On Using Summary Statistics From an External Calibration Sample to Correct for Covariate Measurement Error

Ying Guo,[a] *Roderick J. Little,*[b] *and Daniel S. McConnell*[c]

Background: Covariate measurement error is common in epidemiologic studies. Current methods for correcting measurement error with information from external calibration samples are insufficient to provide valid adjusted inferences. We consider the problem of estimating the regression of an outcome Y on covariates X and Z, where Y and Z are observed, X is unobserved, but a variable W that measures X with error is observed. Information about measurement error is provided in an external calibration sample where data on X and W (but not Y and Z) are recorded.

Methods: We describe a method that uses summary statistics from the calibration sample to create multiple imputations of the missing values of X in the regression sample, so that the regression coefficients of Y on X and Z and associated standard errors can be estimated using simple multiple imputation combining rules, yielding valid statistical inferences under the assumption of a multivariate normal distribution.

Results: The proposed method is shown by simulation to provide better inferences than existing methods, namely the naive method, classical calibration, and regression calibration, particularly for correction for bias and achieving nominal confidence levels. We also illustrate our method with an example using linear regression to examine the relation between serum reproductive hormone concentrations and bone mineral density loss in midlife women in the Michigan Bone Health and Metabolism Study.

Conclusions: Existing methods fail to adjust appropriately for bias due to measurement error in the regression setting, particularly when measurement error is substantial. The proposed method corrects this deficiency.

(*Epidemiology* 2012;23: 165–174)

Submitted 6 April 2011; accepted 23 August 2011.

From [a]Merck & Co., Inc., Rahway, NJ; and the Departments of [b]Biostatistics and [c]Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI.

Supported in part with funding from the American Chemistry Council and the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health. The authors reported no other financial interests related to this research.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Ying Guo, Merck & Co., Inc., 126 East Lincoln Ave, RY34–316, Rahway NJ 07065. E-mail: ying.guo2@merck.com.

ISSN: 1044-3983/12/2301-0165
DOI: 10.1097/EDE.0b013e31823a4386

Many studies in epidemiology involve biomarkers recorded with measurement error, which distorts inferences. Specifically, in regression analysis, regression coefficients of variables subject to measurement error are attenuated, and treatment effects are potentially estimated with bias when variables subject to measurement error are included as covariates.[1–5] However, adjustments to correct these biases are rarely applied in epidemiologic studies.[6]

Information about measurement error is often contained in a calibration experiment such as a bioassay, where samples with known values of the variable are analyzed by a measuring instrument, and the regression of the measured values on the true values is estimated, yielding a calibration curve.[7] Low values with high measurement error are often reported as below the limit of detection, and other values are estimated from this calibration curve and treated as the true values in the main analysis. Browne and Whitcomb[8] provide a review of methods for determining the limit of detection and related quantities. Simulations have shown that this approach, which we call classical calibration, yields biased regression estimates when the measurement error is substantial.[9] This usual way of providing information from calibration experiments to users does not allow valid statistical inferences unless the measurement error is small. Better methods would allow useful information from calibrations with relatively high measurement error to be included in analysis.

We consider data from a main study and a calibration sample in the form of Figure 1. The main analysis concerns the regression of Y on X and Z, where Y is the outcome of interest, X is the true value of the biomarker of interest, and Z denotes other covariates, assumed to be measured without error. The main study data are a random sample on Y, Z, and W, where W is the measured version of the biomarker, the true value X measured with error. Information relating W and X is gained from a calibration sample that includes measurements on W and X. The shaded cells in Figure 1 represent unobserved values.

Figure 1 contrasts 2 calibration sample designs, which we call "internal calibration" and "external calibration." In internal calibration (Fig. 1A), values of X and W are available for a subsample of the main study participants, and Y and Z are also recorded for this subsample. External calibration is carried out independently of the main study, for example by an assay manufacturer, so values of Y and Z are not recorded

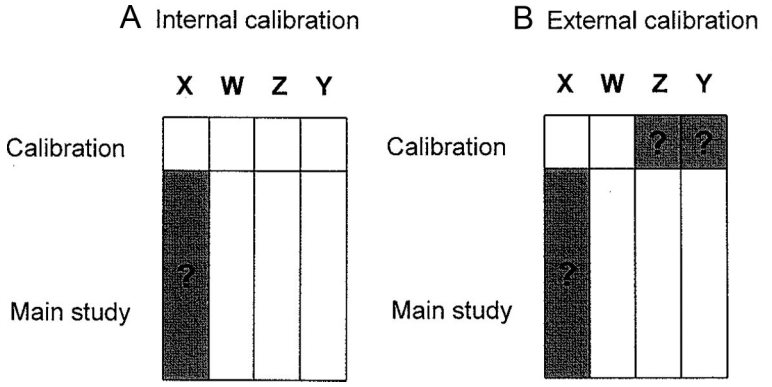

FIGURE 1. A, an internal calibration/main study design, and B, an external calibration/main study design. Shading indicates cells with missing data.

for the calibration sample, yielding the sparser data pattern of Figure 1B.

The literature on measurement error adjustments has largely concerned the internal calibration design. In that setting, regression calibration substitutes the estimated conditional expectation of the true biomarker given the observed surrogate and the other covariates into the primary regression model.[10] This method yields consistent estimates of the main regression parameters, under the nondifferential measurement error assumption that Y is independent of W given Z and X, which we denote as NDME(Y, $W \mid Z$, X). If the calibration data are available, standard errors of the estimates can be calculated using either bootstrap or sandwich estimation methods. When direct estimates of the regression of Y on (X, Z) are available from the calibration sample, they can be combined with the regression calibration estimates, yielding the method known as efficient regression calibration.[11]

An alternative to regression calibration and efficient regression calibration is multiple imputation for internal calibration,[12] where values of the biomarker X are imputed as draws from the conditional distribution of X given W, Z, and Y, estimated from the calibration subsample. This imputation step is repeated to create multiple completed data sets. Each completed data set is then analyzed using standard complete-data procedures, and estimates and standard errors from these analysis are combined using multiple-imputation combining rules given by Rubin.[13] One attraction of multiple imputation is that it is increasingly available in statistical software (SAS PROC MI, IVEware, MICE). Cole et al[14] propose this approach with a survival outcome model when one covariate is measured with error. Raghunathan[15] applies multiple imputation to data from the National Health and Nutrition Examination Survey to handle measurement error in self-reports of health conditions. He and Zaslavsky[16] address underreporting of cancer therapies in registry systems by using multiple imputation to impute correct treatment status, using information from medical records in a calibration sample. Freedman et al[17] compare regression calibration, efficient regression calibration, and multiple imputation under the multivariate normal model, for the internal calibration design. They show in simulations that regression calibration, efficient regression calibration, and multiple imputation for internal calibration have minimal bias and that efficient regression calibration is more efficient than the other 2 approaches. Guo and Little[9] show that more efficient versions of multiple imputation for internal calibration are available that exploit the nondifferential measurement error assumption, and these are as efficient or more efficient than efficient regression calibration. They also extend multiple imputation for internal calibration to handle measurement error with nonconstant variance, a situation that regression calibration and efficient regression calibration are ill-equipped to handle.

Our focus here is on the external calibration design in Figure 1B. Because biomarkers are commonly calibrated by assay producers independently of the main study, this situation is much more common than that of the internal calibration design, but methods for this case have received limited attention. The classical calibration method uses only the information of X and W and hence can be applied to external calibration data. However, it is known from previous studies that classical calibration is biased when the measurement error is substantial.[9] The regression calibration, efficient regression calibration, and multiple imputation for internal calibration methods all require information in the calibration sample that is not available with the external calibration design: for regression calibration, the values of Z; and for efficient regression calibration and multiple imputation for internal calibration, the values of both Y and Z. Our simulation study shows that versions of multiple imputation for internal calibration or regression calibration based only on the distribution of X given W, which can be applied to external calibration data, both yield biased inferences for the regression of Y on X and Z.

We propose multiple imputation for external calibration, a new method that addresses these problems. It requires only summary statistics from the calibration sample, an important consideration because the microdata (ie, subject-

level or unit-level external calibration data) are generally not made available from an external calibration sample. The method yields valid multiple imputation inferences for the regression of Y on X and Z, although values of Y and Z are missing in the calibration sample. Like multiple imputation for internal calibration, it is based on a multivariate normal model, but it is not the standard version of multiple imputation, as implemented in programs like PROC MI in SAS; that method is actually not feasible for external calibration data, because there is insufficient information to estimate all the imputation model parameters. The multiple imputation for external calibration method solves this problem by exploiting parameter restrictions based on the nondifferential measurement error assumption. The method is remarkably simple, because it is a direct simulation method that does not require iterative computations. More statistical details on multiple imputation for external calibration, and R code to implement it, are provided in the Appendix and the eAppendix (http://links.lww.com/EDE/A525), respectively.

Our proposed method is illustrated using data to assess the association between X = concentration of sex hormone-binding globulin (SHBG) and Y = bone mineral density loss, adjusting for Z = age and body mass index (BMI), for midlife women from the 2008 Michigan Bone Health and Metabolism Study.[18] The true SHBG concentration X for each participant is unobserved, but an assay measure W is collected, which can be viewed as an error-contaminated version of the true concentration X. Calibration data on the joint distribution of X and W are also available.

We describe the model that underlies multiple imputation for external calibration, and outline the algorithm for creating multiple imputations. We provide a simulation study comparing multiple imputation for external calibration with competing methods and present sensitivity analysis to examine the robustness of multiple imputation for external calibration. We also show an application to real data from the Michigan Bone Health and Metabolism Study.

PROPOSED METHOD

We write $U = (Y, Z)$, a vector of p variables, for the set of q outcomes Y and r covariates Z other than X, where $p = q + r$. As q and r may be greater than 1, the formulation covers multivariate regression with one or more dependent variables and one or more covariates. We assume here that X and its surrogate W are scalar, although our method can be extended to handle more than 1 variable subject to measurement error.

We assume that in the main sample and the calibration sample, the conditional distribution of U and X given W has a joint $(p + 1)$-variate normal distribution with a mean that is linear in W and a constant covariance matrix. This conditional distribution is assumed to be the same in the main study sample and the calibration sample, although the distribution of W can differ in the 2 samples. This indispensable assumption is related to the transportability across studies assumption.[19] Further, we make the following nondifferential measurement error assumption:

NDME(U, $W \mid X$): the distribution of U given W and X does not depend on W.

That is, the measurement error in W is assumed to be unrelated to values of $U = (Y, Z)$, conditional on the true value X. This assumption is stronger than the nondifferential measurement error assumption for internal calibration, which assumes the measurement error is unrelated to Y conditional on X and the covariates Z. The stronger assumption is needed given the more limited information available in external calibration data. However, we believe the assumption is plausible in many bioassays. The NDME(U, $W \mid X$) assumption will hereafter be referred to simply as the NDME assumption.

Our method generates imputations of X from the conditional distribution of X given the observed variables in the main study sample, namely Y, Z, W; let $\phi = (\lambda, \delta)$ where λ denotes the vector of regression coefficients of regression X on (Y, Z, W) and δ denotes the residual standard deviation for that regression. For data set d, a draw $\phi^{(d)} = (\lambda^{(d)}, \delta^{(d)})$ is taken from the posterior distribution of ϕ given the data. This draw can be computed rather simply from the main sample data and summary statistics from the external calibration sample, namely the sample size, sample mean, and sum of squares and cross products matrix of X and W. The missing value x_i of X for the ith observation in the study sample is then imputed by

$$\hat{x}_i^{(d)} = E(x_i \mid y_i, z_i, w_i, \lambda^{(d)}) + z_i^{(d)} \delta^{(d)} \quad (1)$$

where $E(x_i \mid y_i, z_i, u_i, \lambda^{(d)})$ is the conditional mean of x_i given (y_i, z_i, w_i), the values of (Y, Z, W) for case i, and $z_i^{(d)}$ is a draw from the standard normal distribution. This method is proper in the sense defined by Rubin,[13] as it takes into account uncertainty in estimating ϕ. An alternative approach is to replace $\phi^{(d)}$ by the maximum likelihood estimate $\hat{\phi}$ in the given formula, but this method is improper—it does not propagate uncertainty in estimating ϕ—and hence is inferior to the proper method.

The key aspect of the method is deriving p estimates of the partial covariances between X and the components of U, given W. These p parameters cannot be estimated directly from the data in Figure 1B because X and U are never observed together. However, the nondifferential measurement error assumption implies that the p coefficients of W in the regression of U on W and X are zero. These p parameter restrictions allow the missing covariances to be expressed in terms of parameters that can be estimated from the available data, and then multiple imputations of the missing values to be created. More statistical details on how $\hat{\phi}$ and $\phi^{(d)}$ are computed, and software (R source code) to implement the proper version of

multiple imputation for external calibration, are provided in the eAppendix (http://links.lww.com/EDE/A525).

In this article, we consider the situation in which the external calibration data are not available for inclusion in the postimputation analysis. Reiter[20] shows that in this situation, the standard variance estimator obtained from the multiple-imputation combining rules[13] is positively biased and confidence interval coverage exceeds 95%. Here, we follow Reiter's two-stage imputation procedure to generate imputations that enable consistent estimation of variances. Specifically, we first draw d values of model parameters $\phi^{(d)}$; then, for each $\phi^{(d)}$, $d = 1, \ldots, m$, we construct n imputed data sets by generating n sets of draws of X. Finally, this procedure yields a collection of $M = m \times n$ imputed data sets, which can be analyzed by standard complete data inference. The results from all imputed data sets are combined to obtain valid inferences using the following combining rules suggested by Reiter.[20] For $d = 1, \ldots, m$ and $l = 1, \ldots, n$, let $\hat{\gamma}^{(d,l)}$ and $var(\hat{\gamma}^{(d,l)})$ be the estimate of parameters of interest and the corresponding estimated variance computed with the (d, l) data set, respectively. The multiple imputation estimate of γ, $\hat{\gamma}_{MI}$, and associated variance T_{MI} are calculated as

$$\hat{\gamma}_{MI} = \sum_{d=1}^{m}\sum_{l=1}^{n} \hat{\gamma}^{(d,l)}/(mn) = \sum_{d=1}^{m} \bar{\gamma}_n^{(d)}/m$$

$$T_{MI} = U - W + (1 + 1/m)\,B - W/n$$

with

$$W = \sum_{d=1}^{m}\sum_{l=1}^{n} (\hat{\gamma}^{(d,l)} - \bar{\gamma}_n^{(d)})^2/(m(n-1))$$

$$B = \sum_{d=1}^{m} (\bar{\gamma}_n^{(d)} - \hat{\gamma}_{MI})^2/(m-1)$$

$$U = \sum_{d=1}^{m}\sum_{l=1}^{n} var(\hat{\gamma}^{(d,l)})/mn$$

The 95% confidence intervals for the multiple imputation estimate are calculated as $\hat{\gamma}_{MI} \pm t_{0.975,\upsilon}\sqrt{T_{MI}}$, with degrees of freedom $\upsilon = \left[\frac{((1+1/m)B)^2}{(m-1)T_{MI}} + \frac{((1+1/n)W)^2}{(m(n-1))T_{MI}}\right]^{-1}$. When $T_{MI} < 0$, the variance estimator is recalculated as $(1 + 1/m)B$, and inferences are based on a t-distribution with $(m - 1)$ degrees of freedom. In this study, we choose $m = 12$ and $n = 3$.

SIMULATION STUDY

We now describe simulation studies to compare the proposed multiple imputation for external calibration method with existing methods.

Simulation Design and Parameter Settings

We assume a linear regression model for outcome Y on covariate X measured with error and covariate Z measured without error,

$$f(Y \mid X, Z, \psi) \sim N(\gamma_0 + \gamma_X X + \gamma_Z Z, \tau^2) \quad (2)$$

where $\psi = (\gamma_0, \gamma_X, \gamma_Z, \tau^2)$ denotes the vector of regression coefficients and residual variance. Our objective is inference for $\gamma = (\gamma_X, \gamma_Z)$. The covariates X and Z have mean 0, variance 1, and correlation $\rho = 0.3$ (low correlation) and $\rho = 0.6$ (high correlation). In the regression model, $\gamma_0 = 0$, $\gamma_Z = 0.4$, $\tau^2 = 1$, and $\gamma_X = 0.4$ and 1.2, corresponding to a small and large covariate effect, respectively. In the main study data, $n_{\text{main}} = 400$ observations are generated on Y, Z, and W, a surrogate measure for X related to X by the measurement error model

$$f(W \mid Y, X, Z) \sim N(\beta_0 + \beta_1 X, \sigma^2), \quad (3)$$

where $\beta_0 = 0$ and $\beta_1 = 1.1$, so that W is a linear biased surrogate for X, and σ^2 is set to 0.25, 0.5, and 0.75 to represent small, moderate, and large measurement errors, respectively. For the calibration data, $n_{\text{calib}} = 100$ observations on (X, W) are sampled from the measurement error model. We generate 1000 main and calibration data sets for each combination of parameter values.

Methods Compared

Multiple Imputation for External Calibration

We apply the proper version of the proposed method described earlier in the text and compare it with the following existing methods.

Naive Regression: The coefficients of the regression of Y on X and Z are computed by least squares on the main sample substituting $X = W$, that is, ignoring the measurement error in W.

Classical Calibration: We fit a linear regression curve of W on X based on the calibration data and then estimate the unknown value of X by $\hat{X}_{CC} = (W - \hat{\beta}_0)/\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope obtained from the regression of W on X. The classic calibration estimate of γ and associated standard error are obtained by least squares regression of Y on $\hat{X}_{CC}$ and Z, computed on the main study data.

Regression Prediction: We compute least squares estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ of the coefficients α_0 and α_1 of the linear regression of X on W using the calibration sample and then replace unknown values of X in the main sample by predictions $\hat{X}_{RP} = \hat{\alpha}_0 + \hat{\alpha}_1 W$. The coefficients γ are then estimated by least squares method from the regression of Y on $\hat{X}_{RP}$ and Z, based on the main sample. Standard errors for the estimate of γ can be found by bootstrap methods, if the calibration

data are available. The regression prediction method corresponds to the usual regression calibration method when there is no covariate *Z*.

Results

The results of the simulation studies are shown in Table 1. Inferences about γ_X and γ_Z, the regression coefficients of *X* and *Z* from the regression of *Y* on *X* and *Z*, are assessed for each method. Performance for a parameter γ is summarized using (*a*) the empirical bias, $bias(\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma)$; (*b*) the root mean square error, $RMSE(\hat{\gamma}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_i - \gamma)^2}$; and (*c*) the empirical noncoverage rate of the 95% confidence interval, that is the number of simulated data sets for which the 95% confidence intervals (CIs) do not contain the true parameter values. Given 1000 simulated data sets, the nominal value of noncoverage is equal to 50. We compare various methods with respect to the absolute bias, root mean square error, and empirical noncoverage of 95% confidence intervals. A good method is anticipated to have small absolute bias, low root mean square error, and a nominal level noncoverage.

TABLE 1. Empirical Bias, RMSE, and Noncoverage Rate (Noncov.) for the Estimates of γ_x and γ_z Based on 1000 Simulations

Simulation Parameters						*X*				*Z*			
γ_x	γ_z	β_1	σ^2	ρ		Naive	CC	RP	MI–EC	Naive	CC	RP	MI–EC
0.4	0.4	1.1	0.25	0.3	Bias	105	75	7	1	23	23	23	0
					RMSE	113	90	60	61	57	57	57	54
					Noncov.	664	347	44	28	66	66	61	32
0.4	0.4	1.1	0.5	0.3	Bias	151	126	10	3	38	38	38	1
					RMSE	156	134	68	71	65	65	65	56
					Noncov.	961	784	48	42	111	111	117	41
0.4	0.4	1.1	0.75	0.3	Bias	185	163	13	5	49	49	49	3
					RMSE	189	169	75	80	72	72	72	59
					Noncov.	998	952	53	45	153	153	151	36
0.4	0.4	1.1	0.25	0.6	Bias	27	99	36	2	60	60	60	1
					RMSE	135	113	76	76	85	85	85	69
					Noncov.	710	416	83	36	159	159	165	50
0.4	0.4	1.1	0.5	0.6	Bias	181	158	56	9	95	95	95	6
					RMSE	186	166	92	95	112	112	112	79
					Noncov.	981	868	124	47	350	350	354	48
0.4	0.4	1.1	0.75	0.6	Bias	217	198	70	18	119	119	119	15
					RMSE	221	204	105	122	133	133	133	100
					Noncov.	1000	975	145	49	499	499	506	34
1.2	0.4	1.1	0.25	0.3	Bias	312	223	18	5	67	67	67	10
					RMSE	316	233	85	88	88	88	88	61
					Noncov.	1000	949	63	37	210	210	217	59
1.2	0.4	1.1	0.5	0.3	Bias	451	375	29	12	113	113	113	5
					RMSE	453	383	111	119	128	128	128	73
					Noncov.	1000	999	72	38	418	418	428	53
1.2	0.4	1.1	0.75	0.3	Bias	552	487	36	17	147	147	147	9
					RMSE	554	493	132	144	159	159	159	86
					Noncov.	1000	1000	77	34	621	621	621	53
1.2	0.4	1.1	0.25	0.6	Bias	377	294	104	9	177	177	177	6
					RMSE	381	303	137	111	188	188	188	85
					Noncov.	1000	988	248	37	766	766	760	47
1.2	0.4	1.1	0.5	0.6	Bias	539	472	166	25	284	284	284	20
					RMSE	541	478	197	166	291	291	291	124
					Noncov.	1000	1000	392	34	987	987	987	45
1.2	0.4	1.1	0.75	0.6	Bias	647	591	206	44	355	355	355	37
					RMSE	649	596	239	213	362	362	362	166
					Noncov.	1000	1000	466	37	999	999	1000	47

The true value of γ_X is 0.4 or 1.2; the true value of γ_Z is 0.4.
All values are multiplied by 1000.
Naive indicates naive linear regression of *Y* on *W* and *Z*; CC, classical calibration; RP, regression prediction; MI-EC, multiple imputation for external calibration; RMSE, root mean square error.

We first consider inferences for γ_X. As expected, the naive regression estimate is attenuated towards 0, and empirical noncoverage rate of 95% confidence intervals seriously exceeds the nominal level in all simulation scenarios. The classical calibration method also performs very poorly, with substantial bias and high noncoverage rate, particularly when the measurement error is large. Regression prediction has small empirical bias when the correlation between X and Z is low, but it is biased with poor confidence coverage when the correlation is high, with the bias and noncoverage increasing with the size of covariate effect and the measurement error. Under all simulation scenarios considered here, the multiple imputation for external calibration method has small empirical bias, and confidence interval coverage close to the nominal level. The root mean square error of multiple imputation for external calibration is generally lower than that of regression prediction, but it is a little larger than that of regression prediction in some situations. We conjecture that the loss of precision of multiple imputation for external calibration relative to regression prediction arises because the former takes into account the correlation between X and Z, as is necessary to get consistent estimates. This conjecture was confirmed by assessing the performance of a modified version of multiple imputation for external calibration that assumes (like regression prediction) that X and Z are uncorrelated. This method had smaller root mean square error than regression prediction when the bias from assuming X and Z are uncorrelated is small.

The performance of inferences for the regression coefficient γ_Z of the covariate Z is also shown in Table 1. The estimates obtained by the naive method are biased, with bias increasing with the measurement error, the size of covariate effect, and the correlation between X and Z. The classical calibration and regression prediction methods also exhibit substantial bias, high root mean square error, and high noncoverage rates. In contrast, our multiple imputation for external calibration method performs well in all simulation scenarios.

The results presented in Table 1 are based upon the two-stage imputation parameter setting $(m, n) = (12, 3)$. We also examined the performance of our method under other combinations of m and n settings, namely (20, 3) and (12, 5). The results from the simulation study under those settings are close to those in Table 1, although the combination of (20, 3) results in a slightly higher than nominal coverage rate.

We also assessed through simulation the performance of the usual multiple imputation variance estimator when the calibration data are used for imputation but not for the postimputation analysis. We constructed 36 completed data sets (the same number as those generated using the two-stage imputation procedure) by using the standard multiple imputation method (ie, creating one imputed dataset per draw of model parameters). We then applied Rubin's standard multiple imputation combining rules,[13] without including the calibration data. We found that the Rubin's estimator was positively biased by 5%–30% over the settings we examined.

SENSITIVITY TO DEVIATIONS FROM MULTIVARIATE NORMALITY

Our proposed method is based on assumptions of nondifferent measurement error and multivariate normality. The nondifferential measurement error assumption is common and often reasonable for assay data,[11,17,21] and it is required to identify the parameters.[12] Therefore, we focus on a sensitivity analysis to evaluate the robustness of our method to violation of the normality assumption. We consider 2 forms of misspecification: the case where a binary covariate Z is misspecified as normal and the case where the covariate X is specified as normal when in fact it is log normal. As in the previous section, we examine the performance of our method and others under various choices of the measurement error variance (σ^2), correlation between X and Z (ρ), and the covariate effect of X (γ_X).

Simulation results for a binary covariate are presented in Table 2. We first generate (X, Z^*) from a bivariate normal distribution with mean 0, variance 1, and correlation ρ. The binary variable Z is then set to 1 if $Z^* \geq 0.8$, and to 0 otherwise; this setting results in the marginal probability $Pr.(Z = 1) = 0.2$, a moderately low value. We also examined several cut points other than 0.8, namely 0.5, 0.6, and 0.7, with results similar to those presented below. The surrogate W is generated from a simple unbiased measurement error model given as $W \mid X, Y, Z \sim N(X, \sigma^2)$. The outcome Y is related to X and Z by a linear regression model, as in (2). To be consistent with previous setup, the sample size of the main study is chosen to be 400, and the sample size of the calibration study is chosen as 100. Under each simulation setting, we generate 1000 simulated data sets.

Table 2 summarizes the empirical bias, root mean square error of the estimates for the regression parameters (γ_X, γ_Z), and the noncoverage rate of 95% confidence interval. In all simulation settings, the multiple imputation for external calibration method yields estimates with small empirical bias, and noncoverage rates close to the 50 nominal level. The method appears quite robust to this form of model misspecification.

Table 3 presents results for misspecification of a log-normal covariate. The true covariate X is generated from a log-normal distribution given as $X \sim LN(0, \omega^2)$, and Z is generated from a standard normal distribution with zero correlation between X and Z. We consider various degrees of skewness and heavy tails of the distribution of X by varying the parameter ω. We set ω equal to 0.25, 0.5, and 1 to represent low, moderate, and high skewness, respectively. Except for the scenario where the distribution of X is highly skewed and has a very heavy tail (and hence deviates seriously from normality), multiple imputation for external cal-

TABLE 2. Sensitivity to Multivariate Normality Assumption in the Binary Case

Simulation Parameters						*X*				*Z*			
γ_x	γ_z	β_1	σ^2	ρ		Naive	CC	RP	MI–EC	Naive	CC	RP	MI–EC
0.4	0.4	1	0.25	0.3	Bias	84	84	4	1	45	45	45	0
					RMSE	95	97	61	62	136	136	136	131
					Noncov.	439	452	49	31	64	64	65	30
0.4	0.4	1	0.5	0.3	Bias	139	138	5	3	74	74	74	2
					RMSE	145	146	70	73	149	149	149	137
					Noncov.	903	857	45	42	80	80	87	36
0.4	0.4	1	0.25	0.6	Bias	95	95	17	2	100	100	100	1
					RMSE	106	108	67	69	169	169	169	147
					Noncov.	487	504	59	43	109	109	111	38
0.4	0.4	1	0.5	0.6	Bias	153	153	27	5	161	161	161	7
					RMSE	159	160	77	81	210	210	211	160
					Noncov.	930	886	69	52	212	212	212	40
1.2	0.4	1	0.25	0.3	Bias	250	250	9	5	133	133	134	1
					RMSE	255	260	89	93	195	195	196	153
					Noncov.	996	976	60	39	145	145	148	71
1.2	0.4	1	0.5	0.3	Bias	414	413	13	12	220	220	220	7
					RMSE	417	420	120	125	267	267	267	182
					Noncov.	1000	1000	61	39	309	309	308	73
1.2	0.4	1	0.25	0.6	Bias	282	282	49	7	298	298	298	7
					RMSE	287	291	103	103	332	332	332	176
					Noncov.	998	986	100	37	499	499	509	52
1.2	0.4	1	0.5	0.6	Bias	457	456	77	18	480	480	480	25
					RMSE	460	462	139	141	504	504	504	234
					Noncov.	1000	1000	143	28	860	860	856	44

The table shows empirical bias, root mean square error, and noncoverage rate (Noncov.) for estimates of regression parameters (γ_X, γ_Z).
All values are multiplied by 1000.
Naive, naive linear regression of *Y* on *W* and *Z*; CC, classical calibration; RP, regression prediction; MI-EC, multiple imputation for external calibration.

ibration generally outperforms the other methods with respect to bias and confidence coverage. This finding further suggests robustness of our method, except when the distribution of *X* is highly skewed.

APPLICATION TO THE MICHIGAN BONE HEALTH AND METABOLISM STUDY

We illustrate the proposed method in this section with data from the Michigan Bone Health and Metabolism Study. One of the goals of this study is to assess the association between serum reproductive hormone concentrations and bone mineral density loss in midlife women. We consider here the relationship between sex hormone binding globulin concentration (*X*), which is the primary plasma transport protein for sex hormones, and bone mineral density loss (*Y*), adjusting for covariates Z = age and BMI. For a variety of reasons, including assay imprecision, sex hormone binding globulin concentration has substantial measurement error; what is measured is a noisy version of *X*, namely *W* in the notation defined earlier in the text. The main study included measures of *W*, *Z*, and *Y* in 81 white women, aged 44–64 years, from the Michigan Bone Health and Metabolism Study cohort in 2008. The calibration data consisted of duplicate assay measures *W* at 4 true concentrations *X* of SHBG, from a competitive immunoassay run on the Bayer Diagnostic ACS: 180 automated analyzer (Bayer Diagnostics Corp, Tarrytown, NY) using chemiluminescent technology. The scatter plot of the calibration data in Figure 2 shows clear evidence of measurement error.

We estimated parameters in the linear regression of bone mineral density loss on the logarithm of sex hormone binding globulin concentration, age, and BMI by 5 different methods: the "naive" analysis (where SHBG concentrations are represented by assay measures), classical calibration, 2 versions of regression calibration (namely RP1, with standard errors based on the bootstrap, and RP2, with naive standard errors that ignore the measurement error in the predictions of *X*), and multiple imputation for external calibration. Table 4 presents the estimates and associated standard errors for the regression coefficients of log sex hormone binding globulin concentration, age, and BMI. The naive analysis yields a positive effect of sex hormone binding globulin concentration on bone mineral density loss, 0.0516 (0.0269). Classical calibration, RP1, RP2, and multiple imputation for external calibration result in stronger

TABLE 3. Sensitivity to Multivariate Normality Assumption in the Skew Case

Simulation Parameters			*X*				*Z*			
ϕ	Ratio		Naive	CC	RP	MI–EC	Naive	CC	RP	MI–EC
0.25	0.25	Bias	82	82	0	0	0	0	0	0
		RMSE	188	190	218	214	49	49	50	49
		Noncov.	70	72	56	49	55	55	52	18
	0.5	Bias	135	135	3	2	0	0	0	0
		RMSE	204	207	239	238	49	49	50	49
		Noncov.	131	143	49	60	54	54	51	18
0.5	0.25	Bias	81	81	8	3	0	0	0	0
		RMSE	110	111	100	98	49	49	50	49
		Noncov.	187	200	51	48	52	52	58	19
	0.5	Bias	134	134	15	9	0	0	0	0
		RMSE	116	115	49	49	50	50		
		Noncov.	509	527	58	52	57	22		
1	0.25	Bias	86	86	35	25	1	1	1	1
		RMSE	92	93	81	78	52	52	52	53
		Noncov.	937	877	56	200	49	49	51	42
	0.5	Bias	141	140	69	45	1	1	1	1
		RMSE	145	146	131	126	54	54	54	59
		Noncov.	998	986	45	256	44	50	50	49

The table shows empirical bias, root mean square error, and noncoverage Rate (Noncov.) for estimates of regression parameters (γ_X, γ_Z).
All values are multiplied by 1000.
Naive indicates naive linear regression of *Y* on *W* and *Z*; CC, classical calibration; RP, regression prediction; MI-EC, multiple imputation for external calibration.

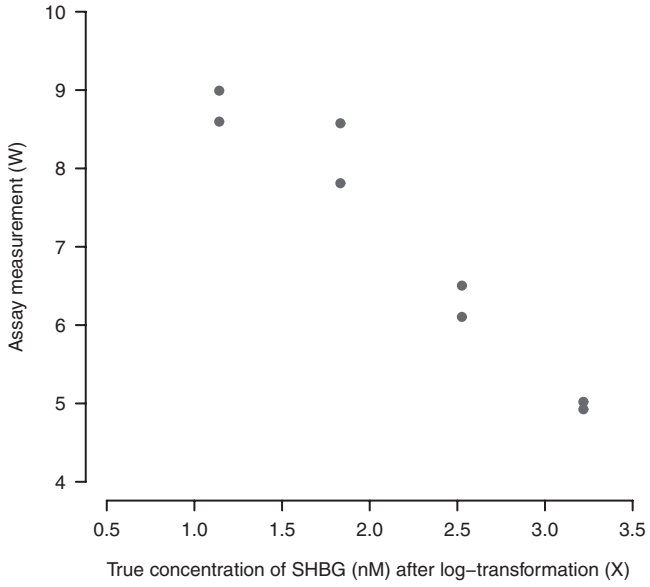

FIGURE 2. Calibration data for sex hormone–binding globulin (SHBG) from a competitive immunoassay. The graph shows duplicate assay measurements (W) versus log-transformed known (true) concentrations of sex hormone binding globulin (X). The assay measurement can be viewed as error-contaminated version of true concentration of sex hormone binding globulin. The calibration data provided an external source of information on the magnitude of measurement error when measuring true concentrations using immunoassay techniques.

TABLE 4. Application to the MBHMS: Parameter Estimates in Linear Regression of BMD on the Logarithm of SHBG Concentration, With and Without Adjustment for Covariates Age and BMI

	log SHBG		Age		BMI	
Methods	Estimates	SE	Estimates	SE	Estimates	SE
With covariate adjustment						
Naive	0.0516	0.0269	−0.0092	0.0048	0.0021	0.0030
CC	−0.0995	0.0518	−0.0092	0.0048	0.0021	0.0030
RP1	−0.1085	0.0610	−0.0086	0.0055	0.0026	0.0027
RP2	−0.1054	0.0549	−0.0092	0.0048	0.0021	0.0030
MI–EC	−0.1072	0.0587	−0.0095	0.0048	0.0012	0.0031
Without covariate adjustment						
Naive	0.0538	0.0258				
CC	−0.1035	0.0498				
RP1	−0.1224	0.0565				
RP2	−0.1096	0.0528				
MI–EC	−0.1160	0.0563				

Naive indicates naive linear regression; CC, classical calibration; RP, regression prediction; MI-EC, multiple imputation for external calibration; MBHMS, the Michigan Bone Health and Metabolism Study; BMD, bone mineral density; SHBG, sex hormone–binding globulin.

negative estimates of the regression coefficient of log sex hormone binding globulin concentration: −0.0995(0.0518), −0.1085(0.0610), −0.1054(0.0549), and −0.1072(0.0587), respectively. In contrast, the multiple imputation for external

calibration estimate is approximately 8% larger than the classical calibration estimate, with a standard error that is almost 1.3 times larger than that of the classical calibration estimate. Here, we do not see much difference between RP1 and multiple imputation for external calibration, although we note that RP1 method requires the full calibration data set, not just summary statistics. The RP2 estimate has smaller standard error than the RP1 estimate, as expected because unlike RP1, it fails to account the uncertainty due to measurement error. Use of RP2 will tend to yield confidence intervals that are too narrow. The naive, classical calibration, and RP1 estimates of the coefficient for BMI are approximately twice as large as the multiple imputation for external calibration estimate, showing that measurement-error adjustment affects the estimated coefficients of other covariates measured without error.

Table 4 also presents estimates of the coefficient for log sex hormone binding globulin concentration from a simple regression of bone mineral density loss on log sex hormone binding globulin concentration without including age and BMI as covariates. These estimates are slightly larger than those obtained from the multivariate regression with age and BMI as covariates.

CONCLUSIONS AND DISCUSSION

Our simulations are consistent with previous findings that the classical calibration method for incorporating information from an external calibration sample yields biased estimates in the regression setting when measurement error is substantial. The higher settings of measurement error in our simulations may exceed that found in many real settings. We suspect that assays with high measurement error are less likely to be implemented in practice, as the widespread use of the classical calibration method inhibits the ability to make use of them, although they may still provide useful information. A method that works only when the problem it is trying to solve is very minor is not a good method.

We propose a simple multiple imputation method that corrects for covariate measurement error in regression analysis, when the calibration data provide information only about X and W. Our simulation studies suggest that our method is markedly superior to existing methods for adjusting for covariate measurement error, eliminating bias, and providing confidence interval coverage close to nominal levels. Its superiority is most pronounced when the covariates X and Z are highly correlated, the covariate effect is large, or the measurement error is large. By general theoretical properties of multiple imputation, inferences for other parameters of the joint distribution of the variables are also valid, under the stated assumptions (ie, the assumptions of multivariate normality and nondifferential measurement error).

The proposed procedure is simple and fast to compute, and requires only simple summary statistics for the joint distribution of (X, W) in the calibration sample. Hence, it is a viable method for external calibration data, where the microdata from the calibration sample are typically not available to the analyst. Of course, the method requires that summary statistics from the calibration sample be made available, which is not yet common in practice. Without these statistics, we know of no valid method of correcting for measurement error in the regression setting, and we do not believe such a method exists.

When the calibration data are not included in the postimputation analysis, we use the two-stage imputation procedure and apply Reiter's multiple-imputation combining rules for valid statistical inference. Reiter's variance estimator T_{MI} could be negative, particularly when measurement error is substantial. In general, we found that negative values of T_{MI} can be avoided by making the m and n large. For fixed M, where $M = m \times n$, making m large is more likely to reduce the chance of negative values than making n large.

Our method rests primarily on the assumptions of nondifferential measurement error, equivalence of the distribution of (U, $X \mid$ W) in the calibration and study samples, and normality of this distribution. The first 2 assumptions are crucial and necessary to identify the parameters, and our simulations suggest some degree of robustness to the normality assumption. This assumption could be relaxed, but at the expense of requiring more information from the calibration sample. This is a topic for future research. We also assume here that X and W are scalar; in the future, we plan to extend the proposed method to handle more than one covariate subject to measurement error.

APPENDIX

Multiple Imputation for External Calibration Algorithm

We first describe the improper method with parameters estimated by maximum likelihood. We then discuss the adopted method where parameters are drawn from their posterior distribution.

The maximum likelihood estimate $\hat{\theta}$ is computed as follows:

Step (1): Let $\theta = (\theta_1, \theta_2, \sigma_{ux \cdot w})$, where θ_1 represents parameters of the normal distribution of X given W, θ_2 represents parameters of the normal distribution of U given W, and $\sigma_{ux \cdot w}$ represents the set of p partial covariances between U and X given W. Estimate θ_1 by $\widehat{\theta}_1$, the maximum likelihood estimates based on the calibration sample on (X, W), and θ_2 by $\widehat{\theta}_2$, the maximum likelihood estimates based on the main study sample on (U, W). These are the normal linear regression maximum likelihood estimates for complete data and involve standard least squares calculations. Also note that $\widehat{\theta}_2$ can be computed from summary statistics on the calibration sample, namely the sample size, sample mean, and sum of squares and cross products matrix of X and W.

Step (2): Estimate

$$\hat{\sigma}_{ux \cdot w} = \hat{\beta}_{uw \cdot w} \hat{\sigma}_{xx \cdot w} / \hat{\beta}_{xw \cdot w},$$

where $\hat{\beta}_{uw \cdot w}$ is the $(p \times 1)$ vector of regression coefficients of U on W, estimated from the main sample, and $\hat{\beta}_{xw \cdot w}$ and $\hat{\sigma}_{xx \cdot w}$ are the regression coefficient of W and residual variance from regression of X on W, estimated from the calibration sample. This expression follows because, from properties of the multivariate normal distribution, $\beta_{uw \cdot w} - \sigma_{ux \cdot w} \beta_{xw \cdot w} / \sigma_{xx \cdot w}$ equals the set of regression coefficients of W in the regression of U on W and X, which are zero because of the nondifferential measurement error assumption.

Step (3): The maximum likelihood estimates of the parameters of the distribution of $(U, X) = (Y, Z, X)$ given W are fully specified by the estimates in Steps (1) and (2). In fact, the method is maximum likelihood because the number of parameter restrictions from the nondifferential measurement error assumption, namely p, is the same as the number of parameters in $\sigma_{ux \cdot w}$ that are not estimable from the main and calibration samples—the model is technically "just identified."[12] The parameter ϕ of the regression of X on Y and Z is a vector function $\phi(\theta_1, \theta_2, \sigma_{ux \cdot w})$ of the parameters $(\theta_1, \theta_2, \sigma_{ux \cdot w})$. The maximum likelihood estimate of ϕ is then $\hat{\phi} = \phi(\hat{\theta}_1, \hat{\theta}_2, \hat{\sigma}_{xx \cdot w})$, obtained by substituting maximum likelihood estimates of $(\theta_1, \theta_2, \sigma_{ux \cdot w})$ in this function. The details of this transformation are discussed by Little and Rubin.[12] Computation is straightforward using the SWEEP operator,[12] which facilitates switching between parameters of different regressions derived from the multivariate normal distribution.

This completes the description of the maximum likelihood algorithm, except for one minor caveat. The estimate of the residual variance of X given (Y, Z, U) could be negative, given the fact that estimates are being combined from 2 samples. If this happens, the residual variance should be set to zero. This is unlikely to happen unless X and W are weakly correlated, in which case, the calibration data have limited utility.

As noted earlier in the text, the imputations based on this procedure have the limitation that they do not reflect uncertainty in the maximum likelihood estimates of ϕ. Fortunately, it is relatively easy to overcome this limitation by replacing maximum likelihood estimates $\hat{\phi}$ of the parameters ϕ for the d^{th} imputed data set by a draw $\phi^{(d)}$ from the posterior distribution of ϕ. A noninformative Jeffreys prior is assumed for the parameter (θ_1, θ_2). Then the maximum likelihood estimates $(\hat{\theta}_1, \hat{\theta}_2)$ in Step (1) are replaced by draws $(\theta_1^{(d)}, \theta_2^{(d)})$ from their complete-data posterior distributions based on the calibration and main study samples, respectively. Draws from these posterior distributions are easily computed using χ^2 and normal deviates, as described in Little and Rubin.[12] Steps (2) and (3) are then as given earlier in the text, except that draws of $\sigma_{ux \cdot w}^{(d)}$, $\phi^{(d)}$ for $\sigma_{ux \cdot w}$ and ϕ are created using the draws $(\theta_1^{(d)}, \theta_2^{(d)})$ rather than $(\hat{\theta}_1, \hat{\theta}_2)$.

ACKNOWLEDGMENTS

We thank the referees for helpful comments.

REFERENCES

1. Morgan TM, Elashoff RM. Effect of covariate measurement error in randomized clinical trials. *Stat Med*. 1987;6:31–41.
2. Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. *Stat Med*. 1993; 12:1703–1722.
3. Zidek JV, Wong H, Le ND, Burnett R. Causality, measurement error and multicollinearity in epidemiology. *Environmetrics*. 1996;7:441–451.
4. Fung KY, Krewski D. On measurement error adjustment methods in Poisson regression. *Environmetrics*. 1999;10:213–224.
5. Sarkar S, Qu Y. Quantifying the treatment effect explained by markers in the presence of measurement error. *Stat Med*. 2007;26:1955–1963.
6. Jurek AM, Maldonado G, Greenland S, Church TR. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Eur J Epidemiol*. 2006;21:871–876.
7. Higgins KM, Davidian M, Chew G, Burge H. The effect of serial dilution error on calibration inference in immunoassay. *Biometrics*. 1998;54:19–32.
8. Browne RW, Whitcomb BW. Procedures for determination of detection limits: application to high-performance liquid chromatography analysis of fat-soluble vitamins in human serum. *Epidemiology*. 2010;21:S4–S9.
9. Guo Y, Little RJ. Regression analysis with covariates that have heteroscedastic measurement error. *Stat Med*. 2011;30:2278–2294.
10. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *J Am Stat Assoc*. 1990;85:652–663.
11. 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.
12. Little RJ, Rubin DB. *Statistical Analysis With Missing Data.* 2nd ed. Hoboken, NJ: Wiley-Interscience; 2002.
13. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.; 1987.
14. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol*. 2006;35:1074–1081.
15. Raghunathan TE. Combining information from multiple surveys for assessing health disparities. *Allgemeines Stat Arch*. 2006;90:515–526.
16. He Y, Zaslavsky AM. Combining information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies. *Biometrics*. 2009;65:946–952.
17. Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Stat Med*. 2008;27: 5195–5216.
18. Sowers MR, Jannausch M, McConnell D, et al. Hormone predictors of bone mineral density changes during the menopausal transition. *J Clin Endocrinol Metab*. 2006;91:1261–1267.
19. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* 3rd ed. New York: Chapman Hall/CRC; 2006.
20. Reiter JP. Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*. 2008;95:933–946.
21. Guolo A, Brazzale AR. A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. *Stat Med*. 2008;27:3755–3775.

