerations that define the external intervention preventing exposure. For example, in an occupational setting, if the exposure is localized to a defined work area or department, then information for those employed in "clean" work areas or departments at a defined facility (i.e., locations at which the occupational exposure of interest was not present) may serve this purpose.

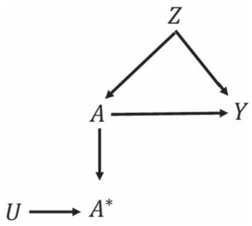

Figure 1. Illustration of relationship between exposure, A, outcome, Y, covariate, Z, and error-prone exposure measure, A^*.

Let $R = 1$ denote the reference population while $R = 0$ denotes the target population of individuals all of whom a priori had an opportunity for exposure. In the reference population, covariates Z and outcome Y are observed. Note that, as will be formalized below, we do not a priori assume that the target and reference populations are necessarily random samples from a common underlying population. Rather, the target and reference populations may differ with respect to distributions of measured and unmeasured covariates. However, as will be formalized below, we will require that certain key population characteristics remain invariant across reference and target populations.

Prognostic score. Let $F(Z)$ denote the expected value of Y conditional on Z among individuals in the reference population, $E[Y|Z, R = 1]$. This type of function is sometimes referred to as a prognostic score or disease risk score. Here, we focused on the setting in which $F(Z)$ is estimated using an external sample, or historical reference sample, that a priori lacked the opportunity for exposure to A and is therefore unexposed, an approach discussed by Hansen and Desai et al. (9, 10) in the context of estimation of prognostic scores. An important distinction can be made between a prognostic score that captures the relationship between covariates and outcome among the subset of persons who happened to be unexposed (and is therefore an estimate of the Z-Y association conditional on exposure equal to zero), and (as in our setting) a prognostic score that represents the Z-Y association in a reference population in which the exposure is set to zero for the entire reference population. Note that by virtue of treatment exclusion in the reference population, one has that $F(Z) = E[Y^{(a=0)}|Z, R = 1]$, where Y^a denotes a potential outcome that would be observed if exposed to treatment value a. An estimate of $F(Z)$ may be obtained by fitting a regression of Y on Z to data for the reference population, and, using the estimated regression model coefficients and information on Z in the target population, we can compute $\hat{F}(Z)$ for all individuals in our study.

Bespoke instrumental variable. Let $\tilde{Y}$ denote an $\hat{F}(Z)$-centered version of the outcome variable, which on the additive scale implies $\tilde{Y} = Y - \hat{F}(Z)$. Under conditions formalized below, $\hat{F}(Z)$ captures the dependence between the average potential outcome in the absence of exposure and Z in both reference and target populations, in which case $\tilde{Y}$ and Z can be expected to be mean independent in the target population. Because Z are observed variables that predict A but not $\tilde{Y}$ in the target population, we refer to Z as "bespoke" instrumental variables (IVs) (8).

The conditional mean independence afforded by $\tilde{Y}$ permits an IV-type analysis that addresses exposure measurement error in a linear or log-linear regression.

Sufficient conditions for bespoke IV identification. Suppose that each person in the reference and target populations has a potential outcome variable Y^a that would be observed if exposed to treatment value a. Further suppose that $Z = (Z_1,\ Z_2)$ and that rather than taking all of Z as candidate bespoke IVs, we take Z_1 only as a bespoke IV, and Z_2 are additional covariates that we adjust for. Below we establish identification of the semiparametric marginal structural model $E[Y^a - Y^{a=0}|a, z_2, R = 0] = \beta(z_2)a$ leveraging the instrument-like properties of Z_1, to account for the fact that one observes A^* instead of A.

For a linear structural nested model, we make the following assumptions:

1) Consistency, such that $Y^a = Y$ if $A = a$ and $R = 0$.
2) A degenerate reference population with $R = 1$, in which we have,

$$E[Y|R = 1, Z] = E\big[Y^{(a=0)}|R = 1, Z\big].$$

3) Partial population exchangeability, such that

$$E\big[Y^{a=0}|R = 0, Z_1 = z_1, Z_2\big] - E\big[Y^{a=0}|R = 0, Z_1 = 0, Z_2\big]$$

$$= E\big[Y^{a=0}|R=1, Z_1=z_1, Z_2\big] - E\big[Y^{a=0}|R=1, Z_1=0, Z_2\big],$$
for all values of z_1.

4) Partial additive causal effect homogeneity (i.e., no interaction between A and Z_1 on the scale of interest) in causing the outcome, such that

$E\big[Y^a - Y^{a=0}|A = a, z_1, z_2, R = 0\big] = E\big[Y^a - Y^{a=0}|A = a, z_1 = 0, z_2, R = 0\big]$ for all z_1.

5) Bespoke IV relevance:

$E\big[A|z_1, z_2, R = 0\big] - E\big[A|z_1 = 0, z_2, R = 0\big] \neq 0$ for each observed z_2.

6) Mean partially unbiased measurement error:

$$E\big[A^*|R = 0, Z_1 = z_1, Z_2\big] - E\big[A^*|R = 0, Z_1 = 0, Z_2\big]$$
$$= E[A|R = 0, Z_1 = z_1, Z_2] - E[A|R = 0, Z_1 = 0, Z_2].$$

Notably, the condition that we refer to as "partial population exchangeability" is substantially weaker than conditional population exchangeability of the target and reference populations (i.e., $E\big[Y^{a=0}|R = 0, Z_1, Z_2\big] = E\big[Y^{a=0}|R = 1, Z_1, Z_2\big]$) or full population exchangeability that the joint distribution of $(Y^{a=0}, Z_1, Z_2)$ is the same in both populations. This condition essentially requires that the Z_1-Y association in an unexposed reference population matches the Z_1-$Y^{a=0}$ association in the target population conditional on Z_2. Bespoke instrument relevance (assumption 5), in which the instrument has a causal effect on A, essentially requires that Z_1 predicts A within strata of Z_2. Assumption 6 relaxes the classical measurement error model. Under assumptions 1–6, we prove the following result in the Web Appendix 2:

$E\big[Y^a - Y^{a=0}|A = a, z_1, z_2, R = 0\big] = \beta\big(z_2\big)a$ is point identified by the bespoke IV estimand,

$$E\big[Y^a - Y^{a=0}|\, a, Z_2, R = 0\big]$$
$$= \frac{E[Y - F(Z)|R=0, Z_1=z_1, Z_2] - E[Y - F(Z)|R=0, Z_1=0, Z_2]}{E[A^*|R=0, Z_1=z_1, Z_2] - E[A^*|R=0, Z_1=0, Z_2]}.$$

In Web Appendix 2, we establish an analogous identification result under a semiparametric marginal structural log-linear model provided a stronger identification condition that the measurement error, U, is conditionally independent of (A, Z_1, Y), given Z_2 in the target population.

Linear regression. Suppose that, focusing on the target population ($R = 0$), we wish to estimate the parameter α_1 in a linear model of the form $= \alpha_0 + \alpha_1 A + \alpha_2 Z + \varepsilon$, where $E(\varepsilon|A, Z) = 0$; however, we only observe the mismeasured surrogate A^* rather than A. We can derive an estimate of α_1 by treating Z as a bespoke IV (for clarity here, suppressing notation to indicate conditioning on $R = 0$) and using a familiar IV estimator, $\alpha_1^{IV} = \frac{\operatorname{cov}(\tilde{Y},Z)}{\operatorname{cov}(A^*,Z)}$, which yields a consistent estimate of the parameter of interest, α_1, noting that,

$$\alpha_1^{IV} = \frac{\operatorname{cov}(\tilde{Y},Z)}{\operatorname{cov}(A^*,Z)}$$
$$= \frac{\operatorname{cov}(\alpha_0+\alpha_1 A+\alpha_2 Z+\varepsilon-F(Z),Z)}{\operatorname{cov}(A+U,Z)}$$

$= \frac{\operatorname{cov}(\alpha_0+\alpha_1 A+\alpha_2 Z+\varepsilon-E(Y|Z,R=1),Z)}{\operatorname{cov}(A+U,Z)}$, and, under the conditions above,

$$\begin{aligned} &= \frac{\operatorname{cov}(\alpha_0+\alpha_1 A+\alpha_2 Z+\varepsilon-E(Z,A=0,R=0),Z)}{\operatorname{cov}(A+U,Z)} \\ &= \frac{\operatorname{cov}(\alpha_0+\alpha_1 A+\alpha_2 Z+\varepsilon-(\alpha_0+\alpha_2 Z),Z)}{\operatorname{cov}(A+U,Z)} \\ &= \frac{\operatorname{cov}(\alpha_1 A+\varepsilon,Z)}{\operatorname{cov}(A+U,Z)} \\ &= \frac{\alpha_1\sigma_{A,Z}}{\sigma_{A,Z}} = \alpha_1 \end{aligned},$$

noting that it suffices that $E(U|Z, R = 0) = 0$.

Suppose that $Z = (Z_1,\ Z_2)$. Rather than taking all of Z as candidate bespoke IVs, we might opt to use only one measured confounder, Z_1, as a bespoke IV (while Z_2 are additional covariates that we adjust for) because, as discussed, point identification requires a homogeneity assumption for the effect of A across levels of the bespoke IV.

Web Appendix 3 provides a brief description of a g-estimation approach to an IV estimator (11), incorporating a control function to create a "bespoke" IV, and SAS (SAS Institute, Inc., Cary, North Carolina) code, for estimation of the average effect of A on Y obtained under an additive structural mean model (11, 12). The approach follows by treating Z_1 (centered at its mean conditional on Z_2) as an instrument, and g-estimation may proceed using generalized method of moments to obtain an estimate, $\hat{\alpha}_1^{IV}$, that results in lack of association between Z_1 (centered at its mean conditional on Z_2) and $H(\hat{\alpha}_1^{IV}, \hat{\alpha}_0) = \tilde{Y} - \hat{\alpha}_1^{IV} A^* - \hat{\alpha}_0$.

Log-linear regression. Suppose that we are interested in estimation of the parameter, β_1, under a log-linear model of the form that was shown in equation 2 (again focusing on the target population, $R = 0$). However, we only observe the mismeasured surrogate A^* rather than A. We can derive an estimate of β_1 by treating Z as a bespoke IV and using g-estimation of the proposed "bespoke" IV estimator by identifying an estimate, $\hat{\beta}_1^{IV}$, that results in lack of association between Z_1 (centered at its mean conditional on Z_2) and $H(\hat{\beta}_1^{IV}, \hat{\beta}_0) = \tilde{Y} \exp\left(-\hat{\beta}_1^{IV} A^* - \hat{\beta}_0\right) - 1$, where $\tilde{Y} = Y/F(Z)$. Web Appendix 3 provides SAS (SAS Institute) code for estimation under a multiplicative structural mean model, implemented using generalized method of moments, and SAS code to obtain bootstrap-based confidence intervals.

Simulation example. Data were simulated for 1,000 studies, with 5,000 people in each study sample and 5,000 people in each reference sample. Each person was randomly assigned a covariate value Z_1 by sampling from a uniform distribution, $Z_1 \sim Uniform(-1, 1)$, and a covariate value Z_2 by sampling from a Bernoulli distribution, $Z_2 \sim Bern(0.5)$. We assigned A as a continuous variable that took a value of $\exp(-1 + 0.5Z_1 + 0.5Z_2 + \varepsilon_1)$, where $\varepsilon_1 \sim N(0{,}0.5)$; in the reference sample, A was set to 0. In the first set of simulations, the outcome variable, Y, was a continuous variable that took a value of $1 + 1 \times Z_1 + 1 \times Z_2 + 1 \times A + \varepsilon_2$, where $\varepsilon_2 \sim N(0{,}1)$. In the second set of simulations, the outcome variable, Y, was a binary variable that took a value of 1 with probability $0.1 + 0.1Z_1 + 0.1Z_2 + 0.1A$ (with resampling if the expression yielded a probability less than 0 or greater than 1).

We generated data under 3 scenarios. In the first scenario, a surrogate exposure, A^*, was generated under a model in which errors conform to the classical measurement error model, $A^* = A + U, U \sim N(0, \sigma_U)$. Simulations were conducted for scenarios where σ_U=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 (13–15). In the second scenario, a surrogate exposure, A^*, was generated under a model in which errors are proportional to the magnitude of the true exposure, $A^* = A + U, U \sim N(0, A)$. In the third scenario, a surrogate exposure, A^*, was generated under a model in which error depended upon the outcome variable, $A^* = A + U, U \sim N(0, e^{0.1Y})$.

For each simulated data set, we fitted a regression model for Y conditional on A, Z_1, and Z_2 to summarize the association in the study sample under the data generating model. We fitted a model for Y conditional on A^*, Z_1, and Z_2 to summarize the biased estimate of association when fitting a model using the error-prone variable A^* rather than A. Finally, we estimated the proposed IV estimator using the approach described in the text (and SAS code in Web Appendix 3). We summarized results from the simulated studies by computing the mean of the estimated association, the estimated standard deviation of the estimates (the empirical standard error, ESE), and the average squared difference between the estimated association and the specified true effect of A on Y (the mean squared error, MSE). In Web Appendix 4, we further report the average standard error and coverage probability of bootstrap-based 95% confidence intervals for simulations under the first scenarios. Additional simulations were conducted, assuming smaller study and reference samples (Web Appendix 5), and simulations were conducted under a log-linear model, which may be preferred when the outcome variable can take only positive values (Web Appendix 6).

Example. We illustrate the proposed method in empirical data that were derived from the Orinda Longitudinal Study of Myopia, 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 (16). The exposure of primary interest is self-reported hours per week reading for pleasure (READHR, in units of hours), and the outcome of interest is spherical equivalent refraction (SPHEQ, in units of diopter, a measure of the eye's effective focusing power). Covariates include age at study entry (AGE, in years); year of study entry (STUDYYEAR, in years); gender (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 considered those who reported a complete absence of reading for pleasure (0 hours) as an accurate indication of the absence of exposure. For the purposes of illustrating the proposed approach, we assumed that those who reported a complete absence of reading for pleasure were subject to an intervention preventing exposure (e.g., absence of books for pleasure reading in the household). Among those who reported reading for pleasure, we assumed that exposure estimates suffered error proportional to the true value, $\log(X^*) = \log(X) + U, U \sim N(0, \sigma_U)$, such that the imperfect exposure measure also was non-negative. 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. We derived an estimated prognostic 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 fitted a regression model for log(READHR) as a function of AGE, STUDYYEAR, GENDER, MOMMY, and DADMY. We derived an estimated exposure score as the predicted value of log(READHR) given the fitted model and observed covariates. We then estimated the diopter change per log-unit increase in hourly reading by fitting a regression model for SPHEQ as a function of the exposure score, with the estimated prognostic 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 log(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.

RESULTS

Simulation

Table 1 reports the simulation results for analyses of the association between a continuous exposure variable and a continuous outcome variable, where the exposure variable, A, had a mean of approximately 0.57 (standard deviation = 0.40), and the imperfect exposure measure A^* suffered exposure measurement error for scenarios, σ_U (0.2, 0.5, or 1.0). In all simulations, the average estimated association under a model for Y conditional on A, Z_1, and Z_2 was equal to 1. In all simulations, the average Z-conditional estimate of association obtained using the error-prone proxy A^* was biased towards the null. In all scenarios, the proposed estimator of the exposure-outcome association was approximately unbiased. As the degree of exposure measurement error, σ_U, increased, there was no increase in bias of the proposed estimator, but there was increasing bias in the covariate-conditional estimator of association between the error prone proxy exposure measure and outcome (as expected). The empirical standard errors of the covariate-conditional models tended to be slightly smaller than of the proposed estimator; the root mean square error of the Z-conditional estimate of association obtained using the error-prone proxy A^* was larger than that of the proposed estimator; when σ_U was larger, the root mean square error of the Z-conditional estimate of association obtained using the error-prone proxy A^* was substantially larger than that of the proposed estimator. Bootstrap-based 95% confidence intervals had close to nominal coverage (Web Table 1).

Table 1 also reports the simulation results for analyses of the association between a continuous exposure and a continuous outcome variable, where the proxy exposure variable suffers measurement error that is proportional to the true exposure, or where the proxy exposure variable suffers measurement error that depends upon the outcome. In these scenarios, the proposed estimator of the exposure-outcome association also was approximately unbiased.

Table 2 reports simulation results obtained in simulations where the outcome was a binary variable; overall patterns of results were similar to those observed in Table 1. In all scenarios, the proposed estimator of the exposure-outcome association was approximately unbiased.

Additional simulations were conducted in which the study and reference populations were smaller; again, the proposed estimator of the exposure-outcome association was approximately unbiased, albeit with larger empirical standard errors and root mean square error than in simulations where sample sizes were larger (Web Table 2). Simulations also were conducted under a log-linear model form, and in all scenarios, the proposed estimator of the exposure-outcome association was approximately unbiased (Web Tables 3 and 4).

Example results

The covariate conditional estimate of the association between reading for pleasure and spherical equivalent refraction was −0.03 (95% confidence interval (CI): −0.11, 0.05)-diopter change per log-unit increase in hourly reading for pleasure each week. The proposed corrected estimate of the association between reading for pleasure and spherical equivalent refraction, corrected for measurement error, was −0.12 (bootstrap 95% CI: −0.22, −0.02)-diopter change per log-unit increase in hourly reading for pleasure each week.

DISCUSSION

This paper discusses regression analysis of an exposure-response association with an error-prone exposure variable. We illustrated that, under the conditions examined, the proposed approach led to a notable reduction in bias due to measurement error and mean square error compared with a covariate-adjusted estimator of association between the error-prone exposure variable and outcome.

The approach requires information for a reference population in which the exposure is set by a possibly hypothetical intervention (and not by a selection process). We refer to a "hypothetical" intervention because we are assuming that there is some external force that reveals the treatment-free potential outcome, $Y^{a=0}$ in the reference population, and that, if this same external force were applied to the target population, the $Y^{a=0}$ that would be revealed would be similar to the treatment-free potential outcome $Y^{a=0}$ experienced by those in the reference population. Importantly, the proposed approach requires only "partial population exchangeability" of the target and reference populations, a condition substantially weaker than full exchangeability of the external reference and target populations. The target and reference populations may differ with respect to distributions of measured and unmeasured covariates; "partial population exchangeability" simply requires, given consistency and a degenerate reference population, that the Z_1-Y association in the reference population equals the Z_1-Y association among the unexposed in the target population.

Our proposed approach shifts the challenge for measurement error correction from that of obtaining validation data (in which exposure measures may be made using a gold-standard measurement tool), replicate measurements of the exposure, or another source of information about the structure and magnitude of the measurement error problem to that of identifying a suitable reference population that meets the necessary partial exchangeability conditions. In some settings a suitable reference population may be based on observations prior to the exposure occurring. For example, in environmental and occupational epidemiology, if an agent has emerged recently or if the exposure arose due to a change in industrial process, then a reference group may be based upon recent historical data. When drawing upon historical data, attention should be paid to possible changes over time in coding study data, standards of care, and other factors that could result in a violation of the partial exchangeability assumption. In other settings, a suitable reference population may arise from spatial or physical considerations that define

Table 1. Simulations of Associations Between Exposure, A, Mismeasured Surrogate Exposure, A^*, Covariates, Z, and Continuous Outcome, Y

Distribution of U Model	Simulation Setup Estimates Obtained From Fitted Models[a]		
	Estimate	ESE	MSE
$N(0, 0.2)$			
1 $E[Y\|Z, A]$	1.00	0.04	0.032
2 $E[Y\|Z, A^*]$	0.73	0.04	0.265
3 Proposed approach	1.00	0.10	0.077
$N(0, 0.5)$			
1 $E[Y\|Z, A]$	1.00	0.04	0.032
2 $E[Y\|Z, A^*]$	0.31	0.03	0.694
3 Proposed approach	1.00	0.10	0.080
$N(0, 1.0)$			
1 $E[Y\|Z, A]$	1.00	0.04	0.032
2 $E[Y\|Z, A^*]$	0.10	0.01	0.901
3 Proposed approach	1.00	0.12	0.092
$N(0, A)$			
1 $E[Y\|Z, A]$	1.00	0.04	0.032
2 $E[Y\|Z, A^*]$	0.18	0.02	0.816
3 Proposed approach	1.00	0.11	0.086
$N(0, e^{0.1Y})$			
1 $E[Y\|Z, A]$	1.00	0.04	0.032
2 $E[Y\|Z, A^*]$	0.07	0.01	0.935
3 Proposed approach	1.00	0.13	0.101

Abbreviations: ESE, empirical standard error; MSE, root mean square error.

[a] Results shown for 3 magnitudes of exposure measurement error, σ_U, and 2 scenarios where measurement error depends upon other covariates. Each scenario involves simulation of 1,000 studies with 5,000 people in the study sample and 5,000 people in an unexposed reference sample.

the hypothetical external intervention preventing exposure. When there is knowledge regarding the distribution or environmental transport of the agent, a reference group may be defined using that information. For example, potential exposure to a hazard may be limited to certain areas, while the remainder is unexposed and may serve as a reference group. Of course, some exposures are ubiquitous, such as the radioactive debris from atmospheric nuclear weapons testing, and this means that one cannot identify a completely unexposed reference group. In such circumstances one may conceive of a reference group that has experienced an "external intervention" that set exposure to a nonzero baseline value. The proposed method can accommodate a known baseline exposure common to all members of the reference group by simply centering A^* accordingly for the purposes of g-estimation.

The proposed approach is illustrated using data from a study of risk factors for myopia (16) that we have drawn upon previously to illustrate methods to minimize bias due to exposure measurement error (6). For the purposes of illustrating the proposed approach we assumed that those who reported a complete absence of reading for pleasure were subject to an intervention preventing exposure (e.g., absence of books for pleasure reading in the household), acknowledging that we assumed treatment exclusion in the reference population. These illustrative data provide an example of the proposed method with publicly available data and allow us to compare findings using the proposed approach to those obtained using a previously published method to minimize bias due to exposure misclassification. The covariate conditional estimate of association between reading for pleasure and spherical equivalent refraction was −0.03 (95% CI: −0.11, 0.05)-diopter change per log-unit increase in hourly reading for pleasure each week. In previous work, we illustrated a method to reduce bias due to exposure measurement error by standardization (6), yielding an estimate of −0.07 (95% CI: −0.14, −0.00)-diopter change per log-unit increase in hourly reading for pleasure each week. Here, employing our proposed method to correct for measurement error yields an estimate of −0.12 (bootstrap 95% CI: −0.22, −0.02)-diopter change per log-unit increase in hourly reading for pleasure each week. Of course, there are limitations to this illustrative example, and caution is warranted in interpretation. For example, in addition to

Table 2. Simulations of Associations Between Exposure, *A*, Mismeasured Surrogate Exposure, A^*, Covariates, *Z*, and Binary Outcome, *Y*

Distribution of *U* Model	Simulation Setup Estimates Obtained From Fitted Models[a]		
	Estimate	ESE	MSE
$N(0,0.2)$			
1 $E[Y\|Z,A]$	0.10	0.02	0.016
2 $E[Y\|Z,A^*]$	0.07	0.02	0.027
3 Proposed approach	0.10	0.03	0.026
$N(0,0.5)$			
1 $E[Y\|Z,A]$	0.10	0.02	0.016
2 $E[Y\|Z,A^*]$	0.03	0.01	0.069
3 Proposed approach	0.10	0.03	0.026
$N(0,1.0)$			
1 $E[Y\|Z,A]$	0.10	0.02	0.016
2 $E[Y\|Z,A^*]$	0.01	0.01	0.090
3 Proposed approach	0.10	0.03	0.026
$N(0,A)$			
1 $E[Y\|Z,A]$	0.10	0.02	0.016
2 $E[Y\|Z,A^*]$	0.02	0.01	0.082
3 Proposed approach	0.10	0.03	0.026
$N(0, \mathrm{e}^{0.1Y})$			
1 $E[Y\|Z,A]$	0.10	0.02	0.016
2 $E[Y\|Z,A^*]$	0.01	0.01	0.090
3 Proposed approach	0.10	0.03	0.026

Abbreviations: ESE, empirical standard error; MSE, root mean square error.

[a] Results shown for 3 magnitudes of exposure measurement error, σ_U, and 2 scenarios where measurement error depends upon other covariates. Each scenario involves simulation of 1,000 studies with 5,000 people in the study sample and 5,000 people in an unexposed reference sample.

treatment exclusion in the reference population, we assumed that the effects of age, sex, and maternal and paternal myopia on a child's measure of spherical equivalent refraction in the reference and target population are exchangeable; this assumption of partial exchangeability may be reasonable to the extent that age and parental effects may plausibly operate in a similar fashion in the target and reference samples.

We have recently demonstrated that a bespoke IV can be used to address biased due to unmeasured confounders (8). Here we demonstrate that a bespoke IV also can serve to address bias due to exposure measurement error. It is worth noting that the assumptions for identification in this paper do not rule out unmeasured confounding, so that the current work essentially extends our prior work by simultaneously also accounting for exposure measurement error. Therefore, the proposed approach will also afford control for unmeasured confounders of the *A*-*Y* association.

The proposed approach may come at a cost in terms of statistical efficiency; however, as we illustrate, the proposed approach will often result in reduction of mean squared error when measurement error is sizable. Consequently, the proposed estimator may provide another useful approach to address bias due to exposure measurement error in a range of epidemiologic study settings.

ACKNOWLEDGMENTS

Author affiliations: Department of Environmental and Occupational Health, Program in Public Health, University of California, Irvine, Irvine, California, United States (David B. Richardson); 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, Jessie K. Edwards, Stephen R. Cole); and Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, United States (Eric J. Tchetgen Tchetgen).

D.B.R. was supported by the National Institute for Occupational Safety and Health of the Centers for Disease Control and Prevention (grant R01 OH011409). E.J.T.T. was supported by the National Institute on Aging (grant R01 AG065276).

The data used to illustrate the method are publicly available and come from Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Hoboken, New Jersey: Wiley; 2013.

Conflict of interest: none declared.

REFERENCES

1. Fuller WA. *Measurement Error Models*. Hoboken, N.J: Wiley-Interscience; 2006.
2. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. London, UK: Chapman & Hall; 2006.
3. 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.
4. 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.
5. 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.
6. Richardson DB, Keil AP, Cole SR, et al. Reducing bias due to exposure measurement error using disease risk scores. *Am J Epidemiol.* 2020;190(4):621–629.
7. Richardson DB, Keil AP, Cole SR. Amplification of bias due to exposure measurement error. *Am J Epidemiol.* 2022; 191(1):182–187.
8. Richardson DB, Tchetgen Tchetgen EJ. Bespoke instruments: a new tool for addressing unmeasured confounders. *Am J Epidemiol.* 2022;191(5):939–947.
9. Hansen BB. The prognostic analogue of the propensity score. *Biometrika.* 2008;95(2):481–488.
10. Desai RJ, Glynn RJ, Wang S, et al. Performance of disease risk score matching in nested case-control studies: a simulation study. *Am J Epidemiol.* 2016;183(10):949–957.
11. Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology.* 2006;17(4):360–372.
12. Tchetgen Tchetgen E, Vansteelandt S. Alternative identification and inference for the effect of treatment on the treated with an instrumental variable. *Harvard University Biostatistics Working Paper Series* 2013;Working Paper 166. https://biostats.bepress.com/harvardbiostat/paper166. Accessed May, 2022.
13. 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):28.
14. 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.
15. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Stat Med.* 2008;27(30):6332–6350.
16. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley; 2013.