Research Article

Statistics in Medicine

Received 18 February 2010, Accepted 12 March 2011 Published online 17 May 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4261

Regression analysis with covariates that have heteroscedastic measurement error

Ying Guo*† and Roderick J. Little

We consider the estimation of the regression of an outcome Y on a covariate X, where X is unobserved, but a variable W that measures X with error is observed. A calibration sample that measures pairs of values of X and W is also available; we consider calibration samples where Y is measured (internal calibration) and not measured (external calibration). One common approach for measurement error correction is Regression Calibration (RC), which substitutes the unknown values of X by predictions from the regression of X on W estimated from the calibration sample. An alternative approach is to multiply impute the missing values of X given Y and W based on an imputation model, and then use multiple imputation (MI) combining rules for inferences. Most of current work assumes that the measurement error of W has a constant variance, whereas in many situations, the variance varies as a function of X. We consider extensions of the RC and MI methods that allow for heteroscedastic measurement error, and compare them by simulation. The MI method is shown to provide better inferences in this setting. We also illustrate the proposed methods using a data set from the BioCycle study.

Keywords: calibration sample; data argumentation; heteroscedastic; measurement error; multiple imputation; regression calibration

1. Introduction

Measurement error is common in many empirical studies, arising from assay or instrumental error, biological variation, or errors in questionnaire-based self-report data. It is well known that in a regression analysis the estimated effect of a predictor variable may be attenuated when it is measured with substantial error. In particular, Kipnis *et al.* [1] find that measurement error in dietary intake assessment by using the Food Frequency Questionnaires (FFQ) leads to severe attenuation in the estimate of disease relative risk in a biomarker study. Cotton *et al.* [2] show that measurement error can reduce the chance of accurate diagnosis of appropriate educational placement for children with reading difficulties. Wannemuehler *et al.* [3] indicate that the impact of measurement error can be substantial in the assessment of the association between pollutant exposure and a health outcome, using surrogates for unobserved measurements of ambient concentrations.

Most of the research on measurement error in covariates assumes that the variance of measurement error is constant. However, the variance of measurement error often increases with the true underlying value, as evidenced by the fact that the limit of quantification in assays is often defined in terms of the coefficient of variation rather than the standard deviation. We consider here methods for correcting for heteroscedastic covariate measurement error. Our motivating example is provided by the BioCycle study, a study where one of the primary goals is to investigate the association between fat-soluble vitamins (e.g. β-carotene) and progesterone in human serum [4]. The fat-soluble vitamins are measured with error using high-performance liquid chromatography (HPLC). Guo *et al.* [5] model calibration data on eight fat-soluble vitamin analytes with measurement variance $\sigma^2 g^2(X, \alpha)$, where X is the true value, and $g(X, \alpha) = X^\alpha$, a power function of X. They find that the constant variance assumption ($\alpha = 0$) is clearly violated, with estimates of α ranging from 0.5 to 0.8.

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, U.S.A.
**Correspondence to: Ying Guo, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, U.S.A.*
†E-mail: guoy@umich.edu

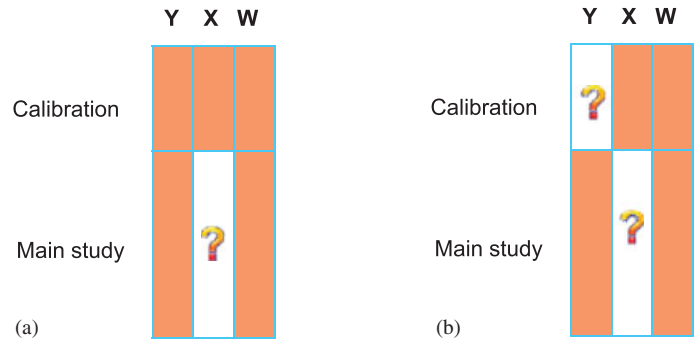

Figure 1. Calibration/main study design: (a) internal and (b) external.

We consider data in the form displayed in Figure 1, where Y denotes a response variable, X denotes the covariate of interest, W denotes the error-prone measurement of X, and question marks denote unobserved values. The main study data consist of a sample of independent and identically distributed observations on (Y, W). The calibration data consist of a sample of independent and identically distributed observations on (X, W). Information about the measurement error is contained in calibration data with measured and true values of the covariate both recorded. We call the calibration data *internal* when they are a random sample from the main study, so they also contain observations of Y, as in Figure 1(a). We call the calibration data *external* when they are from another source, and information of Y is not available, as in Figure 1(b). We consider inference for the parameters of the regression of Y on X. The common case where there are additional error-free covariates is discussed in Section 8.

Comprehensive reviews of statistical methods for adjusting the measurement error include Fuller [6] for linear models and Carroll *et al.* [7] for nonlinear models. One commonly used and simple method is regression calibration (RC) [8]. The unobserved value X is imputed by the expected value of X given W, with coefficients estimated from the calibration data, and the regression of Y on X is then estimated using the filled-in data in the main study. A related method is moment reconstruction, where imputed values are constructed to match the first two moments of Y and X [9]. This method is equivalent to RC in the linear regression case. For the case of internal calibration data, the estimate from RC can be combined with the estimate of the regression of Y and X computed directly from the calibration data, weighting the estimates from two sources according to their precisions. This method is known as efficient RC (ERC); see [10].

An alternative approach is to use imputation or multiple imputation (MI) methods from the missing data literature to fill in the true values. Cole *et al.* [11] consider the MI method to remove the bias in the estimation of the hazard ratio for chronic kidney disease due to mismeasured covariates in a prospective cohort study. Schenker *et al.* [12] describe MI to correct for measurement error of self-report data on health conditions in large-scale population surveys. Other Bayesian approaches for covariate measurement error are given in [13–17].

Freedman *et al.* [18] evaluate the performance of a number of these methods by simulation, for the case of internal calibration data. Several different scenarios are considered, including different choices of the measurement error variance and the strength of the response–covariate relationship. Data are simulated assuming non-differential measurement error (NDME), meaning that Y and W are independent given X. Their findings suggest that the ERC is the preferred method. However, we note that unlike ERC, the version of MI used in this simulation does not exploit the NDME assumption. The MI methods described in this article are more efficient since they are based on an imputation model that makes the NDME assumption.

In addition, Freedman *et al.* [18] assume that the variance of measurement error is constant, and do not assess methods when the measurement error is heteroscedastic. In that situation, existing error correction methods yield biased estimates, as our simulations demonstrate. Spiegelman *et al.* [19] propose Taylor series-based modifications of RC for heteroscedastic measurement error, under the assumption that $g(X, \alpha) = g(X)$ is known. We propose extensions of the methods compared in [18] to correct for heteroscedastic measurement error. We compare these methods through a simulation study, concluding that MI is the best of the methods compared in this setting.

The outline of the rest of this article is as follows. In Section 2, we specify models for the calibration data and main study data. In Section 3, we describe various measurement error correction methods.

We propose a simple extension of standard RC to deal with nonconstant variance. We also propose MI methods under both constant and nonconstant measurement variance, with and without the NDME assumption. MI methods are developed for both internal and external calibration designs. Our MI methods are Bayesian. The unobserved values of covariate X are replaced by draws generated using data augmentation [20], and (in the nonconstant variance case) a Metropolis–Hastings (MH) algorithm [21]. A simplified approximate method that avoids the MH step is also developed. In Section 4, a simulation study is described, considering both constant and nonconstant measurement error variances, and both internal and external calibration study designs. Results from simulations are reported in Section 5. Sensitivity analysis is presented to examine the performance of the MI methods to misspecification of the prior distribution of X in Section 6. In Section 7, we illustrate the use of proposed methods on a real data example from the BioCycle study, where the effect of oxidative stress on female fecundity and fertility is investigated. In Section 8, we conclude with a discussion of the results and extensions of the proposed methods.

2. Models

Measurement error adjustments require an error model linking the true variable X to the surrogate measure W, which requires careful consideration in the context of the specific application [22, 23]. The classical measurement error model assumes that

$$W = X + \xi \tag{1}$$

where ξ is a random error with mean zero and constant variance [24–26]. In some epidemiological studies, W and X are transformed by taking logarithms, that is: $\log(W) \sim \mathrm{N}(\log(X), \sigma^2)$ [27]. In our work, we consider a linear mean function and heteroscedastic measurement error, specifically:

$$p(W|X, \theta) \sim \mathrm{N}(\beta_0 + \beta_1 X, \sigma^2 X^{2\alpha}) \tag{2}$$

where $\theta = (\beta_0, \beta_1, \alpha, \sigma^2)$, and the parameter α models heteroscedasticity. The measurement error variance is constant when $\alpha = 0$. In the main study, we assume a linear regression model of Y on X:

$$p(Y|X, \psi) \sim \mathrm{N}(\gamma_0 + \gamma_X X, \tau^2) \tag{3}$$

where $\psi = (\gamma_0, \gamma_X, \tau^2)$, although more generally nonlinear relationships between Y and X can be modeled.

We assume that (Y, W) given X are bivariate normal with constant correlation ρ. Under the NDME assumption that Y and W are independent given X, $\rho = 0$ [18]. This assumption is often reasonable when measurement error arises from bioassay techniques or laboratory experiments. NDME is less reasonable in retrospective case–control studies, where the disease status of subjects is known and the data about their exposure to risk factors are collected retrospectively, since recall error of past exposures is often thought to be more likely for cases than for controls (e.g. mothers of babies with a deformity may, on the average, have a different recall error about their early pregnancy drug intake than mothers of normal infants).

Further, we assume that the error model of W given X, and the regression model of Y given X hold with the same parameter values in both the main study sample and the calibration sample, when using external calibration data to assess measurement error. This assumption naturally holds for the internal calibration sample since it is a subsample of the main study sample. Therefore, the final analysis can be applied to the completed data including both samples. The situation where the model for measurement error in the external calibration study may differ from that in the main study is discussed later in this article.

3. Methods

3.1. Conventional approach

The conventional approach (CA) fits an appropriate regression curve of W on X to the calibration data and estimates the true value of X using the value on the predicted calibration curve [28–30]. For

example, assuming a linear relationship between the true and measured values, the estimate of the true value given the measured value W is $\hat{X}_{\mathrm{CA}}=(W-\hat{\beta}_0)/\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope obtained from the regression of W on X using the calibration data. The estimate $\hat{X}_{\mathrm{CA}}$ is then substituted for the unknown X in the main study data, and the regression model (3) is fitted to the data, yielding the CA estimate $\hat{\gamma}_{X,\mathrm{CA}}$. The CA approach is biased for the regression coefficients but is nevertheless widely used in practice.

3.2. Regression calibration

RC estimates the regression of X on W using the calibration data, and then substitutes the unknown values X in the main study with predictions $\hat{X}_{\mathrm{RC}}=E(X|W)$ from this regression. The RC estimate $\hat{\gamma}_{X,\mathrm{RC}}$ is then obtained by regressing Y on $\hat{X}_{RC}$. The RC method rests on the assumption that measurement error is non-differential. If this assumption is violated, the RC estimates are biased.

The standard error of the RC estimate can be estimated using asymptotic calculations [10], or by bootstrapping the main and calibration samples. We create bootstrap samples from the calibration data and the main study data separately, and then combine them to compute RC estimates of the regression parameters. This procedure is repeated $B=200$ times. The sample variance of the resulting B estimates is used to estimate the variance.

When the calibration data are internal, two estimates of the regression coefficient are available, the RC estimate $\hat{\gamma}_{X,\mathrm{RC}}$, and the least squares estimate $\hat{\gamma}_{X,\mathrm{LSCalib}}$ from fitting the linear regression model (3) to the calibration sample data on (Y, X). The ERC estimate is the inverse-variance-weighted average of these two estimates,

$$\hat{\gamma}_{X,\mathrm{ERC}}=w_{\mathrm{RC}}\hat{\gamma}_{X,\mathrm{RC}}+(1-w_{\mathrm{RC}})\hat{\gamma}_{X,\mathrm{LSCalib}}$$

with weight

$$w_{\mathrm{RC}}=\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{RC}})^{-1}[\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{RC}})^{-1}+\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{LSCalib}})^{-1}]^{-1}$$

where $\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{RC}})$ and $\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{LSCalib}})$ are the estimated variances of $\hat{\gamma}_{X,\mathrm{RC}}$ and $\hat{\gamma}_{X,\mathrm{LSCalib}}$, respectively. The variance of $\hat{\gamma}_{X,\mathrm{ERC}}$ is computed approximately as $\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{ERC}})=[\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{RC}})^{-1}+\widehat{\mathrm{var}}(\hat{\gamma}_{X,\mathrm{LSCalib}})^{-1}]^{-1}$. ERC is more efficient than RC, particularly when the calibration data set is large.

We propose a modified version of RC, weighted RC (WRC), for situations where measurement error is heteroscedastic. This method estimates the parameters of the regression model of X on W by weighted least squares. Specifically, we assume the regression of X on W can be approximated by the weighted regression model

$$p(X|W,\eta,\pi,\lambda)\sim\mathrm{N}(\eta_0+\eta_1 W,\pi^2 W^{2\lambda}) \tag{4}$$

The estimate $\hat{\gamma}_{X,\mathrm{WRC}}$ is obtained by

- Estimating λ as the slope of a simple regression of the logarithm of squared residuals of the regression of X on W on the logarithm of squared W using the calibration data.
- Estimating $\hat{\eta}_0$ and $\hat{\eta}_1$ by weighted least squares.
- Substituting unknown values X in the main study with the prediction, i.e. $\hat{X}_{\mathrm{WRC}}=\hat{\eta}_0+\hat{\eta}_1 W$.
- Estimating the coefficient $\gamma_{X,\mathrm{WRC}}$ of the regression of Y on $\hat{X}_{\mathrm{WRC}}$ from the main study data.

The associated standard error of the WRC estimate can be estimated using the bootstrapping approach mentioned above. We estimate λ using a simple regression approach. This approach is easy to implement, and yields estimates of λ close to the ML estimates (using an iterated conditional modes algorithm), as demonstrated when applied to the BioCycle study data in our previous study [5].

We can also modify the ERC estimate to account for heteroscedastic measurement error by replacing $\hat{\gamma}_{X,\mathrm{RC}}$ with $\hat{\gamma}_{X,\mathrm{WRC}}$. We call this the weighted ERC (WERC) estimate.

3.3. Multiple imputation

We now develop MI methods based on a fully Bayesian model for the joint distribution of X, W and Y, which we write as $p(X, W, Y)$. These methods all create MIs of the missing values – for the internal calibration design (Figure 1(a)), the missing values of X in the main sample, and for the external calibration design (Figure 1(b)), the missing values of X in the main sample and the missing values

of Y in the calibration sample. In the work described in this article, we combine both the calibration sample and the main study sample in the final (post-imputation) analysis to provide inferences for the regression of Y on X. The MI estimates and associated standard errors are then obtained by applying standard MI combining rules to M multiply-imputed data sets including both samples [31]. Specifically, the MI estimate of γ_X is

$$\hat{\gamma}_{X,\mathrm{MI}} = \frac{1}{m}\sum_{m=1}^{M} \hat{\gamma}_{X,\mathrm{MI}}^{(m)}$$

and the corresponding variance is

$$\mathrm{Var}(\hat{\gamma}_{X,MI}) = \bar{W} + B \times (M+1)/M$$

where $\hat{\gamma}_{X,\mathrm{MI}}^{(m)}$ is the estimate of the coefficient of $X_{\mathrm{MI}}^{(m)}$ in the regression of Y on $X_{\mathrm{MI}}^{(m)}$ in the mth imputed data set; B is the between-imputation variance, calculated as $B = \sum_{m=1}^{M}(\hat{\gamma}_{X,\mathrm{MI}}^{(m)} - \hat{\gamma}_{X,\mathrm{MI}})^2/(M-1)$; $\bar{W}$ is the average of the within-imputation variance, calculated as $\bar{W} = \sum_{m=1}^{M} \mathrm{var}(\hat{\gamma}_{X,\mathrm{MI}}^{(m)})/M$; and $\mathrm{var}(\hat{\gamma}_{X,\mathrm{MI}}^{(m)})$ is the standard variance estimator obtained from the mth imputed data set. We apply this method to $M = 16$ multiply-imputed data sets.

The form of the MI method depends on the assumption made about α and ρ. When $\alpha = 0$, that is, the measurement error variance is constant, MI can be performed easily using standard Bayesian techniques for normal data described by Little and Rubin [32]. We assume uniform priors for the location parameters and log variances. The methods are labeled with a '0' to indicate the assumed value of α. For internal calibration data (Figure 1(a)), the imputation of X can be created assuming NDME ($\rho = 0$)—we label this method MIND0, or not assuming NDME ($\rho \neq 0$)—we label this method MI0. The MI0 method is assessed in the simulation study of [18], but we also consider MIND0 to assess the gain of efficiency from basing imputations on the NDME assumption. For the external calibration design, the parameter ρ is not identified, so we only consider the MIND0 method that assumes NDME ($\rho = 0$).

For the cases where $\alpha \neq 0$ (i.e. the measurement error variance is not constant), we consider two approaches for estimating α. One approach is to generate draws of this parameter from its posterior distribution, through a Metropolis–Hastings step. We describe here MI under the NDME assumption, where the joint distribution of W and Y given X can be factored as:

$$p(W, Y|X, \gamma, \tau, \beta, \sigma, \alpha) = p(W|X, \beta, \sigma, \alpha)p(Y|X, \gamma, \tau)$$

We add prior distributions for the marginal distribution of X and the parameters $(\gamma, \tau, \beta, \sigma, \alpha)$ to complete the fully Bayesian specification. Specifically, we assume

$$p(X, \gamma, \tau, \beta, \sigma, \alpha) = p(X)p(\gamma, \tau, \beta, \sigma, \alpha)$$

where the prior distribution of X is a normal distribution with mean μ_x and variance σ_x^2, and the prior distribution of parameters $(\gamma, \tau, \beta, \sigma, \alpha)$ is a noninformative prior

$$p(\gamma, \log\tau, \beta, \log\sigma, \alpha) = \text{const.}, \quad -2 < \alpha < 2$$

where the range $(-2, 2)$ for α includes values of that parameter thought likely to be of interest; the proper prior distribution for α is to ensure a proper posterior distribution [5]. In this article, we consider a hierarchical normal structure for X, and assume a noninformative prior distribution for hyperparameters μ_x and σ_x^2, with $p(\mu_x, \log\sigma_x) = \text{const.}$, letting the posterior inferences be dominated by the observed data. Other choices of the prior distribution for X are discussed in the concluding section.

Draws can be conveniently computed using the data augmentation algorithm, which iteratively imputes missing values given observed data and draws of the parameters (*the imputation step*), and then draws parameters of the model from their posterior distribution given imputed values and observed data (*the posterior step*). A Metropolis–Hastings step is required to generate draws of X. Details are given in Appendix A. We label MI inferences from this algorithm MINDα.

We also develop a simpler version of MI that does not require the MH step, and can be viewed as a MI analog of the WRC method. The measurement error model is reformulated as the model (4), and the estimate $\hat{\lambda}$ of the parameter λ is substituted. For known λ, MI can be performed easily using standard

Bayesian approaches for normal data. The estimate of λ can be computed using a simple regression approach, where λ is estimated as the slope of a regression of logarithm of squared residuals of the regression of X on W on the logarithm of squared W using the calibration data. We take into account the uncertainty of the estimation of λ by bootstrapping the calibration data to assure that MI is proper. We label this simplified MI method SMINDα.

4. Simulation study

We assess the performance of the above methods by a simulation study. We consider both internal and external calibration data designs, vary the strength of the association of Y and X, the size of measurement error, and consider both homoscedastic and heteroscedastic measurement error. Simulation scenarios are generated by the following combinations of parameters:

Main study data: $\gamma_0=0$; $\gamma_X=0.3$ or 0.6, $\tau^2=1$.

Measurement error data: $\beta_0=0$; $\beta_1=0.5$, 1 or 2; $\sigma^2=0.25$, 0.5 or 1; $\alpha=0$ (homoscedastic measurement error) or 0.4 (heteroscedastic measurement error). The cases of $\sigma^2=0.5$ and 1 are investigated only for $\beta_1=0.5$, which results in five combinations. Different combinations of β and σ^2 values result in varying degrees of the correlation ρ_{XW} between X and W.

To clarify the notation, 'calib' and 'main' will be attached to subscript to denote the calibration study and main study, respectively. We simulate $n_{\text{calib}}=100$ observations in the calibration sample and $n_{\text{main}}=400$ observations in the main sample. For each scenario, 500 simulated data sets are generated.

The true X is first generated from the normal distribution with variance 1. Each main study data set is simulated by randomly generating the values for the response variable Y_i and the observed error-prone variable W_i for $i=1,\ldots,n_{\text{main}}$, based on the models (2) and (3) respectively. For the external calibration data design, we randomly generate values of W_i from the measurement error model (2), for $i=1,\ldots,n_{\text{calib}}$. For the internal calibration data design, we also generate responses Y_i, $i=1,\ldots,n_{\text{calib}}$, using the model (3) with the same values of γ and τ^2 used to simulate the corresponding main study data.

For each of the 500 simulations across each of the simulation scenarios, we estimate the parameter of interest γ_X for each of the measurement error correction methods described above. All methods are compared with respect to bias, root mean squared error (RMSE) of the estimates and empirical non-coverage of 95 per cent confidence intervals. The empirical non-coverage is calculated as the proportion of simulated data sets for which the 95 per cent confidence interval does not include the true value of γ_X. The proportions are multiplied by 1000 to avoid decimal points, and hence a nominal level of non-coverage is equal to 50.

5. Results

In Tables I–III, we examine the performance of the naïve regression of Y on W (i.e. ignoring measurement error), and measurement error correction techniques CA, RC, and MI. We focus on the performance of various methods on inferences for the regression coefficient γ_X. We compare the methods in situations where the association between X and Y is weak ($\gamma_X=0.3$) and strong ($\gamma_X=0.6$), and where the measurement error, as measured by the correlation ρ_{XW} between X and W, is small ($\rho_{XW}>0.9$) and large ($\rho_{XW}<0.6$). Table I compares Naïve, CA, RC, ERC, LSCalib, MI0, and MIND0 for the case of internal calibration data with homoscedastic measurement error ($\alpha=0$). Table II compares Naïve, CA, RC, MI0, MIND0, WRC, WERC, LSCalib, MINDα, and SMINDα for internal calibration data with heteroscedastic measurement error ($\alpha=0.4$). Table III compares Naïve, CA, RC, MIND0, WRC, MINDα, and SMINDα for external calibration data with heteroscedastic measurement error ($\alpha=0.4$); we do not evaluate the LSCalib, ERC, and WERC methods since they are not applicable for this data structure.

Table I presents the results for the case of homoscedastic measurement error. As theory predicts, Naïve estimates are attenuated towards zero, with the degree of attenuation varying with the magnitude of measurement error and the response–covariate association. The non-coverage rate of Naïve is much higher than the nominal level of 50 in most of simulation scenarios. CA also performs poorly, with substantial bias and poor confidence interval coverage, particularly when the measurement error is large. RC is much less biased and has much better coverage than Naïve and CA, but has very large RMSE when the measurement error variance is large, suggesting that it is not very efficient. ERC, LSCalib,

Table I. Empirical bias, RMSE and non-coverage of 95 per cent confidence interval (nominal = 50) of estimates of γ_X with the internal calibration data based on 500 simulations, when the variance of measurement error is constant.

γ_X	β	σ^2	ρ_{XW}	Inference	Naïve	CA	RC	LSCalib	ERC	MI0	MIND0
0.3	2	0.25	0.97	Bias	159	19	1	0	1	2	1
				RMSE	160	51	52	98	46	50	45
				Non-coverage	1000	82	68	58	52	42	50
0.3	1	0.25	0.89	Bias	61	62	1	1	2	3	2
				RMSE	77	77	60	101	50	61	49
				Non-coverage	268	266	68	58	52	44	52
0.3	0.5	0.25	0.71	Bias	2	152	6	0	3	3	1
				RMSE	73	157	82	99	67	71	61
				Non-coverage	58	952	52	58	44	42	44
0.3	0.5	0.5	0.58	Bias	101	202	16	1	5	3	2
				RMSE	118	204	110	100	70	73	66
				Non-coverage	398	998	48	58	38	58	44
0.3	0.5	1	0.44	Bias	181	241	23	1	4	4	3
				RMSE	187	243	315	99	79	76	70
				Non-coverage	970	1000	40	58	46	52	40
0.6	2	0.25	0.97	Bias	318	38	1	0	0	3	1
				RMSE	319	63	54	98	45	51	45
				Non-coverage	1000	110	64	58	48	40	48
0.6	1	0.25	0.89	Bias	122	124	2	0	2	3	2
				RMSE	130	134	66	98	54	62	54
				Non-coverage	742	652	64	58	40	46	48
0.6	0.5	0.25	0.71	Bias	4	303	18	1	3	3	3
				RMSE	71	307	105	96	69	75	66
				Non-coverage	50	1000	52	58	42	48	42
0.6	0.5	0.5	0.58	Bias	201	401	40	0	4	4	3
				RMSE	209	404	151	98	77	79	69
				Non-coverage	894	1000	42	58	40	52	38
0.6	0.5	1	0.44	Bias	359	480	54	1	5	4	3
				RMSE	362	481	531	101	86	80	71
				Non-coverage	1000	1000	40	58	57	50	42

All values are multiplied by 1000. Naïve, naïve linear regression of Y on W; CA, conventional approach; RC, regression calibration; LSCalib, linear regression of Y on X using calibration data only; ERC, efficient regression calibration; MI0, multiple imputation without the NDME assumption; MIND0, multiple imputation with the NDME assumption.

MI0, and MIND0 methods have little empirical bias. When measurement error is small or moderate (e.g. $\rho_{XW}>0.6$), ERC generally has smaller RMSE than RC, LSCalib, and MI0, a finding consistent with the simulation results in [18]. ERC has smaller RMSE than MI0 and similar or greater RMSE than MIND0 in small measurement error settings. For example, when $\gamma_X=0.3$, $\beta=1$ and $\sigma^2=0.25$, the RMSEs of ERC, MI0, and MIND0 are 50, 61, and 49, respectively. For large measurement error and strong response–covariate association, ERC has higher RMSE than both MI0 and MIND0. Overall, MIND0 is superior to other methods, with small empirical bias, low RMSE and close to nominal levels of confidence coverage.

Table II concerns heteroscedastic measurement error for the internal calibration data design, so ERC, WERC, and LSCalib are available for comparison. The MI methods taking into account heteroscedastic measurement error perform better than other methods with respect to bias, RMSE and confidence interval coverage, for all simulation scenarios considered. The simplified MI method SMINDα is comparable in performance to MINDα. Naïve and CA are both seriously biased for γ_X with high non-coverage. RC is also badly biased with inflated RMSE when measurement error and the response–covariate effect are both large. WRC has less empirical bias than RC, but some bias remains, and it has large RMSE when the measurement error is large. WERC has smaller bias and lower RMSE than WRC, especially for large measurement error. For example, when $\gamma_X=0.6$, $\beta_1=0.5$ and $\sigma^2=1$, the RMSE of WERC is 0.102, compared with 1.655 for WRC. The good performance of WERC possibly arises because the

Table II. Empirical bias, RMSE and non-coverage of 95 per cent confidence interval (nominal = 50) of estimates of γ_X with the internal calibration data based on 500 simulations, when the variance of measurement error is heteroscedastic.

γ_X	β	σ^2	ρ_{XW}	Inference	Naïve	CA	RC	LSCalib	MIND0	MI0	WRC	WERC	MINDα	SMINDα
0.3	2	0.25	0.92	Bias	175	47	3	5	2	3	1	0	0	0
				RMSE	177	69	61	96	50	59	61	53	48	49
				Non-coverage	1000	176	70	46	58	60	60	68	58	46
0.3	1	0.25	0.76	Bias	130	129	8	5	5	7	7	2	2	3
				RMSE	136	136	81	96	64	76	80	64	58	59
				Non-coverage	912	824	66	46	58	64	60	66	56	44
0.3	0.5	0.25	0.50	Bias	152	226	29	5	2	8	22	6	3	4
				RMSE	161	228	156	96	76	88	147	80	69	72
				Non-coverage	822	1000	48	46	54	60	46	48	46	50
0.3	0.5	0.5	0.38	Bias	216	258	116	5	8	7	20	10	4	5
				RMSE	220	259	1335	96	81	92	699	88	71	79
				Non-coverage	1000	1000	42	46	50	78	42	60	46	44
0.3	0.5	1	0.28	Bias	250	274	251	3	6	9	47	11	6	7
				RMSE	252	274	1951	106	83	94	882	99	72	81
				Non-coverage	1000	1000	40	60	58	68	40	58	52	44
0.6	2	0.25	0.92	Bias	349	94	5	5	3	6	3	0	0	1
				RMSE	350	109	68	96	54	63	67	57	52	53
				Non-coverage	1000	430	64	46	52	60	60	64	56	52
0.6	1	0.25	0.76	Bias	259	257	16	8	7	6	16	2	4	3
				RMSE	263	262	101	96	70	80	101	72	60	67
				Non-coverage	1000	988	72	46	46	70	56	54	46	48
0.6	0.5	0.25	0.50	Bias	302	450	57	5	6	13	45	8	3	9
				RMSE	307	451	230	96	78	90	221	87	70	78
				Non-coverage	998	1000	38	46	50	62	38	54	48	48
0.6	0.5	0.5	0.38	Bias	430	514	220	5	8	13	52	13	8	12
				RMSE	432	515	2196	96	82	94	1327	94	72	78
				Non-coverage	1000	1000	42	46	52	74	40	62	56	44
0.6	0.5	1	0.28	Bias	505	550	397	3	7	11	129	12	7	13
				RMSE	506	551	3457	106	87	96	1655	102	76	85
				Non-coverage	1000	1000	36	60	58	70	40	64	44	38

All values are multiplied by 1000. Naïve, naïve linear regression of Y on W; CA, conventional approach; RC, regression calibration; LSCalib, linear regression of Y on X using calibration data only; WRC, weighted regression calibration; WERC, weighted ERC; MI0, multiple imputation without the NDME assumption; MIND0, multiple imputation with the NDME assumption. MINDα, multiple imputation with the NDME and $\alpha \neq 0$ assumptions; SMINDα, simplified version of MINDα.

inflated standard error of WRC reduces its effect on WERC, and the least squares estimate stabilizes the estimation, as WERC is an inverse-variance-weighted average of WRC and the least squares estimate. When measurement error is small, WERC is comparable to MINDα, with small bias and low RMSE. When measurement error is large and the response-covariate association is strong, WERC has larger RMSE than MINDα. The MI methods that assume constant measurement error variance (MI0 and MIND0) have larger RMSE and higher non-coverage than the MI methods that allow for nonconstant variance.

Results of the external calibration design are presented in Table III. We observe similar trends to those seen in Table II. Naïve and CA are both biased, with above nominal confidence interval non-coverage.

Table III. Empirical bias, RMSE and non-coverage of 95 per cent confidence interval (nominal = 50) of estimates of γ_X with the external calibration data based on 500 simulations, when the variance of measurement error is heteroscedastic.

γ_X	β	σ^2	ρ_{XW}	Inference	Naïve	CA	RC	MIND0	WRC	MINDα	SMINDα
0.3	2	0.25	0.92	Bias	172	46	3	5	2	0	0
				RMSE	174	66	57	61	57	52	58
				Non-coverage	1000	180	64	88	60	54	52
0.3	1	0.25	0.76	Bias	129	125	4	13	3	2	6
				RMSE	134	132	80	81	78	74	79
				Non-coverage	910	836	60	100	56	46	52
0.3	0.5	0.25	0.50	Bias	148	223	28	55	23	7	11
				RMSE	156	224	157	177	150	129	135
				Non-coverage	822	1000	40	128	52	56	46
0.3	0.5	0.5	0.38	Bias	212	255	72	86	58	8	14
				RMSE	216	256	389	246	378	162	171
				Non-coverage	1000	1000	44	196	42	48	42
0.3	0.5	1	0.28	Bias	250	274	251	68	109	10	12
				RMSE	252	274	1951	306	888	213	239
				Non-coverage	1000	1000	40	212	40	56	38
0.6	2	0.25	0.92	Bias	346	92	5	8	3	2	2
				RMSE	347	106	63	68	62	61	65
				Non-coverage	1000	422	66	106	64	58	54
0.6	1	0.25	0.76	Bias	255	253	5	27	4	6	7
				RMSE	259	258	98	107	97	92	100
				Non-coverage	1000	998	58	120	56	54	50
0.6	0.5	0.25	0.50	Bias	299	447	51	82	42	8	13
				RMSE	304	449	252	210	234	145	166
				Non-coverage	1000	1000	40	228	48	52	46
0.6	0.5	0.5	0.38	Bias	427	513	197	99	91	18	23
				RMSE	429	513	1211	241	1016	171	194
				Non-coverage	1000	1000	42	268	40	56	58
0.6	0.5	1	0.28	Bias	510	556	314	93	166	21	28
				RMSE	511	556	2526	306	1774	236	251
				Non-coverage	1000	1000	48	274	52	58	54

All values are multiplied by 1000. Naïve, naïve linear regression of Y on W; CA, conventional approach; RC, regression calibration; WRC, weighted regression calibration; MIND0, multiple imputation with the NDME assumption. MINDα, multiple imputation with the NDME and $\alpha \neq 0$ assumptions; SMINDα, simplified version of MINDα.

When measurement error is large, RC performs poorly with large bias and RMSE, and WRC has less empirical bias but inflated RMSE. MIND0 has generally greater RMSE than MINDα. Overall, MINDα and its simplified version SMINDα dominate all other methods.

The results of the MI methods presented in Tables I–III are based upon $M = 16$ multiply-imputed data sets. We also examined the performance of MI for $M = 10$ and 20. The results were similar, although coverage with $M = 10$ was slightly below the nominal level.

6. Sensitivity to the choice of prior distribution for X

In this work, we have specified a 'default' normal prior distribution of X with a noninformative prior on the mean and log variance. Ideally, the prior distribution of X would be tailored to the particular context, but we are deliberately considering an 'off-the-shelf' method that does not entail this level of customized specification, since it is not needed for the compared methods. Given this orientation, we conduct a limited investigation of the robustness of the proposed MI methods to misspecification of the prior distribution for X. We consider simulation designs similar to those described in Richardson *et al.* [33]. Two forms of misspecification of the prior distribution are examined. The first case is referred to as 'bimodal', and simulates X from a well-separated symmetric bimodal mixture given by

Table IV. Sensitivity to normality assumption in the bimodal and the skew cases.

ρ_{XW}	σ^2	Inference	Naïve	CA	RC	WRC	WERC	MINDα
					Bimodal case			
0.88	1.340	Bias	121	120	5	7	5	8
		RMSE	123	125	40	42	33	36
		Non-coverage	1000	970	50	52	50	62
0.77	0.670	Bias	202	200	10	11	9	10
		RMSE	203	204	49	51	39	38
		Non-coverage	1000	1000	48	62	46	58
0.64	0.340	Bias	300	299	14	17	10	12
		RMSE	301	301	66	68	43	41
		Non-coverage	1000	1000	42	50	48	54
					Skew case			
0.89	4.000	Bias	139	140	1	2	2	4
		RMSE	149	152	75	75	65	63
		Non-coverage	760	738	46	48	50	34
0.81	1.800	Bias	241	242	8	6	5	5
		RMSE	246	249	98	97	78	72
		Non-coverage	998	986	48	46	52	54
0.69	0.900	Bias	347	348	18	14	12	11
		RMSE	350	352	139	135	88	84
		Non-coverage	1000	1000	36	40	54	56

$0.5\mathrm{N}(-2.0, 1.0)+0.5\mathrm{N}(2.0, 1.0)$. The second case is referred to as 'skew', and simulates X from an asymmetric normal mixture given by $0.5\mathrm{N}(0.19, 0.082)+0.2\mathrm{N}(1.05, 0.22)+0.3\mathrm{N}(2.0, 0.482)$.

In both cases, the surrogate W is generated from the heteroscedastic measurement error model $\mathrm{N}(X, \sigma^2 X^{2\alpha})$ with $\alpha=0.4$, and values of σ^{-2} were simulated corresponding to different values of correlation between W and X, representing varying sizes of measurement error (Table IV). The outcome Y is related to X by a linear regression model $Y \sim \mathrm{N}(\gamma_0+\gamma_X X, \tau^2)$ with $\gamma_0=0$, $\gamma_X=0.6$, and $\tau^2=1$. To be consistent with simulation design reported in Section 4, the sample size of the main study is 400, and the sample size of the calibration study is 100. Under each simulation setting, the number of simulated data sets is set to 500. We consider the case of internal calibration data with heteroscedastic measurement error, so all proposed methods are available for comparison. We mainly focus on the performances of Naïve, CA, RC, WRC, WERC, and MINDα.

Table IV summarizes the results corresponding to the bimodal and the skew cases with respect to the bias, RMSE, and non-coverage rate per 1000 samples of the 95 per cent confidence intervals (target = 50). In both the bimodal and skew cases, the MI method provides satisfactory or conservative confidence coverage, and has lower RMSE than the other methods. These simulations suggest that our MI methods have satisfactory frequentist performance, at least for these choices of misspecification of the prior distribution.

7. Application

In this section, we perform an analysis of the data set from the BioCycle study. This study was designed to assess the relationship between endogenous hormones and biomarkers of oxidative stress during the menstrual cycle. Two hundred and fifty-nine regularly menstruating pre-menopausal women were followed for two menstrual cycles. The goal of the study was to investigate the association between carotenoids (β-carotene) and progesterone.

The calibration data were obtained from calibration experiments for human serum fat-soluble vitamins, measured by using HPLC. Three replicate calibration experiments were performed, with each experiment analyzing eight samples with known concentrations of the analyte carotene, from standard reference materials (SRMs) obtained from the National Institute of Standards and Technology. For each of the three replicate experiments, each of eight samples was analyzed 10 times by using HPLC, yielding a total of 30 replicate measures for each sample. In the calibration data, the true concentrations of carotene X are known, and HPLC measurements can be viewed as error-contaminated versions of X, denoted by W. This calibration data are external, since it includes only the information of W and X.

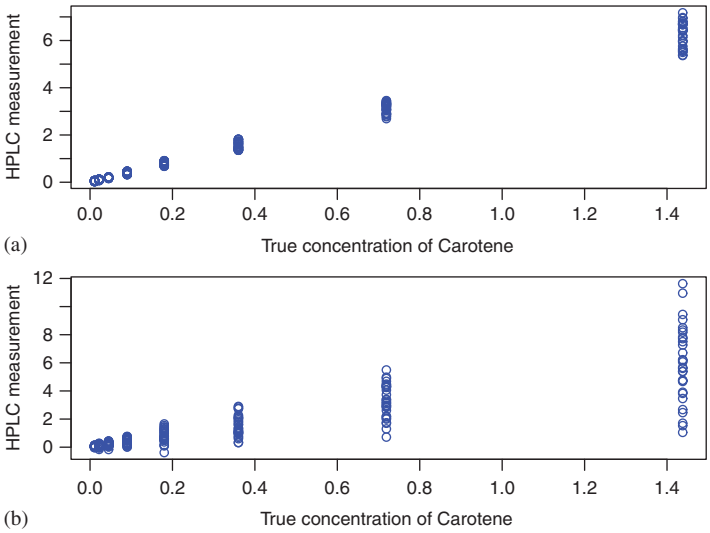

Figure 2. Calibration data of carotene from the BioCycle study: (a) Original calibration data and (b) Modified calibration data.

Table V. BioCycle data: estimated regression coefficients in a linear regression model with carotene as the covariate and progesterone as the dependent variable using data collected at visiting time F1.

Parameters	Naïve	CA	RC	WRC	MINDα
Intercept	0.5089 (0.0238)	0.5089 (0.0238)	0.5095 (0.0244)	0.5099 (0.0242)	0.5089 (0.0239)
Carotene	−0.0065 (0.0024)	−0.0281 (0.0104)	−0.0289 (0.0107)	−0.0286 (0.0106)	−0.0282 (0.0107)

The calibration data of carotene, which is generated from SRMs 968 C1, is external. Standard error is shown in parentheses.

The main study consisted of 211 individual samples with complete information on both progesterone and carotene. Each individual sample had one measurement of outcome progesterone Y and one HPLC measurement of carotene W, but the true concentration of carotene X was unknown. We use the data collected at the visiting time $F1$.

Figure 2(a) shows the original calibration data for carotene. It is clear that the variance of HPLC measurements increases as the true concentration of carotene increases, suggesting that the measurement error variance is not constant. To estimate the linear regression of progesterone Y on carotene X, we apply the naïve regression of Y on W, and four different measurement error correction methods (i.e. CA, RC, WRC, and MI) to the data. Standard errors of the CA, RC, and WRC methods are calculated using the bootstrapping method. Note that these methods are applied to correct only for measurement error from the assay, but not from other sources such as biological variation.

Table V presents the estimates and associated standard errors for regression coefficients. The naïve estimate indicates a weak association between progesterone and carotene—the change of progesterone is 0.0065 when carotene changes one unit. Adjusting for measurement error by the CA, RC, WRC, and MI methods, we find a stronger association between carotene and progesterone. The error-corrected estimates are all four-fold greater than the uncorrected estimates. In particular, the CA method estimates a change of progesterone of 0.0281 for one unit change in carotene. The estimates obtained from the RC, WRC, and MI methods are similar, and slightly larger than the CA estimate.

Our simulation studies suggest that the performance of CA, RC, and MI is more differentiated when the magnitude of measurement error is large. Differences between the correction methods are minor in this example, we think because the magnitude of measurement error is not large. Specifically, the maximum likelihood estimators of the measurement error variance and the effect size are $\widehat{\sigma}^2 = 0.117$ and $\hat{\beta}_1 = 4.353$, indicating a high correlation between X and W ($\rho_{XW} = 0.98$).

To better illustrate our proposed approaches, we created a modified BioCycle data set, where some random noise was added to HPLC measurements to increase the magnitude of measurement error. Figure 2(b) shows the modified calibration data, which maintain the same pattern as the original

Table VI. Modified BioCycle data: estimated regression coefficients in a linear regression model with carotene as the covariate and progesterone as the dependent variable using data collected at visiting time F1.

Parameters	Naïve	CA	RC	WRC	MINDα
Intercept	0.5068 (0.0238)	0.5066 (0.0238)	0.5110 (0.0242)	0.5085 (0.0231)	0.5128 (0.0253)
Carotene	−0.0061 (0.0024)	−0.0255 (0.0099)	−0.0339 (0.0123)	−0.0297 (0.0113)	−0.0285 (0.0105)

The calibration data are external. Standard error is shown in parentheses.

calibration data but are more dispersed, with $\widehat{\sigma}^2 = 2.496$, resulting in a relatively low correlation ($\rho_{XW} = 0.77$), compared with the original calibration data. We analyze the modified data with our proposed methods, and results are summarized in Table VI.

As shown in Table VI, the naïve analysis attenuates the association between carotene and progesterone toward the null, as expected. Noticeably, when applying our proposed error correction methods to the modified BioCycle data, there is an appreciable difference in the estimates of regression coefficient of carotene. In particular, the estimates of the CA, RC, WRC, and MINDα methods are −0.0255, −0.0339, −0.0298, and −0.0280, respectively. We also observe that the RC estimate has a larger standard error than the WRC or MI estimate.

8. Discussion

The study in [18] shows that ERC outperforms MI for the case of homoscedastic measurement error. The reason is that ERC exploits the NDME assumption, whereas the version of MI (MI0) considered by these authors does not. The results reported in Table I indicate that a version of MI that exploits the NDME assumption (MIND0) is similar or superior to ERC for the simulation conditions compared. This finding is to be expected, given the asymptotic efficiency of MI under a correctly specified model, as the number of imputations tends to infinity. Our simulation results also show the efficiency gains of MIND0 over MI0, demonstrating the utility of taking into account the NDME assumption when it is substantively reasonable. The RMSE of the MIND0 estimate is generally 10–15 per cent smaller than that of the MI0 estimate in our simulation settings.

The main focus of this article is on extending methods to the case of heteroscedastic measurement error, a situation where existing methods are biased. In particular, the RC method, which imputes a conditional mean of X given W, does not yield consistent estimates when the measurement error variance is not constant. Our modification WRC of RC, based on estimating the conditional means by weighted least squares, reduces but does not solve this problem. In contrast, the MI methods, which impute draws rather than means, can readily allow for nonconstant measurement variance by simply modifying the imputation model to reflect this feature. Our simulations support that the MI methods are superior to other existing methods, particularly when measurement error is substantial. The higher level of measurement error in our simulation studies may be more substantial than in many real settings, but the widespread use of a flawed technique (e.g. the CA method) inhibits the ability to make use of assays that have relatively high levels of measurement error. Our method allows inferences in noisy assays that currently would not be regarded as acceptable. We have also studied the performance of MINDα (developed to deal with heteroscedastic measurement error) under the scenario of homoscedastic measurement error using the same simulation settings presented in Section 4. Our simulation results show that there is zero or a small loss of efficiency (less than 6 per cent) for all simulation scenarios we considered.

We consider inference for the parameters of the regression Y on X by applying standard MI combining rules to multiply-imputed data sets. The MI analysis is appealing since it allows standard analysis methods to be used on the filled-in data sets. However, it is certainly possible to base inferences directly on the posterior distributions of model parameters—these inferences should be similar to MI, since MI combining rules are based on simulation approximations to the posterior distribution.

The RC method and its extensions rely on the assumption that measurement error is non-differential, and in this article we focus on the performance of MI methods based on the NDME assumption. However, we note that the MI methods can also handle differential measurement error, provided that internal calibration data are available to identify the model parameters. The MINDα method designed under the NDME assumption can be easily extended to work in the case where the assumption may

not hold, i.e. allowing for differential measurement error by remodeling the measurement error as $p(W|X, Y)$, instead of $p(W|X)$.

Although the MINDα method performs well in the simulation studies, it is comparatively complex computationally, given its use of Markov Chain Monte Carlo (MCMC) with a MH step. We provide an alternative, SMINDα, which is much simpler computationally since it avoids the MH step. It is based on an approximate model similar to WRC, so it lacks statistical rigor. However, SMINDα performed similarly to MINDα in the simulations, and it takes much less time to compute.

Our Bayesian MI methods require specification of prior distributions for the model parameters, and also for X. We choose simple noninformative priors. Our main intent here is not to come up with the best possible Bayesian model, but to compare frequentist performance of Bayesian inferences under noninformative priors with more standard frequentist methods, in the heteroscedastic measurement error setting. We are seeking an 'off-the-shelf' all-purpose method rather than a method that tailors prior to the particular data set, since this form of tailoring is not required by the competing methods. We assume that the prior distribution for X is normally distributed with noninformative priors on the mean and log variance, and our simulations suggested a degree of robustness to misspecification of this prior, at least in the cases investigated. However, other more tailored choices of prior distribution for X may lead to better inferences. For example, for linear measurement error models, mixtures of normal distributions for X have been proposed with a prespecified [34] or unknown [33] number of components. These methods could be adapted to our heteroscedastic setting.

We assume that the measurement error variance model is $\sigma^2 X^{2\alpha}$, indicating that the variance increases as the true value of X increases through a power function. This form of nonconstant variance model is a common and intuitive choice, but the proposed MI methods could be extended to more complex choices of variance function, at the expense of some additional computational complexity. In recent work, Spiegelman *et al.* [19] propose Taylor series-based modifications of RC for heteroscedastic measurement error under the assumption that $g(X, \alpha) = g(X)$ is known. These methods do not appear to improve on RC in their simulations, and RC performs better in their simulations than in the simulations reported here.

We have restricted attention here to the case where a simple regression of Y on X is of interest. It is relatively straightforward to extend our MI methods to allow for other covariates Z, recorded without measurement error, since values of these variables can be conditioned in the MCMC analysis. Extensions to non-normal outcomes, as when Y is binary and follows a probit model, could also be developed without too much difficulty.

A crucial assumption here is that the same measurement error model holds for the calibration and main data sets. In some epidemiologic study designs, the calibration data are supplied by an external source, such as a pure standard sample, and the relationship of the true and measured variables might be different for the calibration and main study data, because the sample from each subject of the main study might have impurities that change the measurement error properties. Methods that allow a different measurement model in the two samples are a worthwhile topic for future research, although we expect problems in identifying the parameters in that case.

The traditional CA method performed poorly in our simulations, but yielded reasonable estimates in the BioCycle data application, where the measurement error was small. The discrepancy between CA and the other methods becomes more substantial as the measurement error increases, as our analysis of the modified BioCycle data demonstrated. The developed MI methods can be expected to yield substantial improvements when the response–covariate association is strong and the measurement error is large.

Appendix A

In this article, we apply the MI approach via data augmentation using a MCMC algorithm. We consider measurement error problems in a missing data context where the true value of X is unobserved. The data augmentation method is a two-step iterative algorithm. The key idea is to generate draws for missing values given observed data and a set of parameters (I-step) and then draws of the model parameters from their posterior distribution given completed data (P-step). These two steps are iterated until convergence, and then the results are used to generate multiply-imputed data sets [35].

Suppose we have data collected from the main study containing n_{main} observations (W_i, Y_i) and from the external calibration study containing n_{calib} observations (X_i, W_i). The observations are assumed

to be independently and identically distributed. We let n denote the number of all observations, $n = n_{\text{calib}} + n_{\text{main}}$.

The posterior predictive distribution of X given W, Y and the parameters cannot be expressed in a closed form; however, by Bayes' Theorem, it can be factorized as

$$p(X|Y,W,\beta,\sigma,\alpha,\gamma,\tau,\mu_x,\sigma_x) \propto p(W|X,Y,\beta,\sigma,\alpha)p(Y|X,\gamma,\tau)p(X|\mu_x,\sigma_x)$$
$$\propto p(W|X,\beta,\sigma,\alpha)p(Y|X,\gamma,\tau)p(X|\mu_x,\sigma x)$$

This factorization includes three models: the measurement error model which links the true variable X and the observed variable W, the main study model which specifies the relationship between the response variable Y and the unobserved covariate X, and the prior distribution for X.

Assuming that the parameter vectors $\theta = (\beta, \sigma, \alpha)$, $\psi = (\gamma, \tau)$ and $\pi = (\mu_x, \sigma_x)$ are distinct and $priori$ independent, the likelihood function can be factorized, and the joint posterior for the parameters given the complete data can be expressed as

$$p(\beta,\sigma,\alpha,\gamma,\tau,\mu_x,\sigma_x|Y,X,W) \propto p(Y,X,W|\beta,\sigma,\alpha,\gamma,\tau)p(\beta,\sigma,\alpha,\gamma,\tau,\mu_x,\sigma_x)$$
$$\propto p(W|X,\beta,\sigma,\alpha)p(\beta,\sigma,\alpha)p(Y|X,\gamma,\tau)p(\gamma,\tau)p(X|\mu_x,\sigma_x)p(\mu_x,\sigma_x)$$

Hence θ, ψ and π can be sampled separately.

For the external calibration/main design, we also need impute missing values for Y. The posterior distribution of Y given the parameters is simply normal with mean $(\gamma_0 + \gamma_X X)$ and variance τ^2. Draws of Y are obtained by substituting drawn values of the parameters in this distribution.

Having obtained the complete-data posteriors for the model parameters, and the predictive distribution for X and Y, our imputation procedure for the external calibration/main study composes of the following steps:

I-1 step: Generate imputed values of Y_i for i corresponding to the ith observation in the calibration study, $i = 1, \ldots, n_{\text{calib}}$, from the posterior density

$$p(Y_i|X_i,\psi) \sim \text{N}(\gamma_0 + \gamma_X X_i, \tau^2)$$

I-2 step: Generate imputed values of X_i for i corresponding to the ith observation in the main study, $i = 1, \ldots, n_{\text{main}}$, from the posterior density specified by

$$p(X_i|Y_i,W_i,\beta,\sigma,\alpha,\gamma,\tau\sigma_x,\mu_x) \propto \tau^{-2}\exp\left(-\frac{1}{2\tau^2}(Y_i-\gamma_0-\gamma_X X_i)^2\right)$$
$$\times\sigma^{-2}X_i^{-2\alpha}\exp\left(-\frac{1}{2\sigma^2 X_i^{2\alpha}}(W_i-\beta_0-\beta_1 X_i)^2\right)$$
$$\times\sigma_x^{-2}\exp\left(-\frac{1}{2\sigma_x^2}(X_i-\mu_x)^2\right)$$

P-1 step: Draw ψ from the posterior density $p(\psi|X,Y)$.
P-2 step: Draw θ from the posterior density $p(\theta|W,X)$.
P-3 step: Draw π from the posterior density $p(\pi|X)$.

For the internal calibration/main study design, we observe (Y, X, W) in the calibration study. Therefore, ML estimates for the regression parameters γ_0, γ_X and τ^2, as well as the measurement error model parameters β_0, β_1 and σ^2 can be computed using the calibration data by standard analysis (e.g. least squares methods), and used as initial values for the data augmentation algorithm. The sample mean and variance of X can also be calculated from the calibration data, and used as initial values for μ_x and σ_x^2. The initial value of α is estimated as the slope of regression of logarithm of squared residuals of the regression of X on W on the logarithm of squared W. For the external calibration/main study design, we still can obtain estimates of the measurement error parameters β_0, β_1, σ^2 and α, using the same approach as for the internal calibration design. However, the initial values of γ_0, γ_X and τ^2 cannot be computed straightforwardly because X and Y are not observed together in the whole data set. We apply the RC method to obtain initial values. Specifically, we first substitute unobserved values of X

in the main study by the expectation $E(X|W)$, and then regress Y on the substituted values to obtain initial estimates for γ and τ^2.

The data augmentation method iterates between the P-step and the I-step for a large number of times until the algorithm converges. We use the method proposed by Gelman and Rubin [36] to diagnose convergence of the model parameter iterates. Initial values of model parameters are chosen to be reasonably overdispersed by bootstrapping the original data set. For each of the bootstrap samples, estimates of model parameters are obtained using the method described above, and used as the initial points for the MCMC chains. After the algorithm converges, we discard data from the initial burn-in period before saving the values. The imputed complete data set is generated by using every dth iteration after an adequate burn-in period, to avoid possible autocorrelation between successive sets of imputed values, so that the imputed data sets can be treated as independent. For all the model parameters, we observe reasonable mixing and convergence after 2000 iterations of the MCMC chains. Hence, we decide to discard the first 2000 iterations, and choose $d=100$. We generate 16 imputed data sets, which can be analyzed by standard complete data methods.

The imputation procedure for the internal calibration/main study is similar to that of the external calibration/main study, except that the I-1 step can be omitted since Y is observed in this design.

The I-1 step can be performed easily by generating a random draw from a normal distribution. The I-2 step is not straightforward, since we do not have an analytical expression for the posterior density of X. We use a Metropolis–Hastings algorithm for this step to generate values of X_i. It consists of the following two steps:

- Generate a value $\boldsymbol{X}_i'$ from an appropriate candidate-generating density $q_i(\boldsymbol{X}_i, \boldsymbol{X}_i')$, where $\boldsymbol{X}_i$ denote the current value
- Set

$$\boldsymbol{X}_i^{(j+1)} = \begin{cases} \boldsymbol{X}_i' & \text{with probability } \kappa = min\left[1, \dfrac{\pi(\boldsymbol{X}_i')q(\boldsymbol{X}_i', \boldsymbol{X}_i^{(j)})}{\pi(\boldsymbol{X}_i^{(j)})q(\boldsymbol{X}_i^{(j)}, \boldsymbol{X}_i')}\right] \\ \boldsymbol{X}_i^{(j)} & \text{otherwise.} \end{cases}$$

  where $\pi(X_i)$ is the target density from which we want to simulate X.

The choice of candidate-generating density is arbitrary, but a correctly specified density can improve the efficiency of this algorithm. In our work, we generate a candidate $\boldsymbol{X}_i'$ from a Gaussian model centered on the current value $\boldsymbol{X}_i^{(j)}$ and variance σ_*^2, equal to a scaled sampling variance of the target density. That is, $q(\boldsymbol{X}_i', \boldsymbol{X}_i^{(j)}) = \mathrm{N}(\boldsymbol{X}_i^{(j)}, \sigma_*^2)$. The algorithm with the normal-generating density is also called 'a random walk Metropolis' algorithm.

The P-step requires generating the draws of the parameters from the complete-data posterior distribution. In P-1 step, by Bayes' rule, the posterior for $\psi = (\gamma, \tau)$ given the data (X, Y) can be factored as a multivariate normal distribution and a scaled inverted χ^2 -distribution, which make it easy to draw the values. In practice, we first draw τ^2 from $\tau^2 \sim \mathrm{inv}-\chi^2(\upsilon, S^2)$; and then draw γ from multivariate normal distribution $\mathrm{N}(\hat{\gamma}, \tau^2(X'X)^{-1})$, where $\hat{\gamma} = (X'X)^{-1}X'Y$ is the ordinary least squares estimate from the regression of all data, $S^2 = (Y - X\hat{\gamma})'(Y - X\hat{\gamma})$ is the corresponding sample variance, and $\upsilon = n-2$ is the degree of freedom.

Drawing parameter θ from the posterior density $p(\theta|X, W)$ is little complicated due to the presence of the unknown parameter α. For known α, draws of β, σ^2 can be readily obtained from their posterior density using an approach similar to the P-1 step, with the least squares estimate replaced by the weighted least squares estimate. Specifically, let $\omega_i = 1/X_i^{\alpha}$, we first generate values of α from its posterior distribution,

$$p(\alpha|\beta_0, \beta_1, \sigma^2, X, W) \propto \left(\prod_{i=1}^{n} \omega_i\right) \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\omega_i^2(W_i - \beta_0 - \beta_1 X_i)^2\right\}, \quad i = 1, \ldots, n$$

using a random walk Metropolis step. Given α, we draw σ^2 from $\sigma^2 \sim \mathrm{inv}-\chi^2(\upsilon, S_\omega^2)$; and then draw β from multivariate normal distribution $\mathrm{N}(\hat{\beta}, \sigma^2(X'\omega X)^{-1})$, where $\hat{\beta} = (X'\omega X)^{-1}X'\omega W$ is the weighted least squares estimate from the regression of all data, ω is a $n \times n$ weighted matrix with the diagonal element ω_i and the other elements equal to zero, $S_\omega^2 = (W - X\hat{\beta})'\omega(W - X\hat{\beta})$ is the corresponding weighted sample variance, and $\upsilon = n-2$.

Acknowledgements

We thank the referees and editors for helpful comments.

References

1. Kipnis V, Subar AF, Midthune D, Freedman LS. Structure of dietary measurement error: results of the open biomarker study. *American Journal of Epidemiology* 2003; **158**:14–21.
2. Cotton SM, Crewther DP, Crewther SG. Measurement error: implications for diagnosis and discrepancy models of developmental dyslexia. *Dyslexia* 2005; **11**:186–202.
3. Wannemuehler KA, Lyles RH, Waller LA, Hoekstra RM, Klein M, Tolbert P. A conditional expectation approach for associating ambient air pollutant exposures with health outcomes. *Environmetrics* 2009; **20**:877–894.
4. Wactawski-Wende J, Schisterman EF, Hovey KM, Howards P, Browne RW, Hediger M, Liu A, Trevisan M. Biocycle study: design of the longitudinal study of the oxidative stress and hormone variation during the menstrual cycle. *Paediatric and Perinatal Epidemiology* 2009; **23**:171–184.
5. Guo Y, Harel O, Little RJ. How well quantified is the limit of quantification? *Epidemiology* 2010; **21**:S1–S7.
6. Fuller WA. *Measurement Error Models*. Wiley: New York, NY, 1987.
7. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models*: *A Modern Perspective* (2nd edn). Chapman & Hall/CRC: Boca Raton, FL, 2006.
8. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in problems with surrogate predictors. *Journal of the American Statistical Association* 1990; **85**:652–663.
9. Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrika* 2004; **60**:172–181.
10. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine* 2001; **20**:139–160.
11. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 2006; **35**:1074–1081. DOI: 10.1093/ije/dyl097.
12. Schenker N, Raghunathan TE, Bondarenko I. Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine* 2009; **29**:533–545. DOI: 10.1002/sim.3809.
13. Clayton DG. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health*, Dwyer JH, Feinleib M, Lippert P, Hoffmeister H (eds). Oxford University Press: Cary, NC, 1992.
14. Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine* 1993; **12**:1703–1722. DOI: 10.1002/sim.4780121806.
15. Richardson S, Green PJ. On bayesian analysis of mixtures with an unknown number of components (with Discussion). *Journal of the Royal Statistics Society. Series B* (*Methodological*) 1997; **59**:731–792.
16. Hossian S, Gustafson P. Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Statistics in Medicine* 2009; **28**:1580–1600. DOI: 10.1002/sim.3552.
17. Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association* 2005; **100**:1123–1132.
18. Freedman LS, Midthune D, Carroll R, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* 2008; **27**:5195–5216. DOI: 10.1002/sim.3361.
19. Spiegelman D, Logan R, Grove D. Regression calibration with heteroscedastic error variance. *The International Journal of Biostatistics* 2011; **7**. DOI: 10.2202/1557-4679.1259.
20. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 1987; **52**:528–540.
21. Hastings WK. Monte carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**:97–109.
22. Heid IM, Kuchenhoff H, Miles J, Kreienbrock L, Wichmann HE. Two dimensions of measurement error: classical and berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology* 2004; **14**:365–377.
23. Guolo A, Brazzale AR. A simulation-based comparison of techniques to correct for measurement error in matched case-control studies. *Statistics in Medicine* 2008; **27**:3755–3775. DOI: 10.1002/sim.3282.
24. Dellaportas P, Stephens DA. Bayesian analysis of errors-in-variables regression models. *Biometrics* 1995; **51**:1085–1095.
25. Hyslop DR, Imbens GW. Bias from classical and other forms of measurement error. *Journal of Business and Economic Statistics* 2001; **19**:475–481.
26. Kuha J, Temple J. Covariate measurement error in quadratic regression. *International Statistical Review* 2003; **71**:131–150.
27. Roeder K, Carroll RJ, Lindsay BG. A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* 1996; **91**:722–732.
28. Rodbard D, Frazier GR. Statistical analysis of radioligand assay data. *Methods of Enzymology* 1975; **37**:839–841.
29. Finney DJ. Radioligand assay. *Biometrics* 1976; **32**:721–740.
30. Higgins KM, Davidian M, Chew G, Burge H. The effect of serial dilution error on calibration inference in immunoassay. *Biometrics* 1998; **54**:19–32.
31. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley: Hoboken, NJ, 1987.

32. Little RJ, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: Hoboken, NJ, 2002.
33. Richardson S, Leblond L, Jaussent I, Green PJ. Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society*, *Series A* 1999; **165**:549–566.
34. Carroll RJ, Maca JD, Ruppert D. Nonparametric regression in the presence of measurement error. *Biometrika* 1999; **86**:541–554.
35. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.
36. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences (with Discussion). *Statistical Science* 1992; **7**:457–472.