

A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression

Laurence S. Freedman^{1,*},†, Douglas Midthune², Raymond J. Carroll³
and Victor Kipnis²

¹*Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel*

²*Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, U.S.A.*

³*Department of Statistics, Texas A&M University, College Station, TX, U.S.A.*

SUMMARY

Regression calibration (RC) is a popular method for estimating regression coefficients when one or more continuous explanatory variables, X , are measured with an error. In this method, the mismeasured covariate, W , is substituted by the expectation $E(X|W)$, based on the assumption that the error in the measurement of X is non-differential. Using simulations, we compare three versions of RC with two other ‘substitution’ methods, moment reconstruction (MR) and imputation (IM), neither of which rely on the non-differential error assumption. We investigate studies that have an internal calibration sub-study. For RC, we consider (i) the usual version of RC, (ii) RC applied only to the ‘marker’ information in the calibration study, and (iii) an ‘efficient’ version (ERC) in which the estimators (i) and (ii) are combined. Our results show that ERC is preferable when there is non-differential measurement error. Under this condition, there are cases where ERC is less efficient than MR or IM, but they rarely occur in epidemiology. We show that the efficiency gain of usual RC and ERC over the other methods can sometimes be dramatic. The usual version of RC carries similar efficiency gains to ERC over MR and IM, but becomes unstable as measurement error becomes large, leading to bias and poor precision. When differential measurement error does pertain, then MR and IM have considerably less bias than RC, but can have much larger variance. We demonstrate our findings with an analysis of dietary fat intake and mortality in a large cohort study. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: differential measurement error; moment reconstruction; multiple imputation; non-differential measurement error; regression calibration

*Correspondence to: Laurence S. Freedman, Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer 52161, Israel.

†E-mail: lsf@actcom.co.il

Contract/grant sponsor: National Cancer Institute; contract/grant number: CA-57030

1. INTRODUCTION

The problem of error in the measurement of covariates to be used in a regression analysis is often important in epidemiological research, where accurate measurements are commonly difficult to achieve. It is now well understood that such error can cause bias in the estimates of regression coefficients, and a large collection of special methods for eliminating such bias is available [1].

Regression calibration (RC) [2, 3] has been one of the most commonly used methods [4]. One of the principal advantages of RC is its simplicity. In place of each individual's mismeasured covariate one uses a substitute value and then runs the same estimation procedure as would be used with a precisely measured covariate. The substituted value is the expected value of the true measurement conditional on the observed measurement and other exactly measured covariates in the model.

Recently, it has been realized that two other 'substitution' methods are available that would possibly share these advantages enjoyed by RC. The first is moment reconstruction (MR), proposed by Freedman *et al.* [5]. The second is imputation (IM) or multiple IM [6], a class of methods that is well established for handling missing data, but that has been proposed for dealing with errors of measurement in covariates [7]. Cole *et al.* [8] applied this method to an epidemiological study with a binary outcome and a mismeasured binary covariate. MR and IM both allow the covariate measurement error to be differential, that is, informative about the outcome variable, whereas RC requires the measurement error to be non-differential. All three substitution methods require information about the measurement error, usually obtained from a calibration study, in which the mismeasured covariate is supplemented by a reference measurement.

In this paper we study and compare these three methods in the context of adjusting for measurement error in dietary intakes, in studies relating diet to disease. We investigate studies that have an internal calibration sub-study. For RC, we consider three variants: (i) the usual version of RC; (ii) RC applied only to the reference measurements in the calibration study; and (iii) an 'efficient' version of RC [9] in which the first two estimators are combined in an efficient manner.

Our method of comparing the methods is primarily through simulation in which the models used for simulation are motivated by nutritional epidemiology. We also give a practical illustration of the methods applied to data from the NIH-AARP Diet and Health Cohort study.

Our results show that 'efficient' RC (ERC) is preferable among the methods we compare when there is no reason to suspect differential measurement error. Under this condition, there are cases where ERC is less efficient than MR or IM, but they rarely occur in epidemiology. We show that the efficiency gain of ERC over MR or IM can sometimes be dramatic, and that the price paid by the other methods for relaxing the assumption of non-differential measurement error is high. The usual version of RC carries similar efficiency gains to ERC over MR and IM, but becomes unstable when the measurement error is large, leading to bias and poor precision. When differential measurement error does pertain, then MR and IM have considerably less bias than RC and ERC, but can have much larger variance.

2. METHODS

In this section we describe the methods that we will investigate. We consider the following situation. Let Y be the disease variable. We will allow Y to be either continuous, as for a disease marker, or binary, as for a disease indicator. Let X be the exposure variable(s) of interest. We would like to measure X exactly, but are not able. Instead we measure W , which is X with the error.

The statistical models linking Y , X and W will consist of two parts, the disease model linking Y and X , and the measurement error model linking X and W . When the measurement error is differential, then the latter model will also include Y .

We express the disease model as

$$E(Y|X) = h(\beta_0 + \beta_X X) \quad (1)$$

where h is either the identity function when Y is continuous or the logistic function when Y is binary. Our aim is to estimate β_X as well as possible. In many applications, there will also be covariates in the model that are measured without error. However, the insights that we hope to gain from our study do not require their inclusion here.

We express the non-differential measurement error model as

$$W = \gamma_0 + \gamma_X X + \delta \quad (2)$$

where δ is a residual error with zero expectation that is independent of X and Y . Such a model has been called the non-classical measurement error model to distinguish it from the classical measurement error model where $\gamma_0 = 0$ and $\gamma_X = 1$. The model is motivated by dietary self-report data that appear to conform to this model after a suitable transformation [10]. Later, in this paper we will discuss calibration studies that may be conducted to estimate the parameters of model (2), which need to be known in order to implement the three statistical methods that we now describe. Later, we will also consider differential measurement error models where equation (2) will include some dependence of W on Y .

2.1. Regression calibration

In this method we first estimate the quantity

$$X_{RC}(W) = E(X|W) \quad (3)$$

and then substitute this quantity into the regression model (1) in place of the unknown X , so as to estimate β_X . Under the assumption of non-differential measurement error (i.e. $f[Y|X, W] = f[Y|X]$), the error term δ in (2) is independent of Y , and the resulting estimate of β_X is known to be consistent for linear regression [1, p. 90], and inconsistent, but usually with small bias, for logistic regression [1, pp. 91–92]. The non-differential measurement error assumption is critical here and RC will often give highly biased estimates if this assumption is violated. Standard errors for the estimate of β_X are most easily found by bootstrap methods, although the stacking equations method may be used at the cost of some algebraic and programming work [1, pp. 387–392].

2.2. Moment reconstruction

In this method the aim is to find a quantity $X_{MR}(W, Y)$ that has the same distribution as X and then substitute this quantity into the regression model (1) in place of the unknown X , so as to estimate β_X . Standard errors for the estimate of β_X are most easily found by bootstrap methods. The quantity $X_{MR}(W, Y)$ is constructed so that its first two moments joint with Y are the same as the first two moments of (X, Y) . Freedman *et al.* [5] gave the expression for $X_{MR}(W, Y)$ as

$$E(W|Y) + G\{W - E(W|Y)\}$$

where $G = \{\text{cov}(X|Y)\}^{1/2}\{\text{cov}(W|Y)\}^{-1/2}$. However, this expression was based on the assumption that W follows a classical measurement error model, in which case $E(X|Y) = E(W|Y)$. For non-classical measurement error $E(X|Y) \neq E(W|Y)$, so that a modification is needed to the expression for $X_{\text{MR}}(W, Y)$ so as to preserve the first-moment relationship $E[X_{\text{MR}}(W, Y)] = E(X)$. This is achieved by modifying the definition to

$$X_{\text{MR}}(W, Y) = E(X|Y) + G\{W - E(W|Y)\} \quad (4)$$

with G defined exactly as before, from which the desired equality of the first two moments follows immediately on taking expectations conditional on Y .

Freedman *et al.* [5] demonstrated that when the measurement error model parameters are known, MR is equivalent to RC in linear regression, and in logistic regression with normally distributed covariates the MR estimate of β_X is, unlike RC, consistent. A further potential benefit of MR is that the conditioning on Y ensures that it can successfully handle differential measurement error, where δ is dependent on Y . Furthermore, the method can be used in more complex situations, such as evaluating the impact of measurement error on classification trees [5].

2.3. Stochastic IM

In this method we estimate the quantity $E(X|W, Y)$ and then compute

$$X_{\text{IM}}(W, Y) = E(X|W, Y) + e \quad (5)$$

where e is a random draw from the distribution of residuals from the regression of X on W and Y . We then substitute this quantity into the regression model (1) in place of the unknown X , so as to estimate β_X . Similar to MR, the method can accommodate differential measurement error since both methods condition on Y .

Each method has several variants that can be considered for use. We have chosen to report on variants that we believe are practical and make efficient use of the available data. The details of each variant will be specified below.

3. IMPLEMENTATION

The description of methods in Section 2 is quite general. However, implementation of the methods requires estimation of the measurement error model parameters, based on a calibration study. We consider here the case of an internal calibration study, where the subjects are a random sample of those in the main study sample. We assume that there is a single explanatory variable X that is measured unbiasedly by a 'marker' M in the calibration study. The measurement of M is considerably more expensive than that of W and can be performed only in the smaller calibration study but not in the main study sample. M is related to X by the classical measurement error model:

$$M = X + u \quad (6)$$

where u is a random error with zero expectation, independent of X , W and Y . We assume that M is measured twice on each person in the calibration study and that the random errors u for the two measurements are independent, so that the variance of u can be estimated. We consider the calibration study design in which W and two values of M and the disease variable Y are measured on each person. If Y is not measured in the calibration study, then the MR and IM methods are

not directly available since they require estimates of moments of X conditional on Y . Indirect versions of MR and IM, based on the assumption of the non-differential measurement error, can be constructed in these circumstances, but we will not pursue these here, deferring comment on these until Section 7.

To simplify the description of the implementation we assume that (Y, X, W, M) has a multivariate normal distribution. In the case where Y is binary, we assume that (X, W, M) has a multivariate normal distribution conditional on Y , and also marginally, so that all subsidiary regressions required for our methods are linear. While the conditional and marginal normality assumptions cannot hold simultaneously, they can be approximately true simultaneously when the disease ($Y = 1$) is rare, which we will assume.

3.1. Details

3.1.1. Regression calibration. We consider here three separate estimates that have previously been termed RC in the literature. The first estimate, $\hat{\beta}_{X,RC1}$ is the standard RC estimate when the calibration study is external to the main study.

The second estimate, $\hat{\beta}_{X,RC2}$, is obtained from the RC estimate based on the individuals in the calibration study using their outcome values, Y , and their repeated marker values, M_1 and M_2 , that are assumed to follow the classical measurement error model. This estimate is not usually considered, but is of interest in its own right since it does not employ values of W , and therefore is valid when the measurement error in W is differential. To distinguish it from the other methods, we call this method ‘calibration study RC’.

The third estimate, proposed by Spiegelman *et al.* [9] is a weighted average of $\hat{\beta}_{X,RC1}$ and $\hat{\beta}_{X,RC2}$. The two estimates are weighted by the inverse of their estimated variances; see Appendix A for details. It is expected that this estimate will be more efficient than both $\hat{\beta}_{X,RC1}$ and $\hat{\beta}_{X,RC2}$ and has been termed ERC [9].

We will denote the above three methods by RC1, RC2 and ERC, respectively. Note that the RC1 and ERC estimates are based on the assumption of non-differential measurement error.

3.1.2. Moment reconstruction. $X_{MR}(W, Y) = E(X|Y) + G\{W - E(W|Y)\}$ may be calculated as follows. $E(W|Y)$ and $\text{var}(W|Y)$ are estimated from the main study. $E(X|Y)$ and $\text{var}(X|Y)$ are estimated from the calibration study data, via the regression of \bar{M} on Y using $\hat{E}(X|Y) = \hat{E}(\bar{M}|Y)$ and $\widehat{\text{var}}(X|Y) = \widehat{\text{var}}(\bar{M}|Y) - \widehat{\text{var}}(u)/2$, where $\widehat{\text{var}}(u) = \widehat{\text{var}}(M_1 - M_2)/2$. G is then estimated by $\sqrt{\widehat{\text{var}}(X|Y)/\widehat{\text{var}}(W|Y)}$ and X_{MR} calculated for each individual in the main study. Finally, we estimate $\hat{\beta}_{X,MR}$ as the coefficient of X_{MR} in the regression of Y on X_{MR} in the main study sample.

3.1.3. Stochastic IM. We follow the general approach described in Appendix 2 of Cole *et al.* [8]. For each person in the main study sample who is not also in the calibration study we impute X using $X_{IM}(W, Y) = E(X|W, Y) + e$, whereas for persons in the calibration study, we impute using $X_{IM}(W, Y, \bar{M}) = E(X|W, Y, \bar{M}) + e^*$. In these formulas, e is a random draw from the distribution of residuals in the regression of X on (W, Y) , whereas e^* is a random draw from the distribution of residuals in the regression of X on (W, Y, \bar{M}) . We repeat the procedure K times, thereby creating a total of K imputed sets of covariates $X_{IM}^{(k)}$. For each k from 1 to K , we then regress Y on $X_{IM}^{(k)}$ in the main study to obtain the estimate $\hat{\beta}_{X,IM}^{(k)}$ and the naïve model-based estimate $\widehat{\text{var}}(\hat{\beta}_{X,IM}^{(k)})$ that ignores the fact that X was imputed.

Finally, we estimate β_X as

$$\hat{\beta}_{X,IM} = \frac{1}{K} \sum_{k=1}^K \hat{\beta}_{X,IM}^{(k)}$$

and we estimate $\text{var}(\hat{\beta}_{X,IM})$ as

$$\widehat{\text{var}}(\hat{\beta}_{X,IM}) = \frac{1}{K} \sum_{k=1}^K \widehat{\text{var}}(\hat{\beta}_{X,IM}^{(k)}) + \frac{K+1}{K(K-1)} \sum_{k=1}^K (\hat{\beta}_{X,IM}^{(k)} - \hat{\beta}_{X,IM})^2$$

The full details of the procedure are described in Appendix A. Cole *et al.* [8] in their Appendix 2 comment that their IM procedure is ‘proper’ in the sense of Little and Rubin [6, p. 214]. The method is expected to give unbiased estimates of the model parameters, and confidence intervals with coverage probabilities at the nominal level.

We investigated both $K = 10$ and 40 , the value suggested by Cole *et al.* [8], in our simulations. Parameter estimates with both $K = 10$ and 40 were essentially unbiased. However, the coverage properties of the confidence intervals with $K = 40$ follow the nominal level, whereas with $K = 10$ the coverage is slightly below the nominal level (see Section 4). The results in our tables (which deal with bias and precision of the estimates and not with confidence interval coverage) are those based on $K = 10$.

3.1.4. Further remarks on the methods. It will become apparent when considering the simulation results given below that the above five estimation methods RC1, RC2, ERC, MR and IM actually fall into three classes. Methods RC2, MR and IM derive all or most of their information from the marker (M) and disease (Y) measurements in the internal calibration study. Method RC1 derives most of its information from the exposure (W) and disease (Y) measurements in the main study. Method ERC combines these two types of information in an efficient manner. When viewed in this manner, one could also ask how efficient combinations of RC1 and MR, or of RC1 and IM, would perform. We content ourselves with studying just one combination method (ERC) in this paper, choosing the method that has already appeared in the literature [9].

4. SIMULATIONS

In this section, unless stated otherwise, we take as given that each method yields unbiased or nearly unbiased estimates of β_X . Our main interest is therefore in the precision of the methods, and we use simulations to compare them. We begin with a simulation where the disease model is a linear regression.

4.1. Linear regression

We generated data according to the following model parameters and conditions:

Disease model (1): Link function $h^{-1} = \text{Identity}$; $\beta_0 = 0$; $\beta_X = 0.3$ or 0.6 ; $\text{var}(X) = 1$; Residual error variance = $\text{var}(Y|X) = 0.91$ or 0.64 (corresponding to $\beta_X = 0.3$ or 0.6). These two choices correspond to (Y, X) having a bivariate standard normal distribution with correlation 0.3 or 0.6 .

Measurement error model (2): $\gamma_0=0$; $\gamma_X=0.5, 0.75$ or 1 ; $\text{var}(\delta)=1, 2$, or 4 ; non-differential error ($\delta \perp Y$). The cases of $\text{var}(\delta)=2$ and 4 were run only for $\gamma_X=0.5$, giving five combinations.

'Marker' (M) model (6): $\text{var}(u)=1.0$.

Main study: sample size (N)=1000; one measurement of Y and W per person.

Calibration study: sample size (n)=100; one measurement of Y and W , two measurements of M per person.

Number of simulations per scenario: 1000.

Methods of estimating β_X : RC1, RC2, ERC, MR and IM.

By varying the regression slope in the disease model, and the error variance in the measurement error model, we compare cases where the disease–diet relationship is strong (larger slope) and weak (smaller slope), and where there is a large measurement error (higher error variance) and a small measurement error (lower error variance).

Note also that we do not yet include differential measurement error in these first simulations.

The results of these simulations are shown in Table I. The table shows the precision of $\hat{\beta}_X$ for the various methods in the 10 (2×5) different scenarios. Examination of the table shows that the naïve method was very biased, whereas RC2, ERC, MR and IM methods all had little or no bias. When the measurement error was high then the ERC estimate was underestimated but by less than 10 per cent in our examples. The precision of the ERC estimate was generally greater than or similar to that of the RC2, MR and IM estimates. For example, when $\gamma_X=1$, $\text{var}(\delta)=1$ and $\beta_X=0.6$, then the empirical standard deviation of the ERC estimate was approximately one-half of those for the RC2, MR and IM estimates. When the measurement error was larger ($\text{var}(\delta)=2$ and 4), then the precisions of the ERC, RC2, MR and IM estimates were closer to each other.

The table also shows the advantage of ERC over the standard RC estimate, RC1. It may be seen that the latter becomes badly biased with inflated standard error as the measurement error increases. However, its combination with the RC2 estimate (yielding the ERC estimate) stabilizes the estimation procedure and yields standard errors considerably smaller than those of the component parts. Note that the good results for ERC occur partly because the bias of RC1 and its standard error increase together as the measurement error increases. As ERC is a weighted average of RC1 and RC2 with weights equal to the inverse of their variance, the RC1 estimate has an ever-decreasing influence on ERC as its bias grows.

We also examined the empirical coverage of the multiple IM estimated confidence interval for β_X . With 40 multiply imputed data sets, the coverage was close to the nominal 95 per cent for all 10 scenarios considered in Table I (range: 94.1–95.7 per cent). With only 10 multiply imputed data sets, the coverage was slightly below the nominal level (range: 91.0–93.8 per cent). Full data are available from the authors.

4.2. Logistic regression

We consider two types of study that may be analyzed by logistic regression: case–control and cohort studies. In case–control studies dietary ascertainment is done after the occurrence (or not) of the disease, resulting in greater opportunity for differential error. Therefore, for this design we simulate scenarios with differential error as well as non-differential error. In cohort studies dietary intake is assessed before any occurrence of the disease, and differential error is much less likely. Therefore, for this design we simulate only non-differential error. Moreover, the sample sizes of these two designs are typically very different. We therefore generated data according to

Table 1. Empirical means of $\hat{\beta}_X$ (with empirical standard deviations in parentheses) in 1000 simulations where the disease model is linear, estimated by naive linear regression of Y on W , regression calibration for W in the main study (RC1), regression calibration for \bar{M} in the calibration study (RC2), efficient regression calibration (ERC), moment reconstruction (MR) and multiple imputation (IM); main study sample size = 1000, calibration sample size = 100.

β_X	γ_X	σ_ϵ^2	ρ_{XW}^*	Naive	RC1	RC2	ERC	MR	IM
0.3	1	1	0.71	0.150 (0.022)	0.307 (0.071)	0.312 (0.135)	0.303 (0.061)	0.310 (0.129)	0.301 (0.115)
0.3	0.75	1	0.60	0.144 (0.025)	0.309 (0.077)		0.299 (0.068)	0.301 (0.127)	0.298 (0.118)
0.3	0.5	1	0.45	0.119 (0.028)	0.323 (0.131)		0.293 (0.084)	0.301 (0.125)	0.296 (0.122)
0.3	0.5	2	0.33	0.067 (0.022)	0.363 (0.422)		0.285 (0.100)	0.300 (0.122)	0.294 (0.121)
0.3	0.5	4	0.24	0.036 (0.015)	0.563 (3.498)		0.279 (0.106)	0.297 (0.118)	0.287 (0.121)
0.6	1	1	0.71	0.300 (0.020)	0.614 (0.104)	0.605 (0.138)	0.595 (0.083)	0.604 (0.124)	0.590 (0.121)
0.6	0.75	1	0.60	0.289 (0.023)	0.621 (0.121)		0.600 (0.095)	0.609 (0.127)	0.596 (0.126)
0.6	0.5	1	0.45	0.239 (0.027)	0.654 (0.252)		0.593 (0.115)	0.612 (0.122)	0.599 (0.124)
0.6	0.5	2	0.33	0.133 (0.020)	0.745 (0.909)		0.585 (0.119)	0.609 (0.124)	0.594 (0.125)
0.6	0.5	4	0.24	0.071 (0.016)	0.984 (4.524)		0.581 (0.130)	0.610 (0.124)	0.594 (0.125)

* ρ_{XW} is the correlation of X and W .

the following model parameters and conditions:

Disease model (1): Link function h^{-1} =Logistic.

Cohort: $\beta_0 = -2.2$; $\beta_X = 0.3$ or 0.6 ; $\text{var}(X) = 1$ and X has normal distribution marginally. (Since we consider only rare disease, X is approximately normally distributed both among the controls and among the cases. The value $\beta_0 = -2.2$ ensures that approximately 10 per cent of the cohort participants are cases. The value $\beta_X = 0.3$ (0.6) corresponds to an odds ratio of 2.2 (4.6) between the 90th and 10th percentiles of X .)

Case-control: Data are generated from the above cohort model, and cases and controls are randomly selected so that there are equal numbers of each in the case-control study.

Measurement error model (2):

Cohort, non-differential error ($\delta \perp Y$): As for linear regression $\gamma_0 = 0$, $\gamma_X = 0.5$, $\text{var}(\delta) = 1, 2$ or 4 .

Case-control, non-differential error: As above.

Case-control, differential error: Three simulations with $\beta_X = 0.3$, as follows (the extra suffix in the symbols below denotes case/control status, with 0 =control and 1 =case):

$$\gamma_{00} = 0, \quad \gamma_{01} = 0.25, \quad \gamma_{X0} = \gamma_{X1} = 1, \quad \text{var}_0(\delta) = \text{var}_1(\delta) = 1$$

$$\gamma_{00} = \gamma_{01} = 0, \quad \gamma_{X0} = 0.5, \quad \gamma_{X1} = 1, \quad \text{var}_0(\delta) = \text{var}_1(\delta) = 1$$

$$\gamma_{00} = \gamma_{01} = 0, \quad \gamma_{X0} = \gamma_{X1} = 1, \quad \text{var}_0(\delta) = 2, \quad \text{var}_1(\delta) = 0.5$$

'Marker' (M) model (6): $\text{var}(u) = 1$, as for linear regression above.

Main study

Cohort: sample size (N)=100000; one measurement of Y and W per person.

Case-control: sample size (N)=1000 cases and 1000 controls; one measurement of Y and W per person.

Calibration study

Cohort: sample size (n)=1000; one measurement of Y and W , two measurements of M per person.

Case-control: sample size (n)=100 cases and 100 controls; one measurement of Y and W , two measurements of M per person.

Number of simulations per scenario: 1000

Methods of estimating β_X : RC1, RC2, ERC, MR, IM.

The results are shown in Table II and show similar trends to those seen in Table I. For cohort studies with non-differential measurement error, the RC1 and ERC estimates are more precise than those of the RC2, MR and IM methods, sometimes dramatically so. The RC1 and ERC estimates are subject to mild bias especially for larger exposure effects ($\beta_X = 0.6$), but in our simulations the bias was less than 5 per cent of the estimate. For larger exposure effects ($\beta_X = 0.6$) and a higher degree of measurement error ($\text{var}(\delta) = 4$), the ERC estimate is considerably more precise than the RC1 estimate.

For case-control studies with non-differential measurement error, ERC is once again more efficient than the RC2, MR and IM methods. The bias in the ERC estimate is a little higher than that in the cohort study simulations but still remains below 10 per cent of the estimate. When the measurement error is large ($\text{var}(\delta) = 4$), the advantage of ERC over MR and IM lessens, but the mean-squared error of the ERC estimate remains smaller than that of the MR and IM estimates. Note that in these simulations standard RC (RC1) is quite biased, particularly as the exposure effect and the measurement error increases.

Table II. Empirical means of $\hat{\beta}_X$ (with empirical standard deviations in parentheses) in 1000 simulations when the disease model is logistic, estimated by different methods: naive logistic regression of Y on W , regression calibration for W in the main study (RC1), regression calibration for \bar{M} in calibration sub-study (RC2), efficient regression calibration (ERC), moment reconstruction (MR) and multiple imputation (IM).

β_X	Γ_{0y}	γ_{Xy}	$\sigma_{\delta y}^2$	ρ_{XW^*}	Naive	RC1	RC2	ERC	MR	IM
Cohort, non-differential error ($N = 100000, n = 1000$)										
0.3	0	0.5	1	0.45	0.119 (0.010)	0.300 (0.034)	0.302 (0.131)	0.299 (0.034)	0.303 (0.132)	0.299 (0.128)
0.3	0	0.5	2	0.33	0.066 (0.007)	0.302 (0.050)		0.300 (0.047)	0.305 (0.133)	0.307 (0.135)
0.3	0	0.5	4	0.24	0.035 (0.005)	0.303 (0.066)		0.297 (0.059)	0.298 (0.136)	0.297 (0.138)
0.6	0	0.5	1	0.45	0.233 (0.009)	0.585 (0.053)	0.601 (0.130)	0.585 (0.050)	0.607 (0.134)	0.603 (0.129)
0.6	0	0.5	2	0.33	0.129 (0.007)	0.587 (0.077)		0.585 (0.066)	0.609 (0.133)	0.603 (0.135)
0.6	0	0.5	4	0.24	0.068 (0.005)	0.592 (0.108)		0.582 (0.078)	0.614 (0.137)	0.608 (0.143)
Case/control, non-differential error ($N = 2000, n = 200$)										
0.3	0	0.5	1	0.45	0.121 (0.042)	0.324 (0.161)	0.317 (0.199)	0.298 (0.114)	0.310 (0.201)	0.310 (0.196)
0.3	0	0.5	2	0.33	0.065 (0.031)	0.381 (0.860)		0.282 (0.139)	0.301 (0.197)	0.301 (0.198)
0.3	0	0.5	4	0.24	0.035 (0.022)	0.419 (1.150)		0.276 (0.152)	0.305 (0.187)	0.303 (0.193)
0.6	0	0.5	1	0.45	0.231 (0.040)	0.683 (1.247)	0.629 (0.213)	0.590 (0.166)	0.618 (0.220)	0.608 (0.210)
0.6	0	0.5	2	0.33	0.131 (0.031)	0.666 (2.203)		0.586 (0.179)	0.623 (0.212)	0.618 (0.223)
0.6	0	0.5	4	0.24	0.068 (0.021)	0.814 (16.94)		0.577 (0.201)	0.622 (0.217)	0.612 (0.222)
Case/control, differential error ($N = 2000, n = 200$)										
0.3	0/0.25	1	1		0.276 (0.033)	0.565 (0.108)	0.317 (0.199)	0.494 (0.095)	0.311 (0.192)	0.303 (0.165)
0.3	0	0.5/1	1		0.175 (0.035)	0.469 (0.178)	0.307 (0.188)	0.374 (0.114)	0.301 (0.182)	0.298 (0.166)
0.3	0	1	2/0.5		0.132 (0.031)	0.414 (0.140)	0.302 (0.192)	0.364 (0.110)	0.298 (0.188)	0.291 (0.165)

* ρ_{XW} is the correlation of X and W .

For case-control studies with differential measurement error, Table II shows that RC1 and ERC estimates have considerable bias, but will often have smaller variance than the almost unbiased RC2, MR and IM methods. In these situations the trade-off between bias and precision will have to be weighed.

In these simulations, we again examined the empirical coverage of the multiple IM estimated confidence interval for β_X for the six scenarios of case-control studies with non-differential measurement error that are listed in Table II. As with the linear regression model, we found that with 40 multiply imputed data sets the coverage was close to the nominal 95 per cent (range: 94.2–94.9 per cent), but that with only 10 multiply imputed data sets, the coverage was slightly below the nominal level (range: 92.0–94.3 per cent).

5. ASYMPTOTIC VARIANCES IN A SIMPLIFIED SITUATION

Some aspects of the results presented in Section 4 were surprising to us and not easily understood intuitively. It was particularly surprising that according to Tables I and II the precision of estimates from the MR and IM methods appeared insensitive to the measurement error model parameters, in contrast to the RC1 and ERC estimates. In order to gain better insight, and also as a check on our results, we developed asymptotic expressions for the standard error of $\hat{\beta}_X$ in linear regression of the methods in a slightly simpler context than our simulations, where M is an exact measure of X , that is, where $\text{var}(u) = 0$. The asymptotic expressions are based on the assumption that the main study sample size N is very large, and much larger than n , the sample size of the calibration study. Thus, the expressions do not include terms of order $1/N$, which is much smaller than the dominant term of order $1/n$. It turns out that the expressions are simple functions of two correlations, ρ_{XW} and ρ_{XY} . The expressions for the standard errors are given below and outlines of proofs are provided in Appendix B. Furthermore, the expressions agree well with empirical results of simulations (see Appendix B).

Standard regression calibration (RC1):

$$\text{SE}(\hat{\beta}_X) = \sqrt{\frac{\beta_X^2(1 - \rho_{XW}^2)}{n\rho_{XW}^2}}$$

Calibration study RC (RC2):

$$\text{SE}(\hat{\beta}_X) = \sqrt{\frac{\beta_X^2(1 - \rho_{XY}^2)}{n\rho_{XY}^2}}$$

ERC:

$$\text{SE}(\hat{\beta}_X) = \sqrt{\frac{\beta_X^2(1 - \rho_{XW}^2)(1 - \rho_{XY}^2)}{n(\rho_{XW}^2 + \rho_{XY}^2 - 2\rho_{XW}^2\rho_{XY}^2)}}$$

Moment reconstruction (MR):

$$SE(\hat{\beta}_X) = \sqrt{\frac{\beta_X^2(1-\rho_{XY}^2)}{n\rho_{XY}^2}(1-2\rho_{XY}^2+2\rho_{XY}^4)}$$

where ρ_{XY} is the correlation coefficient between X and Y .

Imputation (IM):

$$SE(\hat{\beta}_X) = \sqrt{\frac{\beta_X^2(1-\rho_{X|Y,W}^2)}{n\rho_{XY}^2}(1-2\rho_{XY}^2+2\rho_{XY}^2\rho_{X|Y,W}^2)}$$

where $\rho_{X|Y,W}^2$ is the multiple correlation coefficient between X and (Y, W) , $(1-\rho_{X|Y,W}^2) = (1-\rho_{XY}^2)(1-\rho_{XW}^2)/(1-\rho_{YW}^2)$, and, under the assumption of non-differential error, $\rho_{WY}^2 = \rho_{XW}^2\rho_{XY}^2$.

These formulas confirm that the asymptotic standard error for the RC1 and ERC estimators are dependent on parameters ρ_{XW} and ρ_{XY} , whereas those of the RC2 and MR estimators are entirely independent of ρ_{XW} . The reason is clear for RC2. It occurs for MR because the definition of X_{MR} includes calibration study information only on $E(X|Y)$ and $\text{var}(X|Y)$, and these are estimated from data on X and Y , but not W .

Although the asymptotic standard error of the IM estimator does involve ρ_{XW} (through the expression $\rho_{X|Y,W}^2$), one can in fact show that it is bounded above by the asymptotic standard error of the MR estimator. This follows directly from the observation that $\rho_{X|Y,W}^2 \geq \rho_{XY}^2$. In the worst-case scenario that W provides no information about X , we have $\rho_{X|Y,W}^2 = \rho_{XY}^2$, and the standard errors for the MR and IM estimators will be asymptotically equal. This demonstrates that, using our implementation methods, IM is asymptotically superior to MR. It is also clear from the formulas that both MR and IM are asymptotically superior to RC2. As shown in the simulations of Section 4, the difference between RC2, MR and IM is, in practice, often small.

Tabulating these expressions for different values of ρ_{XW} and ρ_{XY} is informative (Table III). In this table we have chosen the values of β_X and n in each row to ensure that $SE(\hat{\beta}_{X,RC})$ equals 1, enabling simple comparison between the standard errors of each method. The table shows that ERC is asymptotically superior to MR and IM when correlations between X and Y are low (0.2), as they generally are in epidemiological studies with binary outcomes. The advantage to ERC increases as the measurement error decreases. A clear asymptotic advantage to MR or IM over ERC is seen only for a high correlation between X and Y (0.6 or higher) in combination with large measurement error (i.e. low correlation between X and W).

6. APPLICATION TO AARP STUDY

The NIH-AARP Diet and Health study is a large cohort consisting of 550 644 individuals (325 176 men and 225 468 women) over the age of 50 years, who completed a food frequency questionnaire (FFQ) in 1995–1996 and have since been followed for mortality and cancer incidence. Details of the study are provided by Schatzkin *et al.* [11]. We examine the question whether dietary fat

Table III. Ratio of standard errors for standard regression calibration (RC1), calibration study RC (RC2), moment reconstruction (MR) and multiple imputation (IM) relative to efficient regression calibration (ERC), as a function of correlations between the true covariate X and outcome variable Y and between X and the measured covariate W .

ρ_{XY}^*	ρ_{XW}^\dagger	RC1	RC2	ERC	MR	IM
0.20	0.20	1.41	1.41	1	1.35	1.32
0.20	0.40	1.10	2.36	1	2.29	2.12
0.20	0.60	1.04	3.81	1	3.69	3.04
0.20	0.80	1.01	6.61	1	6.40	4.00
0.40	0.20	2.36	1.10	1	0.95	0.93
0.40	0.40	1.41	1.41	1	1.21	1.14
0.40	0.60	1.16	1.99	1	1.72	1.50
0.40	0.80	1.05	3.21	1	2.77	1.94
0.60	0.20	3.81	1.04	1	0.76	0.76
0.60	0.40	1.99	1.16	1	0.85	0.84
0.60	0.60	1.41	1.41	1	1.03	0.99
0.60	0.80	1.15	2.04	1	1.49	1.24
0.80	0.20	6.61	1.01	1	0.74	0.74
0.80	0.40	3.21	1.05	1	0.77	0.77
0.80	0.60	2.04	1.15	1	0.84	0.83
0.80	0.80	1.41	1.41	1	1.03	0.93

* ρ_{XY} is the correlation between X and Y .

† ρ_{XW} is the correlation between X and W .

intake is related to mortality. At the time of the analysis, subjects had been followed for a median of 9.6 years, and 65 168 subjects (44 445 men and 20 723 women) had died.

The internal calibration sub-study comprised 1953 subjects (987 men and 966 women), who in addition to completing the FFQ also completed at least one of the two 24-h recalls (24HR) (1890 completed both), to be used as reference measurements. Details of the calibration sub-study are provided by Thompson *et al.* [12]. At the time of the analysis, 208 subjects (114 men and 94 women) in the calibration sub-study had died.

For illustration of our methods, we estimate the parameters in a logistic regression of mortality (Y) on the logarithm of per cent calories from fat in the diet (X) and age. Reported exposure (W) is log per cent calories from fat as measured by the FFQ, and the reference measurements (M_1 and M_2) are log per cent calories from fat as measured by the two 24HR. Note that there is doubt over whether the 24HR measurements will indeed conform to the classical measurement error model (6), but currently there is no measure of fat intake available that is known to be a valid reference measurement (i.e. unbiased with errors that are uncorrelated with Y , X and W).

Prior to the analysis, we excluded 5034 subjects (976 deaths) who reported dietary intakes that were determined to be outliers of W or FFQ log total caloric intake. None of the excluded subjects were in the calibration sub-study. For subjects in the calibration sub-study, we excluded 20 values of M_1 and 23 values of M_2 that were also determined to be outliers. Outliers were defined to be values that fell below the 25th percentile of the distribution of the variable minus two interquartile ranges or above the 75th percentile plus two interquartile ranges.

We estimated parameters in the logistic regression of mortality on log per cent calories from fat and age, separately for men and women, using six different methods: naïve regression of Y on

W, RC1, RC2, ERC, MR and IM. Standard errors were estimated using a bootstrap method with 100 replications.

Table IV presents the estimates of the coefficient for log per cent calories from fat. The naïve estimate indicates a moderate association with an odds ratio of 1.7–2.0 ($\exp(0.55) - \exp(0.71)$) for a 2.7-fold = $\exp(1.0)$ increase in per cent fat intake. This association is highly statistically significant ($z > 20$ for both men and women), because of the very large sample size. Adjustment for the measurement error by RC1 or ERC indicates an even stronger association with an odds ratio of 2.9–4.7 ($\exp(1.06) - \exp(1.54)$) for a 2.7-fold increase in per cent fat intake, which is still highly statistically significant ($z > 10$ for both men and women). However, the MR and IM method estimates have standard errors that are 5–10 times larger than that of the RC1 or ERC estimate, and consequently conventional statistical significance ($z > 1.96$) is no longer seen.

This result appears even stronger than that in the first row of the simulated cohort studies seen in Table II, where standard errors of the MR and IM estimates were approximately 4 times larger than that of the ERC estimate. The key to linking the AARP result to the simulations lies in considering the values of ρ_{XY} and ρ_{XW} for the two cases. In a logistic regression model, a covariate X has a correlation with Y approximated by $\beta_X \sqrt{\text{var}(X)\text{var}(Y)}$. In the simulation in question $\beta_X \sqrt{\text{var}(X)\text{var}(Y)}$ is $0.3 \times 1.0 \times 0.3 = 0.09$ and $\rho_{XW} = 0.45$. In AARP, $\beta_X \sqrt{\text{var}(X)\text{var}(Y)}$ is, for women, $1.54 \times 0.21 \times 0.29 = 0.09$ and ρ_{XW} is 0.64. Similar values are found for men. Thus, in the AARP study although the value of ρ_{XY} is very similar to that of the simulation, the value of ρ_{XW} is larger. Using the asymptotic formulas in Section 5, one may predict that the standard error for MR and IM estimates will be approximately 9 and 7 times, respectively, the standard error for RC1 or ERC, which is not far from the observed ratios of 6.3 and 5.3 seen for women in Table IV.

Our main conclusion is that the MR and IM methods in this context are grossly inefficient compared with RC1 or ERC. The finding, using RC1 or ERC, of a possibly highly important association between per cent fat intake and total mortality needs further examination. Confounding with other factors needs to be considered. The association may partly reflect the known fat intake–cholesterol–heart disease pathway, which may be studied by examining the association for selected causes of death.

7. DISCUSSION

We have described and compared three substitution methods for correcting regression coefficients for measurement error in the covariates, in the context of nutritional epidemiologic studies. We note that in place of our term ‘substitution’, we could have used the word ‘imputation’ (in its general sense), but to do so may have caused confusion. In fact, RC corresponds to the conditional mean IM method described by Little [13], but it is not clear where MR would fit into the array of current IM methods.

We have considered in this paper the case where the calibration study includes information on the disease variable Y . Sometimes this information is not available in the calibration study. In these cases, among the methods we have described, only RC1 is available, as others require knowledge of Y in the calibration study.

The ‘efficient’ version of RC (ERC) that we have used appeared in our simulations to offer a considerable advantage over the usual RC estimator (denoted by RC1 in our tables). Tables I and II show the large advantage of ERC over standard RC when the measurement error is large. We

Table IV. AARP study: estimated regression coefficient for log per cent calories from fat in a logistic regression of mortality on log per cent calories from fat and age, using six methods of estimation: naïve logistic regression of mortality on FFQ, regression calibration for FFQ in the main study (RC1), regression calibration for 24HR in the calibration sub-study (RC2), efficient regression calibration (ERC), moment reconstruction (MR) and multiple imputation (IM); standard errors* are in parentheses.

Gender	<i>N</i> (No. of deaths in main study)	<i>n</i> (No. of deaths in calibration sub-study)	Naïve	RC1	RC2	ERC	MR	IM
Male	322 321 (43 810)	987 (114)	0.55 (0.02)	1.06 (0.07)	1.27 (0.72)	1.06 (0.07)	1.37 (0.75)	1.17 (0.74)
Female	223 278 (20 382)	966 (94)	0.71 (0.03)	1.54 (0.14)	1.17 (0.75)	1.53 (0.13)	1.14 (0.82)	0.74 (0.69)

*Standard errors are estimated using a non-parametric bootstrap method with 100 replications.

also found in simulations not reported here that ERC was preferable to using Fuller's small-sample correction for RC [14].

When the calibration study includes information on disease, and non-differential error pertains, then ERC appears more efficient than MR and IM in almost all of the situations that we have examined in our simulations. The simulations indicated that the gap between the methods narrows as the measurement error variance increases, but we found only one case where the standard error for the ERC estimate was larger than that of the other estimates. Our asymptotic results indicate that when the correlation between X and Y is high ($\rho_{XY} \geq 0.6$) and measurement error is high ($\rho_{XW} < 0.4$), MR and IM can hold an advantage over ERC, but the usual situation in epidemiology is the reverse, with a correlation between X and Y that is low and less than the correlation between X and W . In fact, in a cohort study with a disease prevalence of 10 per cent and a normally distributed covariate, a value of ρ_{XY} equal to 0.6 would correspond to a relative risk of 160 between the upper and lower quintiles.

We believe that the efficiency advantage provided by ERC stems primarily from its assumption of non-differential measurement error. RC2, MR and IM do not make this assumption, and payment in the form of increased variance is extracted for the privilege. In some cases the payment is very high. It is in fact possible, although more complex, to construct versions of MR and IM that are based on the assumption of non-differential measurement error and do not use the knowledge of Y in the calibration study. We have studied this separately and have found in simulations, not reported here, that they perform very similarly to ERC. This reinforces our view that the increased variance of the MR and IM estimates (differential error version) relative to the ERC estimates indeed results from relaxing the non-differential measurement error assumption. The rare cases where MR and IM improve on ERC will occur through the former methods' use of Y , which, if it is highly correlated with X , can supply important extra information for estimating X .

We note that ERC can be viewed as an efficient linear combination of the usual RC estimator (RC1) and an RC estimator applied to the marker data in the calibration study (RC2). The insight that MR and IM also derive most of their information from the marker data in the calibration study, leads to the suggestion of combining the RC1 estimator with MR or with IM, instead of with RC2. As MR and IM will generally have somewhat greater precision than RC2, one would expect the resulting combined estimators to have slightly greater precision than ERC. We have not pursued this line here, as we made our aim to compare methods that have been proposed in the literature, but it is of interest to do so. Examining the combination of RC1 with IM would seem most worthwhile, firstly because IM is slightly more precise than MR and, secondly, because the variance of MR can be determined only by bootstrap, making it more cumbersome to obtain the best weights for the linear combination.

When the differential measurement error pertains, then RC2, MR and IM have considerably less bias than ERC, but can have much larger variance, and the decision which to use has to be weighed according to the expected degree of the bias arising in the ERC method. In the important case of prospective studies, however, differential measurement is less likely and the decision regarding which method to use can be based on the estimated variances, as in the AARP example presented.

The methods of MR and IM perform similarly, but IM has greater precision in some circumstances. Theoretical results indicate that, asymptotically, IM is always as efficient, or more efficient than MR. These results are supported by our simulations, although in many cases there is little practical difference between the two methods. One advantage of the IM estimate is the ability to obtain direct estimates of the standard error without resorting to use of the bootstrap. We found

that the confidence intervals for the model parameters had good coverage properties if they were based on 40 multiply imputed data sets.

APPENDIX A: IMPLEMENTATION OF RC AND MULTIPLE IM

A.1. Regression calibration

The ERC estimate is a weighted average of two available RC estimates of β_X . The first estimate, $\hat{\beta}_{X,RC1}$ is obtained by (i) estimating the linear regression $E(\bar{M}|W) = \lambda_0 + \lambda_1 W$ in the calibration study to get estimates $\hat{\lambda}_0$ and $\hat{\lambda}_1$; (ii) calculating for each individual in the main study, $X_{RC1} = \hat{\lambda}_0 + \hat{\lambda}_1 W$; and (iii) estimating the coefficient of X_{RC1} in the regression of Y on X_{RC1} in the main study sample. For case/control studies, we use only the controls to obtain $\hat{\lambda}_0$ and $\hat{\lambda}_1$ in step (i).

This is the usual RC estimate when the calibration study is external to the main study. However, as we have an internal calibration study, we can improve upon this estimate.

The second estimate, $\hat{\beta}_{X,RC2}$, is obtained by (i) estimating $E(\bar{M})$, $\text{var}(\bar{M})$ and $\text{var}(u) = \text{var}(M_2 - M_1)/2$ in the calibration study, where \bar{M} is the mean of the two determinations of M ; (ii) calculating $\widehat{\text{var}}(X) = \widehat{\text{var}}(\bar{M}) - \widehat{\text{var}}(u)/2$ and $\hat{\lambda}_M = \widehat{\text{var}}(X)/\widehat{\text{var}}(\bar{M})$; (iii) for each individual in the calibration study, calculating $X_{RC2} = \hat{E}(\bar{M}) + \hat{\lambda}_M \{\bar{M} - \hat{E}(\bar{M})\}$; and (iv) estimating the coefficient of X_{RC2} in the regression of Y on X_{RC2} in the calibration study. For case/control studies, we use only the controls to estimate $E(\bar{M})$ and $\text{var}(\bar{M})$ in step (i).

Finally, we combine the two estimates of β_X as follows: (i) we estimate variances of $\hat{\beta}_{X,RC1}$ and $\hat{\beta}_{X,RC2}$, using formulas described in Spiegelman *et al.* [10] and Rosner *et al.* [15]; (ii) we calculate the weight $w_{RC} = \widehat{\text{var}}(\hat{\beta}_{X,RC1})^{-1} / \{\widehat{\text{var}}(\hat{\beta}_{X,RC1})^{-1} + \widehat{\text{var}}(\hat{\beta}_{X,RC2})^{-1}\}$; and we calculate $\hat{\beta}_{X,RC}$ as the weighted average: $\hat{\beta}_{X,RC} = w_{RC} \hat{\beta}_{X,RC1} + (1 - w_{RC}) \hat{\beta}_{X,RC2}$.

A.2. Multiple IM

For each person in the main study sample who is not also in the calibration study we impute X using $X_{IM}(W, Y) = E(X|W, Y) + e$, whereas for persons in the calibration study, we impute using $X_{IM}(W, Y, \bar{M}) = E(X|W, Y, \bar{M}) + e^*$. In these formulas, e is a random draw from the distribution of residuals in the regression of X on (W, Y) , whereas e^* is a random draw from the distribution of residuals in the regression of X on (W, Y, \bar{M}) .

Assuming that X has a normal distribution conditional on W and Y , and that u in (5) has a normal distribution, then

$$\left(\begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \middle| W, Y \right) \sim N \left\{ \begin{pmatrix} \mu(\alpha; W, Y) \\ \mu(\alpha; W, Y) \end{pmatrix}, \begin{pmatrix} \Sigma_{11}(\theta; W, Y) & \Sigma_{12}(\theta; W, Y) \\ \Sigma_{12}(\theta; W, Y) & \Sigma_{11}(\theta; W, Y) \end{pmatrix} \right\} \tag{A1}$$

where $\mu(\alpha; W, Y) = E(X|W, Y)$, $\Sigma_{12}(\theta; W, Y) = \text{var}(X|W, Y)$ and $\Sigma_{11}(\theta; W, Y) = \text{var}(X|W, Y) + \text{var}(u)$. Then $(X|W, Y) \sim N(\mu(\alpha; W, Y), \Sigma_{12}(\theta; W, Y))$.

In addition, $(X|W, Y, \bar{M}) \sim N(\mu(\alpha; W, Y) + R(\theta; W, Y)\{\bar{M} - \mu(\alpha; W, Y)\}, \Sigma_{12}(\theta; W, Y)\{1 - R(\theta; W, Y)\})$ where

$$R(\theta; W, Y) = \text{var}(X|W, Y) / \text{var}(\bar{M}|W, Y) = 2\Sigma_{12}(\theta; W, Y) / \{\Sigma_{11}(\theta; W, Y) + \Sigma_{12}(\theta; W, Y)\}$$

The multiple IM procedure is therefore as follows:

(i) Fit model (A1), where μ , Σ_{11} and Σ_{12} are known functions (defined below) of unknown parameter vectors α or θ , in the calibration study to obtain estimates $\hat{\alpha}$, $\hat{\theta}$, $\widehat{\text{cov}}(\hat{\alpha})$ and $\widehat{\text{cov}}(\hat{\theta})$.

(ii) For $k=1$ to K IMs,

(a) Generate a random draw of the parameter estimates: $\alpha^{(k)} \sim N(\hat{\alpha}, \widehat{\text{cov}}(\hat{\alpha}))$, $\theta^{(k)} \sim N(\hat{\theta}, \widehat{\text{cov}}(\hat{\theta}))$.

(b) For each individual in the main study but not in the calibration study, generate $e^{(k)} \sim N(0, \Sigma_{12}(\theta^{(k)}; W, Y))$ and calculate $X_{\text{IM}}^{(k)} = \mu(\alpha^{(k)}; W, Y) + e^{(k)}$.

(c) For each individual in the calibration study, generate $e^{*(k)} \sim N(0, \Sigma_{12}(\theta^{(k)}; W, Y)\{1 - R(\theta^{(k)}; W, Y)\})$ and calculate $X_{\text{IM}}^{(k)} = \mu(\alpha^{(k)}; W, Y) + R(\theta^{(k)}; W, Y)\{\bar{M} - \mu(\alpha^{(k)}; W, Y)\} + e^{*(k)}$.

(c) Regress Y on $X_{\text{IM}}^{(k)}$ in the main study to obtain the estimate $\hat{\beta}_{X,\text{IM}}^{(k)}$ and the naïve model-based estimate $\widehat{\text{var}}(\hat{\beta}_{X,\text{IM}}^{(k)})$ (that ignores the fact that X was imputed).

(iii) Estimate β_X as $\hat{\beta}_{X,\text{IM}} = (1/K) \sum_{k=1}^K \hat{\beta}_{X,\text{IM}}^{(k)}$.

(iv) Estimate $\text{var}(\hat{\beta}_{X,\text{IM}}^{(k)})$ as

$$\widehat{\text{var}}(\hat{\beta}_{X,\text{IM}}) = \frac{1}{K} \sum_{k=1}^K \widehat{\text{var}}(\hat{\beta}_{X,\text{IM}}^{(k)}) + \frac{K+1}{K(K-1)} \sum_{k=1}^K (\hat{\beta}_{X,\text{IM}}^{(k)} - \hat{\beta}_{X,\text{IM}})^2$$

For continuous Y , the mean and variance functions are

$$\begin{aligned} \mu(\alpha; W, Y) &= \alpha_0 + \alpha_1 W + \alpha_2 Y \\ \Sigma_{11}(\theta; W, Y) &= \exp(\theta_1) \\ \Sigma_{12}(\theta; W, Y) &= \exp(\theta_1) \left(\frac{\exp(\theta_2)}{1 + \exp(\theta_2)} \right) \end{aligned}$$

This parameterization is used to ensure that the estimates of variance are always positive.

For binary Y , the mean and variance functions are

$$\begin{aligned} \mu(\alpha; W, Y) &= \alpha_{Y0} + \alpha_{Y1} W \\ \Sigma_{11}(\theta; W, Y) &= \exp(\theta_{Y1}) \\ \Sigma_{12}(\theta; W, Y) &= \exp(\theta_{Y1}) \left(\frac{\exp(\theta_{Y2})}{1 + \exp(\theta_{Y2})} \right) \end{aligned}$$

APPENDIX B: THEORY FOR UNDERSTANDING THE RESULTS OF THE SIMULATIONS IN SECTION 4

Assume the model

$$\begin{aligned} Y &= \beta_0 + \beta_X X + \varepsilon \\ W &= \gamma_0 + \gamma_X X + \delta \end{aligned}$$

as in models (1) and (2) of the main text.

Assume that Y and W are measured in N individuals where N is very large.

Assume also that Y , W and X are measured in an independent sub-study of n individuals where n is much smaller than N .

This is not exactly the same situation as we simulated (e.g. we assume here that we can measure X exactly, whereas in the simulations we had repeat measurements of an unbiased of X), but we think it is close enough to give us insight into the results of the simulations in Section 4.

Define the following quantities:

$$\hat{X}_{RC} = \frac{\widehat{\text{cov}}_n(X, W)}{\widehat{\text{var}}_n(W)} W$$

$$\hat{X}_{MR} = \hat{E}_n(X|Y) + \sqrt{\frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)}} (W - E_N(W|Y))$$

$$\hat{X}_{IM} = \hat{E}_n(X|Y, W) + \eta \quad \text{where } \eta \sim N(0, \widehat{\text{var}}_n(X|Y, W))$$

where subscript n denotes that the quantity is being evaluated in the calibration sub-study.

Estimates of β_X considered in our paper are given by the OLS regressions of Y on \hat{X}_{RC} , \hat{X}_{MR} and \hat{X}_{IM} in the main study ($\hat{\beta}_{X,RC1}$, $\hat{\beta}_{X,MR}$, $\hat{\beta}_{X,IM}$, respectively). In addition, an estimate of β_X is given by regressing Y on X in the calibration sub-study ($\hat{\beta}_{X,RC2}$). The ERC estimate that we consider in this paper is given by

$$\hat{\beta}_{X,RC} = \frac{\hat{\beta}_{X,RC1}/\widehat{\text{var}}_1 + \hat{\beta}_{X,RC2}/\widehat{\text{var}}_2}{1/\widehat{\text{var}}_1 + 1/\widehat{\text{var}}_2}$$

where $\widehat{\text{var}}_i$ is the estimated variance of $\hat{\beta}_{X,RCi}$ ($i = 1, 2$). We assume in the following that these variances are estimated sufficiently accurately to ignore their own uncertainty. In this case,

$$\text{var}(\hat{\beta}_{X,RC}) = \frac{1}{1/\text{var}_1 + 1/\text{var}_2}$$

Furthermore,

$$\text{var}_1 = \text{var} \left(\frac{\widehat{\text{cov}}_N(Y, \hat{X}_{RC})}{\widehat{\text{var}}_N(\hat{X}_{RC})} \right)$$

$$\text{var}_2 = \text{var} \left(\frac{\widehat{\text{cov}}_n(Y, X)}{\widehat{\text{var}}_n(X)} \right)$$

$$\text{var}(\hat{\beta}_{X,MR}) = \text{var} \left(\frac{\widehat{\text{cov}}_N(Y, \hat{X}_{MR})}{\widehat{\text{var}}_N(\hat{X}_{MR})} \right)$$

$$\text{var}(\hat{\beta}_{X,IM}) = \text{var} \left(\frac{\widehat{\text{cov}}_N(Y, \hat{X}_{IM})}{\widehat{\text{var}}_N(\hat{X}_{IM})} \right)$$

Subscript N indicates that the estimate is being made across the full study.

Our task is to evaluate these variances.

1. *Standard RC (RC1)*:

$$\widehat{\text{var}}_1 = \frac{\widehat{\text{cov}}_N(Y, \hat{X}_{RC})}{\widehat{\text{var}}_N(\hat{X}_{RC})} = \frac{\widehat{\text{cov}}_n(X, W)\widehat{\text{cov}}_N(Y, W)}{\widehat{\text{var}}_n(W) \left(\frac{\widehat{\text{cov}}_n(X, W)}{\widehat{\text{var}}_n(W)} \right)^2 \widehat{\text{var}}_N(W)} = \frac{\widehat{\text{var}}_n(W)\widehat{\text{cov}}_N(Y, W)}{\widehat{\text{cov}}_n(X, W)\widehat{\text{var}}_N(W)}$$

Assuming N very large, using the delta method, the approximate variance of this expression is

$$\text{var}_1 = \frac{\beta_X^2 \text{var}(\delta)}{n\gamma_X^2 \text{var}(X)} = \frac{\beta_X^2 (1 - \rho_{XW}^2)}{n\rho_{XW}^2}$$

2. *Calibration study RC (RC2)*: It is simple to show that

$$\text{var}_2 = \frac{\beta_X^2 (1 - \rho_{XY}^2)}{n\rho_{XY}^2}$$

3. *ERC (RC1)*: From the above results it follows that the variance for $\hat{\beta}_{X,RC}$ is given by

$$\frac{\beta_X^2 (1 - \rho_{XW}^2)(1 - \rho_{XY}^2)}{n(\rho_{XW}^2 + \rho_{XY}^2 - 2\rho_{XW}\rho_{XY}^2)}$$

4. *MR*:

$$\begin{aligned} \frac{\widehat{\text{cov}}_N(Y, \hat{X}_{MR})}{\widehat{\text{var}}_N(\hat{X}_{MR})} &= \left\{ \widehat{\text{var}}_N(Y) \left[\hat{\phi} - \psi \sqrt{\frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)}} \right] + \widehat{\text{cov}}_N(Y, W) \sqrt{\frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)}} \right\} \\ &\div \left\{ \widehat{\text{var}}_N(Y) \left[\hat{\phi} - \psi \sqrt{\frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)}} \right]^2 + \widehat{\text{var}}_N(W) \frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)} \right. \\ &\quad \left. + 2\widehat{\text{cov}}_N(Y, W) \sqrt{\frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)}} \left[\hat{\phi} - \psi \sqrt{\frac{\text{var}_n(X|Y)}{\widehat{\text{var}}_N(W|Y)}} \right] \right\} \end{aligned}$$

where $\hat{\phi}$ is the sub-study estimate of the regression coefficient of X on Y , and ψ is the regression coefficient of W on Y .

For large N , using the delta method liberally, the variance of this quantity simplifies to approximately

$$\frac{\beta_X^2 \text{var}(Y|X)}{n} \left[\frac{\text{var}(X)}{\text{cov}(X, Y)^2} - \frac{2 \text{var}(Y|X)}{\text{var}(Y)^2} \right] = \frac{\beta_X^2 (1 - \rho_{XY}^2)(1 - 2\rho_{XY}^2 + 2\rho_{XY}^4)}{n\rho_{XY}^2}$$

5. *IM*:

$$\frac{\widehat{\text{cov}}_N(Y, \hat{X}_{IM})}{\widehat{\text{var}}_N(\hat{X}_{IM})} = \frac{\widehat{\text{cov}}_N(Y, \hat{E}_n(X|Y, W)) + \widehat{\text{cov}}_N(Y, \eta)}{\widehat{\text{var}}_N(\hat{E}_n(X|Y, W) + \eta)} = \frac{\widehat{\text{cov}}_N(Y, YW)\hat{\gamma}}{\hat{\gamma}^T \widehat{\text{var}}_N(YW)\hat{\gamma} + \widehat{\text{var}}_n\eta}$$

where $\hat{\gamma}$ is the sub-study estimate of the regression coefficients of X on (Y, W) .

Table BI. Comparison of theoretical asymptotic standard errors (upper half of the cell) with empirical values (lower half of the cell) for efficient regression calibration (ERC), moment reconstruction (MR) and multiple imputation (IM) estimators.

var(δ)	var(ε)	SE($\hat{\beta}_{X,RC}$)	SE($\hat{\beta}_{X,MR}$)	SE($\hat{\beta}_{X,IM}$)
1	1	0.032	0.032	0.030
		0.033	0.034	0.033
1	0.04	0.0088	0.0086	0.0084
		0.0088	0.0083	0.0087
1	9	0.042	0.121	0.097
		0.061	0.127	0.099
4	1	0.040	0.032	0.031
		0.039	0.032	0.032

Using the delta method liberally, the variance of this expression for large N turns out to be

$$\frac{\beta_X^2 \text{var}(X|Y, W)}{n} \left[\frac{\text{var}(Y)}{\text{cov}(X, Y)^2} - \frac{2 \text{var}(X|Y, W)}{\text{var}(X)^2} \right] = \frac{\beta_X^2 (1 - \rho_{XY|W}^2) (1 - 2\rho_{XY}^2 + 2\rho_{XY}^2 \rho_{XY|W}^2)}{n \rho_{XY}^2}$$

where under the assumption of non-differential measurement error,

$$1 - \rho_{XY|W}^2 = \frac{(1 - \rho_{XY}^2)(1 - \rho_{XW}^2)}{1 - \rho_{XY}^2 \rho_{XW}^2}$$

To verify the accuracy of the variance expressions for ERC, MR and IM, we compared their values with empirical variances obtained from simulations. In Table BI, the theoretical value is given in the upper half of the cell, and the empirical value in the bottom half. The values $\text{var}(X) = 1$, $\beta_0 = 0$, $\beta_X = 1$, $\gamma_0 = 1$, $\gamma_X = 1$, $N = 10000$, $n = 500$ were fixed throughout.

In most cases the approximate formulas agree well with the empirical values. The formula for RC does not appear to do very well when $\text{var}(\varepsilon)$ is large, i.e. 9. However, with a larger N (100 000) the empirical variance for ERC in this case reduces to 0.046, much closer to the theoretical (asymptotic) value of 0.042.

ACKNOWLEDGEMENTS

Carroll's research was supported by a grant from the National Cancer Institute (CA-57030).

REFERENCES

1. 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.
2. Carroll RJ, Stefanski LA. Approximate quasilielihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 1990; **85**:652–663.
3. Gleser LJ. Improvements of the naïve approach to estimation in non-linear errors-in-variables regression models. In *Statistical Analysis of Measurement Error Models and Applications*, Brown PJ, Fuller WA (eds). American Mathematical Society: Providence, RI, 1990.
4. Pierce DA, Kellerer AM. Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika* 2004; **91**:863–876.

5. Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* 2004; **60**:172–181.
6. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*, Chapter 10 (2nd edn). Wiley: Hoboken, NJ, 2002.
7. Brownstone D, Valletta RG. Modeling earnings measurement error: a multiple imputation approach. *The Review of Economics and Statistics* 1996; **78**:705–717.
8. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology* 2006; **35**:1074–1081.
9. 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.
10. Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano R, Bingham S, Schoeller DA, Schatzkin A, Carroll RJ. The structure of dietary measurement error: results of the OPEN biomarker study. *American Journal of Epidemiology* 2003; **158**:14–21.
11. Schatzkin A, Subar AF, Thompson FE, Harlan LC, Tangrea J, Hollenbeck AR, Hurwitz PE, Coyle L, Schussler N, Michaud DS, Freedman LS, Brown CC, Midthune D, Kipnis V. Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health-American Association of Retired Persons Diet and Health Study. *American Journal of Epidemiology* 2001; **154**:1119–1125.
12. Thompson FE, Kipnis V, Midthune D, Freedman LS, Carroll RJ, Subar AF, Brown CC, Butcher MS, Mouw T, Leitzmann M, Schatzkin A. Performance of a Food Frequency Questionnaire in the U.S. National Institutes of Health-AARP Diet and Health Study. *Public Health Nutrition* 2008; **11**:183–195.
13. Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992; **87**:1227–1237.
14. Fuller WA. *Measurement Error Models*. Wiley: New York, 1987.
15. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology* 1992; **136**:1400–1413.