Published in final edited form as:
*Stat Med*. 2008 December 30; 27(30): 6332–6350. doi:10.1002/sim.3458.

# Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment

**Karen Messer**[1] and **Loki Natarajan**[1,*]

1*Division of Biostatistics, Moores UCSD Cancer Center, University of California, La Jolla, CA 92093*

## SUMMARY

In epidemiologic studies of exposure-disease association, often only a surrogate measure of exposure is available for the majority of the sample. A validation sub-study may be conducted to estimate the relation between the surrogate measure and true exposure levels. In this article, we discuss three methods of estimation for such a main study / validation study design: (i) maximum likelihood (ML), (ii) multiple imputation (MI) and (iii) regression calibration (RC). For logistic regression, we show how each method depends on a different numerical approximation to the likelihood, and we adapt standard software to compute both multiple imputation and maximum likelihood estimates. We use simulation to compare the performance of the estimators for both realistic and extreme settings, and for both internal and external validation designs. Our results indicate that with large measurement error or large enough sample sizes, ML performs as well or better than MI and RC. However, for smaller measurement error and small sample sizes, either ML or RC may have the advantage. Interestingly, in most cases the relative advantage of RC versus ML was determined by the relative variance rather than bias of the estimators. Software code for all three methods in SAS is provided.

### Keywords

measurement error; maximum likelihood; multiple imputation; regression calibration

## 1. Introduction

Epidemiologic studies of exposure-disease association often use a noisy or surrogate measure of exposure on a majority of the sample. For instance, in diet studies food frequency questionnaires (FFQ) or food diaries are used to estimate usual intake of foods consumed. Despite being subject to recall biases and measurement error, these self-report dietary assessment methods are used extensively in nutrition studies because they are inexpensive and easy to administer. If ignored, this measurement error can bias the estimated association between exposure and disease [1,2]. In nutritional epidemiology, the associated bias tends to attenuate the estimated risk toward the null so that for example the “true” diet-cancer relative risk may be substantially under-estimated [3,4]. Hence it is important in the analysis of such data to adjust for the error in the measurement of the exposure variables. Often a validation sub-study is conducted to estimate the relation between the noisy surrogate measure and the true exposure levels. Such main study/validation study designs have given rise to a large literature on methods to adjust for measurement error [5,6,7,8,9,10,11,12,13,14], but relatively few epidemiologic studies use these methods in practice [15].

*Correspondence to: Moores UCSD Cancer Center, University of California, La Jolla, CA 92093-0901.

When the outcome of interest is a binary indicator of disease status, logistic regression is often used to model the association between exposure and disease. In this setting, we discuss three approaches to adjusting for measurement error in a main study/validation study design: (i) regression calibration (RC), (ii) multiple imputation (MI) and (iii) maximum likelihood (ML). Regression calibration is the most popular approach in practice as it is simple to apply and usually performs well. Maximum likelihood estimators will in principle have smaller bias and mean squared error (MSE) than regression calibration, yet maximum likelihood is not often applied in epidemiologic studies, perhaps because the method requires special software that is not part of the standard statistical packages. Multiple imputation, which is generally applicable to missing data problems, has not been studied extensively in the measurement error setting. In this article, we compare these three methods analytically and by simulation under a variety of scenarios. Our goal is to characterize situations where the estimates will be numerically close to one another, or where one method may have smaller MSE than the others. We also show how maximum likelihood estimates can be found using standard software, in particular using the NLMIXED procedure in SAS. For some designs, we show how multiple imputation can be conveniently carried out using PROC MI in SAS.

There are few publications using maximum likelihood estimates for measurement error adjustment. Spiegelman et al. [8] compare the maximum likelihood estimator to regression calibration via simulation for a logistic regression model. They use a numerical algorithm to maximize the likelihood and cite a publicly available Fortran program that can be used for computations. Thoreson & Laake [13] also compare maximum likelihood, probit, and regression calibration for logistic regression in a simulation study. They use numerical methods and a 32-point Gaussian quadrature rule to maximize the likelihood equation. Cole et al. [14] compare regression calibration to multiple imputation via simulations in survival analysis using a Cox proportional hazards model. Other methods that deserve mention but that will not be the focus of this article are SIMEX [9] and Bayesian methods [10,11,12]. To our knowledge, maximum likelihood and multiple imputation have not been compared for measurement error adjustment.

## 2. Methods

### Main study / validation study designs

We consider a logistic regression model for a binary disease outcome $Y$, which depends on a vector of exposure variables $X$ and covariates $Z$. The disease model is

$$P\left(Y=1\middle|X,Z\right)=\pi\left(X\beta_x+Z\beta_z\right)\equiv\frac{\exp\left(X\beta_x+Z\beta_z\right)}{1+\exp\left(X\beta_x+Z\beta_z\right)}. \tag{1}$$

The primary objective is to estimate the parameter vector $\beta_x$, which quantifies the risk of disease $Y$ at a given exposure level $X$. In the *main study* we always observe outcome $Y$, but sometimes observe the surrogate measure $W$ instead of exposure $X$. The surrogate $W$ is related to exposure $X$ by a multivariate normal linear regression model which is often called the *measurement error* model,

$$\begin{aligned} W &= X\alpha_x+Z\alpha_z+\epsilon_w \\ \epsilon_w &\sim N(0,\sigma_w^2 I). \end{aligned} \tag{2}$$

where measurement error $\epsilon_w$ is independent of all other variables except $W$, and the parameters $\alpha$ are matrices of regression coefficients.

Because we will sometimes use surrogate $W$ to predict unobserved exposure $X$, a secondary study objective is to estimate the conditional distribution of $X$ given $W$. We assume the exposure vector $X$ is multivariate normal with covariance matrix $\sum_x$, and, for convenience, mean 0. Then, using standard normal theory [16], the conditional distribution of $X$ given $W$ is also multivariate normal so that $X$ can be written

$$\begin{aligned} X &= W\gamma_w + Z\gamma_z + \epsilon_x \\ \epsilon_x &\sim N(0, \textstyle\sum_{x|w}), \end{aligned} \tag{3}$$

with $\epsilon_x$ again independent of all other variables except $X$ and $\gamma$ again a matrix parameter. The parameters $\gamma$ and $\sum_{x|w}$ can be written in terms of the parameters $\alpha$, $\sigma_w^2$ and $\sum_x$, and conversely. Information on these measurement error parameters is obtained from a *validation study*, a usually smaller study in which we measure both exposure $X$ and surrogate $W$, and may or may not also observe outcome $Y$ .

In the measurement error literature, a distinction is often drawn between *internal* and *external* validation study designs. In each, the object is to estimate $\beta_x$, the logistic regression coefficients of $Y$ on $X$. In *internal* validation designs, the validation study is conducted on a random subsample of subjects in the main study. In the validation subsample, complete data $(Y,X,W)$ are measured on each subject. For the main study sample, $(Y,W)$ are measured but $X$ is not observed. In *external* validation designs, the main study and the validation study use independent samples. In the validation sample $(X,W)$ are observed and in the main study $(Y,W)$ are observed. There are no subjects for which $X$ and $Y$ are observed together.

## The likelihood for inference from Y and observed X, conditional on W

We will assume that surrogates $W$ and covariates $Z$ are observed on all subjects. These $Z$'s are assumed to be error-free and will often be omitted from the notation, even though the disease (1) and measurement error (2) models include $Z$.

**The likelihood for (Y,X,W) observed—**Models (1) and (3) imply that outcome $Y$ is conditionally independent of the surrogate $W$ as long as we observe the true exposure $X$, that is, $P(Y|X,W) = P(Y|X)$. This is "non-differential error" in the measurement error literature, and it implies that the joint distribution of $(Y,X)$ given $W$ factors as $P(Y,X|W) = P(Y|X,W)P(X|W) = P(Y|X)\mathrm{P}(X|W)$. Thus the complete-data log-likelihood is given by

$$l(\beta,\gamma \textstyle\sum_{x|w}; Y,X|W) = l_1(\beta; Y|X) + l_2(\gamma, \textstyle\sum_{x|w}; X|W)$$

with $l_1(\beta; Y|X) = Y\,X\,\beta - \log(1 + \exp(X\,\beta))$ a logistic and

$l_2(\gamma, \sum_{x|w}; X|W) = -(1/2)\,(\log\det(\sum_{x|w}) + (X - W\gamma_w)^t \sum_{x|w}^{-1} (X - W\gamma_w))$ a normal regression likelihood.

**The likelihood for (Y,W) observed, X missing—**The marginal distribution for $P(Y|W)$ is given by integrating over the unobserved $X$ in the joint probability distribution $P(Y,X|W)$,

$$P(Y|W;\beta,\gamma,\textstyle\sum_{x|w}) = \int P\left(Y|x;\beta\right) P(x|W;\gamma,\textstyle\sum_{x|w})\,dx. \tag{4}$$

This is of course also the likelihood for the logistic-normal random effects model

$$\begin{aligned} P\left(Y=1\middle|W,\epsilon_x;\beta,\gamma\right) &= \pi\left((W\gamma)\beta+\epsilon_x\beta\right), \\ \epsilon_x &\sim N(0,\Sigma_{x|w}), \end{aligned}$$

obtained by substituting equation (3) into equation (1). After the change of variable $u = \epsilon_x\beta$ the log-likelihood may be written as the univariate integral

$$l_3(\beta,\gamma,\Sigma_{x|w};Y|W)= \log\left((\beta^t\Sigma_{x|w}\beta)^{-1/2}\int\frac{\exp((W\gamma\beta+u)Y)}{1+\exp(W\gamma\beta+u)}\exp\left(-(1/2)\,u^t(\beta^t\Sigma_{x|w}\beta)^{-1}u\right)\mathrm{d}u\right). \quad (5)$$

**The full sample likelihood—**The full sample log-likelihood is the sum over the study subjects of the individual contributions to the likelihood. For an internal design, this is

$$l(\beta,\gamma,\Sigma_{x|w})=\sum_{i\in\text{validation}}\left(l_1\left(\beta;Y_i\middle|X_i\right)+l_2(\gamma,\Sigma_{x|w};X_i,|W_i))+\sum_{i\in\text{main}}l_3(\beta,\gamma,\Sigma_{x|w};Y_i|W_i)\right. \quad (6)$$

For an external design, only $W$ and $X$ are observed for a subject in the validation study, and so, integrating over the missing $Y$, the terms in $l_1$ become zero.

## Maximum Likelihood Estimators

Likelihood (6) is a mixture of exponential family models, and if the model is correct, the maximum likelihood (ML) estimators of the parameters will be consistent, asymptotically normal and have the smallest asymptotic mean squared error among all "regular" estimators. The exact optimality statement involves the usual sequence of local alternatives; to be precise Theorem 8.8 of [17] holds. The necessary differentiability of (5) in particular follows from [17] example 7.7 and the dominated convergence theorem. Thus, under the given model the ML estimator of β will set the "gold standard" for comparison with other estimation methods in the large sample setting. In addition, the model-based standard errors produced by ML will be correct, although we recommend using bootstrap standard errors for all estimation methods, as these will be correct even if the modeling assumptions are incorrect.

The ML estimate will not be computable in closed form, and will be found as the solution to a numerical optimization problem using an algorithm such as Newton-Raphson. Within the maximization algorithm, an added difficulty is that the likelihood function itself is not available in closed form because the integral in term $l_3$ given in (5) must be approximated numerically each time the likelihood is evaluated. Efficient numerical quadrature formulas for (5) have been extensively studied in the random effects literature [18], and by recognizing that this component of the log-likelihood is a normal-logistic random effects model, we may take advantage of readily available specialized maximization routines written for random effects models. In particular, the likelihood for both internal and external validation designs is easily programmed in PROC NLMIXED in SAS v. 9.1. The sample SAS code used in our simulations is given in Appendix II, for a model with two exposure variables and one covariate. We were unable to find a similar module in R which could be adapted for our use. The simulations for most internal designs ran within a few seconds for a sample size of 500. The most time-consuming case took typically about 20 seconds for each sample, and the Newton-Raphson algorithm used between 25 and 30 iterations.

### Regression Calibration Estimators

Regression calibration (RC) is a widely used two-step approach to estimating β in main study / validation study designs [2]. At the first step, the likelihood from the measurement model is maximized using validation study data to obtain the regression estimates $\tilde{\gamma}$, $\tilde{\sum}_{x|w}$. For both the internal and external designs, this first step uses standard linear regression software to maximize

$$\sum_{i\in\text{validation}} l_2(\gamma, \sum_{x|w}; X_i|W_i). \tag{7}$$

At the second step, these estimated parameters are "plugged in" to form an estimate of $E[X|W]$, $\tilde{E}[X|W] = W\tilde{\gamma}_w + Z\tilde{\gamma}_z$. This is then substituted for $X$ into the logistic regression likelihood whenever $X$ is missing for a main study subject. For an internal validation design the second step uses the approximate likelihood

$$l_{RC}(\beta) = \sum_{i\in\text{validation}} l_1\left(\beta; Y_i|X_i\right) + \sum_{i\in\text{main}} l_1\left(\beta; Y_i, |\tilde{E}\left[X|W_i\right]\right), \tag{8}$$

which is maximized using standard logistic regression software to obtain the regression calibration estimate $\tilde{\beta}$. For an external design, only the second term of (8) is available.

Comparing the likelihood equation (6) to (7) and (8), regression calibration corresponds to 1) ignoring any information about the measurement error parameters γ and ∑ possibly contributed by the mixture distribution $l_3$; and 2), approximating $l_3$ by the logistic likelihood $l_1$ evaluated at $\tilde{E}[X|W]$. Regression calibration is computationally simpler than maximum likelihood, however regression calibration estimators may be asymptotically less efficient than maximum likelihood estimators.

At the first RC step, by maximizing $l_2$ separately and ignoring the information in $l_3$, we can obtain $\tilde{\gamma}$, $\tilde{\sum}_{x|w}$ using least squares regression instead of using a more complicated numerical optimization routine. These estimators will be asymptotically efficient among estimators which use only the validation sample [19], however they ignore any information from the main study contributed by $l_3$; which also depends on γ and ∑. Below, we show that $l_3$ actually contributes additional information on γ and ∑ only when $\beta'\sum\beta$ is large, and so it is only in this case that RC could be expected to lose efficiency at the first step. A less complicated issue is that if data are not balanced so that $X_1$ and $X_2$ are missing on different subsamples, then regression calibration should use Generalized Least Squares methods to simultaneously fit a regression model for $(X_1, X_2)$ given $(W_1, W_2)$.

At the second step, RC maximizes likelihood $l_1(Y|\tilde{E}[X|W])$ instead of likelihood $l_3(Y|W)$. Note that, as the quantity $\beta^t\sum_{x|w}\beta \to 0$, the limiting value of $l_3(Y|W)$ is indeed $l_1(Y|E[X|W])$. (To see this, notice the normal density inside the integral in (5) converges to a delta function at $u = 0$ as $\beta^t\sum_{x|w}\beta \to 0$, and then integrate by parts and apply the dominated convergence theorem.) Hence for $\beta^t\sum_{x|w}\beta$ small enough the approximate likelihood (8) is close to the true main study likelihood in (6), and the RC estimates should be close to the ML estimates. However if $\beta^t\sum_{x|w}\beta$ is not small, the RC estimator may solve the wrong equation. By contrast, the ML estimate can be obtained by maximizing $l_3$ directly using specialized numerical integration routines in PROC NLMIXED in SAS [20], which uses standard Gauss-Hermite polynomials and adaptive quadrature algorithms from the numerical analysis literature [21]. These routines adaptively select the number of quadrature points so that the numerical error in the solution of $l_3$ is within a pre-specified tolerance for any value of $\beta^t\sum_{x|w}\beta$.

Further insight into loss of efficiency at the first RC step can be gained by considering the well known approximation [22,23] to the "missing data" likelihood (4) given by

$$P(Y=1|W,Z;\beta,\gamma,\sum_{x|w}) \approx \pi\left(\frac{E\left(X|W\right)\beta+Z\beta_z}{(1+(1.7)^{-2}\beta^t\sum_{x|w}\beta)^{1/2}}\right), \tag{9}$$

where $\pi(\cdot)$ is given in (1). Formal properties of related approximations are studied in [24] and shown to be inefficient as compared to ML estimation using numerical quadrature; nonetheless (9) can aid our intuition. Thus, neglecting $Z$, the log-likelihood for an external design is approximated by

$$l(\beta,\gamma,\sum_{x|w}) \approx \sum_{i\in\text{validation}} l_2(\gamma,\sum_{x|w};X_i|W_i)+\sum_{i\in\text{main}} l_1\left(\gamma g\left(\beta,\sum\right);Y_i|W_i\right) \tag{10}$$

where

$$g\left(\beta,\sum\right)=\beta/(1+(1.7)^{-2}\beta^t\sum_{x|w}\beta)^{1/2}. \tag{11}$$

RC estimates are based on assuming $g(\beta, \sum) = \beta$ in (10), while ML estimates approximately solve the score equations

$$\begin{aligned}\partial l/\partial\sum &= l_2^{(0,1)}(\gamma,\sum)+l_1'(\gamma g(\beta,\sum))\gamma\partial g/\partial\sum=0\\ \partial l/\partial\gamma &= l_2^{(1,0)}(\gamma,\sum)+l_1'(\gamma g(\beta,\sum))g(\beta,\sum)=0\\ \partial l/\partial\beta &= l_1'(\gamma g(\beta,\sum))\gamma\partial g/\partial\beta=0,\end{aligned} \tag{12}$$

Here, $l^{(0,1)}$ denotes the partial derivative with respect to the second argument, $l^{(1)}$ the ordinary derivative and similarly for related notation. We have omitted the summations in (12) for ease of notation.

Considering (7) we see that the regression calibration estimates $\tilde{\gamma}$ and $\tilde{\sum}$ set the terms involving $l_2$ in (12) to 0. Let $\hat{\delta}$ be the solution to the ordinary logistic regression score equations for the regression of $Y$ on $W$ :

$$l_1'(\widehat{\delta})=0. \tag{13}$$

If $\hat{\beta}$ solves

$$g(\widehat{\beta},\tilde{\sum})=\tilde{\gamma}^{-1}\widehat{\delta}, \tag{14}$$

then $\hat{\beta}$ will set the terms $l_1$ in (12) to 0. In this case the score equations are satisfied and the full ML estimates are (approximately) given by $\hat{\beta}$, $\tilde{\gamma}$, $\tilde{\sum}$ In particular, the RC and ML estimates of $\gamma$ and $\sum$ coincide, and there is no loss of efficiency at step 1. However, if $\sum$ and $\gamma^{-1}\delta$ are large,

then equation (14) will not have a solution in β. In this case the score equations do not uncouple, and the ML and RC estimates of γ and ∑ will differ.

As has been observed by [25] in the random-effects model setting, if equation (14) has solution $\hat{\beta}$, then the RC estimate of β is the solution to $\tilde{\beta} = \tilde{\gamma}^{-1}\hat{\delta}$, and, substituting, we have $g(\hat{\beta}, \tilde{\sum}) = \beta$. Because $|g(\beta,\sum)| < |\beta|$ we see that $\tilde{\beta} < \hat{\beta}$, and so β will be an asymptotically biased estimate, with correspondingly smaller variance but greater MSE than the asymptotically optimal $\hat{\beta}$. However, if $\beta^t \sum_{x|w} \beta$ is small, then $|g(\beta, \sum)| \approx |\beta|$ and the ML and RC estimates will be numerically very close to one another.

## Multiple Imputation Estimators

Multiple imputation is now a standard technique for handling data which are missing at random (MAR) [26,27]. The measurement error problem can be put into this framework. For an internal validation design, complete data (*Y*,*X*,*W*) are observed for each subject in the validation subsample, but only (*Y*,*W*) and not *X* are observed for the remaining subjects. If validation substudy subjects are a simple random sample of main study subjects, the missing observations on *X* will be missing completely at random (MCAR). If the validation subsample is stratified on covariates *Z* which are included in models (1) and (3), to the extent that the models are correct, the missing observations will be MAR [26]. In either case, the key point is that the conditional distribution of *X*|*Y*,*W*,*Z* will be the same for a validation study subject, for whom *X* is observed, as for a main study subject, for whom *X* is not observed.

For an external design, the complete data are (*Y*,*X*,*W*,*Z*) as before, however we do not observe complete data for any subject. We observe (*Y*,*W*,*Z*) on main study subjects and (*X*,*W*,*Z*) on validation study subjects. The MAR requirement is now that the conditional distribution *X*|*Y*,*W*,*Z* should be the same for main study and validation study subjects, with a similar requirement for *Y*|*X*,*W*,*Z*. Unlike the internal validation case, we have no direct observations from either conditional distribution. In practice, the assertion that the data are MAR may depend more strongly on *a priori* modeling assumptions for an external design than for an internal design.

"Multiple imputation" describes a class of procedures in which a set of imputed data is substituted for the missing values over *m* independent iterations, at each iteration the parameters are estimated from the "completed" data, and the mean of the *m* resulting parameter estimates yields the final estimator. Here we use a frequentist version of the original Bayesian multiple imputation algorithm [28], in which the imputed data are repeated draws from the distribution of the missing data conditional on the observed data, and at each iteration this conditional distribution is evaluated at randomly drawn parameter values which reflect the uncertainty in the parameter estimates. The Bayesian and frequentist versions differ only in the distribution from which the randomly drawn parameter values are taken. We use the frequentist version because it facilitates comparison with the ML estimator and its variance is smaller than the Bayesian version, substantially so if a large fraction of information is missing [29]. Our discussion follows [27]; for the close connection to the stochastic EM algorithm see also [29] and the references therein. Other specialized multiple imputation procedures exist which are tailored to monotone patterns of missingness or are not model-based, but we do not explore those here.

We next describe the algorithm for the internal design case. Both the imputed data and the parameter values are constructed using the Gibbs sampling version of Markov Chain Monte Carlo (MCMC). The output of the Markov chain at each iteration consists of a randomly drawn parameter vector (the parameter step) and an imputed missing data vector (the imputation step) [20,27]. At iteration $(t-1)$ let $(\beta, \gamma, \sum_{x|w})^{(t-1)}$ be the current parameters and $\tilde{X}(Y_i, W_i)^{(t-1)}$ the associated imputed missing data for the ith subject, which depends explicitly on the

observed values $Y_i$ and $W_i$ for this subject. At the next iteration of the parameter step, the Markov chain computes the parameter estimates $(\hat{\beta}, \hat{\gamma}, \hat{\sum}_{x|w})^{(t)}$ by maximizing the "completed data" likelihood. For an internal validation design this is

$$\sum_{i\in\text{validation}} l_1(\beta;Y_i,X_i) + l_2(\gamma,\sum\nolimits_{x|w};X_i,W_i) + \sum_{i\in\text{main}} l_1(\beta;Y_i,\tilde{X}(Y_i,W_i)^{(t-1)}) + l_2(\gamma,\sum\nolimits_{x|w};\tilde{X}(Y_i,W_i)^{(t-1)},W_i). \quad (15)$$

Recall that $l_1$ and $l_2$ are from the usual logistic and normal regression models, and so the parameter estimates can be obtained using standard software. Next the Markov chain draws an updated parameter vector $(\beta, \gamma, \sum_{x|w})^{(t)}$ from the normal distribution with mean $(\hat{\beta}, \hat{\gamma}, \hat{\sum}_{x|w})^{(t)}$ and variance given by the inverse of the Hessian of (15) evaluated at $(\hat{\beta}, \hat{\gamma}, \hat{\sum}_{x|w})^{(t)}$. (In the Bayesian multiple imputation literature, this distribution is the posterior distribution of the parameters given the data; here we use an uninformative "improper" prior [27,29].)

At the imputation step, for a main study subject the Markov chain imputes $\tilde{X}(Y_i, W_i)^{(t)}$ as a random draw from the conditional distribution of the missing $X$'s given the observed $Y_i$ and $W_i$, evaluated at $(\beta, \gamma, \sum_{x|w})^{(t)}$. This distribution is proportional to

$$f(x) = \frac{\exp(x\beta^{(t)}Y_i)}{1+\exp(x\beta^{(t)})}\exp\left(-(1/2)(x - W_i\gamma^{(t)})'\left(\sum\nolimits_{x|w}^{(t)}\right)^{-1}(x - W_i\gamma^{(t)})\right), \quad (16)$$

which is, however, not readily available using standard software. We discuss how to approximate this distribution below.

MCMC multiple imputation exploits the fact that under regularity conditions, the output of the Markov chain, $(\beta, \gamma, \sum_{x|w}, \tilde{X})^{(t)}$ for $t = b + k, b + 2k \ldots$, will be approximate independent identically distributed (iid) draws from a limiting distribution. Here the Markov chain is iterated $b$ times, where $b$ is a large burn-in period, and the chain is iterated $k$ times between imputations with $k$ large. An EM-type argument shows that this limiting distribution for the $\tilde{X}^{(t)}$ is probability model (16) evaluated at the ML estimates; see [27,29] and the references therein. Similarly, the parameters $(\beta, \gamma, {}_{x|w})^{(t)}$'s are iid draws from a multivariate normal with mean $(\hat{\beta}, \hat{\gamma}, \sum \wedge_{x|w})$, the ML estimates of likelihood (6), and variance the inverse of Hessian matrix of (6) evaluated at $(\hat{\beta}, \hat{\gamma}, \hat{\sum}_{x|w})$. This is of course the asymptotic sampling distribution of the ML estimator, with parameters evaluated at their ML estimates.

To compute the MI estimate, after a sufficient number of iterations (depending on $b$ and $k$), $m$ imputed "completed" data sets are saved, where $m$ is small (often 3 to 5). Then for each imputed dataset, the "completed data" maximum likelihood estimate $\tilde{\beta}^{(i)}$ is computed using (15), which only requires the standard method that would be applied if the data were complete. The MI estimate $\hat{\beta}_{MI}$ is then constructed as the average of these "imputation" estimates,

$$\tilde{\beta}_{MI} = m^{-1}\sum_{i=1}^{m}\tilde{\beta}^{(i)}.$$

For large $n$, this estimator should have the same asymptotic distribution as the ML estimator computed from (6) [29].

The argument for averaging a small number of imputation estimates is from an application of the Rao-Blackwell theorem. The $\beta^{(i)}$'s are iid draws from a multivariate normal distribution with mean $\hat{\beta}$, the ML estimate. Should we average the $\beta^{(i)}$'s, this would be a consistent estimator of $E[\beta^{(i)}] = \hat{\beta}$. Recall, however, that $\beta^{(i)}$ was constructed as a random draw from $\hat{\beta}^{(i)}+\epsilon$, where $\epsilon$ is a mean-zero error term and $\hat{\beta}^{(i)}$ is the completed-data estimate using $\tilde{X}^{(i-1)}$ from the previous imputation. Clearly then, the $\hat{\beta}^{(i)}$'s have smaller imputation variance than the $\beta^{(i)}$'s. Thus we are better off averaging the $\hat{\beta}^{(i)}$'s than the $\beta^{(i)}$'s . Finally, the imputation estimate $\tilde{\beta}^{(i-1)}$ is the completed-data estimate using $\tilde{X}^{(i-1)}$, and so this recovers $\hat{\beta}^{(i)}$. To see why a small number of imputations is sufficient, note the imputation variability of $\hat{\beta}^{(i)}$ may be small, since it is only the fraction of data that is missing which varies from iteration to iteration. The conditioning argument is formalized by writing $E[\beta^{(i)}] = E[E[\beta^{(i)}|\tilde{X}^{(i-1)}]]$, where the Rao-Blackwell theorem assures that $\text{var}(E[\beta^{(i)}|\tilde{X}^{(i-1)}]) < \text{var}(\beta^{(i)})$; see [27] and references therein. For quantification of the imputation error as a function of $m$, see [29].

A computational difficulty implementing this MI algorithm in practice is that there is no readily available MI software which samples from (16) explicitly. However, in some circumstances a mixture of two normals may be an adequate approximation to this distribution. Note that $\beta$ enters to distort the shape of distribution (16) by the factor $\pi(x\beta)$ if $Y = 1$ and the factor $(1-\pi(x\beta))$ if $Y = 0$, and apart from this in each case (16) is a normal density. In an internal validation design, we can fit a separate normal regression of $X$ on $W$ for observations with $Y = 1$ and with $Y = 0$. This provides an empirical approximation to (16) as a mixture of normals, and corresponds to substituting the mixture log-likelihood

$$l_{2MI}\left(\gamma_1,\gamma_0,\sum\nolimits_1,\sum\nolimits_0;W,X,Y\right)=Yl_2\left(\gamma_1,\sum\nolimits_1;X,W\right)+(1-Y)\,l_2\left(\gamma_0,\sum\nolimits_0;X,W\right)$$

for $l_2$ in the completed data likelihood (15) . With this substitution, however, the $\beta^{(i)}$'s at the parameter step are no longer used in the imputation step. Hence we are free to construct an acceptable set of imputed $\tilde{X}^{(t+i)}$ using only the normal likelihood $l_{2MI}$ , computed separately for subjects with $Y = 1$ and with $Y = 0$. The final $m$ imputed data sets are used to form the multiple imputation estimate $\tilde{\beta}_{MI}$ as before.

SAS code using PROC MI to implement this strategy for an internal validation design is given in Figure III, and is evaluated in the simulations. Here, a standard MCMC normal imputation of the missing $X$ data is run separately for the cases $Y = 1$ and $Y = 0$. An uninformative Jeffery's prior is used, which is the default in SAS. The burn in period is $b = 500$ iterations, and $k = 200$ iterations are run between exporting imputed data sets. The number of imputations is $m = 5$. Simulation results below show $\tilde{\beta}_{MI}$ is competitive with the ML estimator for all but extreme parameter values (Table II). However, we have not explored the adequacy of the usual MI estimate of variance, which might be more sensitive to our approximations. For this reason we suggest using bootstrap estimates of variance rather than imputation based estimates. For a detailed discussion of choice of $b$ and $m$ and of the rate of convergence of the variance of the MI estimator to the variance of the ML estimator, see [29]. Future refinements might include sampling directly from (16) using importance sampling as a referee has suggested, or using MI tailored to the monotone normal case in our approximate approach.

For an external validation design, the solution we propose is not satisfactory. The completed data likelihood (15) still applies, except that $\tilde{Y}$ $(X_i, W_i)$ replaces $Y$ for validation study subjects. Thus, the previous approach could be carried out using a Gibbs sampler with $Y$ imputed as well as $X$ at the imputation step. Although in principle this is straightforward, it is not easily implemented inside PROC MI in SAS and we leave this for future work. An alternative na|ve approach which is well known to work poorly for normal data [30] would be to estimate the

parameters for the conditional distribution of $X|W$ from the validation sample, and use this conditional distribution to impute the missing $X$'s in the main study sample. We adopt this method for the external design and refer to it as naïve MI. This method is easily programmed but as expected performs poorly (Table III). A moment's thought reveals why: the missing $X$'s should be imputed from $X|Y^{\frown}, W$; by using $X|W$ extra noise has been introduced which is independent of $Y$. Rather than adjusting for measurement error, further measurement error has been introduced and the na|ve MI estimate in the external case will be biased towards zero. This is borne out by the simulation results.

## 3. Simulations

### Simulation characteristics

We used simulation to compare the performance of maximum likelihood (ML), regression calibration (RC), and multiple imputation (MI) estimators for a logistic regression model constructed with measurement error, using both internal and external validation designs. We simulated two exposure variables $X_1$ and $X_2$; which will have surrogate measures on a subset of the sample, and one covariate, $Z$, always measured without error. We generated 1000 datasets, each of size 500, from a logistic regression disease model as in (1):

$$\begin{aligned} Y|(X_1,X_2,Z) &\sim Bern(\eta), \text{ with logit}(\eta) = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3Z \\ (X_1,X_2,Z) &\sim N(0,\textstyle\sum). \end{aligned}$$

We then generated a corresponding data set of surrogate exposure measurements $W_1$ and $W_2$ from a normal measurement error model as in (2):

$$\begin{aligned} W_1|X_1 &\sim N(\alpha_{01}+\alpha_{11}X_1,\sigma^2_{w|x}) \\ W_2|X_2 &\sim N(\alpha_{02}+\alpha_{12}X_2,\sigma^2_{w|x}). \end{aligned}$$

The size of the validation set was always 250, i.e., 50% of the entire sample. The choice of validation study size was motivated by previous studies [4,8] where validation sample sizes ranging from 173 to 484 were considered. We then deleted $X_1$ and $X_2$ from the 250 observations in the main study sample, for both simulated internal and external designs. In external validation simulations, we also deleted the outcome $Y$ from the 250 observations in the validation study sample.

To facilitate comparison of diffierent simulated scenarios, the predictor variables $X$ and $Z$ were standardized to have mean zero and variance 1, and $X_2$ and $Z$ were constructed to be independent. For simplicity we also fixed $\alpha_{01} = \alpha_{02} = 0$, $\alpha_{11} = \alpha_{12} = 1$, so that $W_1$ and $W_2$ were unbiased surrogates for $X_1$ and $X_2$ respectively. In the disease model, we set $\beta_0 = -1$ yielding a disease rate of 27% at the mean values of the covariates; $\beta_2 = \beta_3 = 0{:}371$, which corresponds to odds ratios of disease of approximately 1.6 between the highest and lowest quartiles of $X_2$ or $Z$, values that are reasonable for many epidemiologic studies [13]. We varied the remaining model parameters: $\beta_1$, the coefficient of the primary exposure variable, the measurement error standard deviation $\sigma_{w|x}$, and $\rho_{12}$ and $\rho_{1z}$ which set the correlation between $X_1$ and $X_2$ and $X_1$ and $Z$ respectively.

The five models we simulated are summarized in Table 1. In the first four, $\beta_1$; the coefficient of the primary exposure variable, is set to 1. This represents a large exposure effect, in which the odds of disease are increased by a factor of 3.85 between the first and third quartiles of exposure as measured by $X_1$, similar to those considered by Thoreson et al. [13]. Models 1 and

2 correspond to large measurement error, with the standard deviation of the measurement error $\alpha_{w|x} = 3$, while Models 3 and 4 have smaller measurement error with $\sigma_{w|x} = 1$. These $\sigma_{w|x}$ values correspond to correlations of 0.3 or 0.7 between surrogates $W$ and true $X$'s. The choice of $\sigma_{w|x}$ values were based on dietary studies, where food frequency questionnaires (FFQ) are often used to estimate usual intake of various food components. FFQs are inexpensive and easy to administer but are known to be subject to large measurement errors [3,4]. In some instances less error-prone measures using biomarkers, such as doubly labeled water for energy intake or urinary nitrogen for protein intake, are used to calibrate the surrogate FFQ in validation substudies. Correlations between FFQ-based dietary intake and "true" intake calibrated by biomarkers have been reported to range between 0.2 and 0.45 [4,31].

In Models 1 and 3 compared to Models 2 and 4, the prediction error for estimating the missing $X_1$ values from the conditional distribution of $X_1$ given $W$ and $Z$ will be relatively small, helped by a strong correlation of $\rho_{1z} = 0{:}70$ between $X_1$ and $Z$; which is always observed without error. The correlation between $X_1$ and $X_2$ is set to 0.2 in Models 1 and 3. In Models 2 and 4 the situation is reversed, with $\rho_{1z} = 0{:}20$ and $\rho_{12} = 0.70$.

Finally, model 5 examined the performance of the methods in an extreme situation of a very large effect ($\beta_1 = 3$), large measurement errors ($\sigma_{w|x} = 3$ ), and high correlation between the two mismeasured covariates ($\rho_{12} = 0.7$). This last situation, although perhaps extreme in epidemiologic situations, permits a comparison of the methods under a worst-case scenario. For models 1–4, the root mean squared error (RMSE) of the ML estimate of $\beta_1$ with complete data (i.e. $X$ observed on the full sample) was 0.18. For model 5 the RMSE of the complete data estimate of $\beta_1$ was 0.33. Thus, while extreme, this case is not entirely unrealistic.

Table 1 also displays $\beta' \sum_{x|w} \beta$ for each of the 5 models considered. If $\beta' \sum_{x|w} \beta$ is small, then the approximation (10) will be valid and ML and RC estimates should be close. Thus this scalar quantity could be used to summarize scenarios when RC will or will not perform well.

Both external and internal validation designs were considered. For each dataset we computed estimates of model parameters by: (i) maximum likelihood (ML); (ii) regression calibration (RC); and (iii) multiple imputation (MI) with $m = 5$ imputed datasets for each of the 1000 simulations. Using the true $X$'s for the entire sample we also computed the complete data estimates, which serve as the "gold standard" for the simulation results.

## Simulation Results

Table I summarizes the performance of regression calibration and multiple imputation relative to the maximum likelihood estimate for the primary exposure variable $\beta_1$. As a measure of performance we used the ratio of the root mean-squared error (RMSE) of the RC and MI estimates to that of the ML estimate. We denote this as relative RMSE. We used a $\delta$-method to obtain approximate standard errors (SEs) of the relative RMSEs. These approximate SEs are displayed as a footnote to Table I. More detailed output from the simulations, i.e., the bias and standard deviation of estimates with standard errors for all three parameters for the internal and external designs, are presented in Table II and Table III.

In the internal design, for all estimators and all models considered, the sampling distributions of estimates were close to normally distributed with no major skewness or outliers. As expected MI and ML estimators had similar MSEs for $\beta_1$ for models 1 through 4 in the internal design. Closer investigation confirmed that these estimates were numerically very close to one another, indicating that the approximation to the conditional distribution we used in the MI was good enough to yield the maximum likelihood estimates. In fact, ML and MI had little bias and similar SDs for Models 1 – 4 for all parameters (Table II). In the internal validation design, it was not until the extreme case of model 5 that the more tailored numerical approximation to

the likelihood in the ML code gave this method a 15% advantage over our implementation of MI. In Model 5 for the internal design (Table II) the downward bias in MI for estimating the true $\beta_1 = 3$ was 8% (100*0.241/3) compared to 3% for ML (100*0.096/3), leading to a 15% increase in MSE for MI compared to ML (Table I). The relative efficiency (on the variance scale) of MI from using a finite number of imputations, *m*, versus using infinitely many imputations, is approximately $(1+\frac{\lambda}{m})^{-1}$ where λ is the fraction of the missing information [28]. In our internal validation set-up $\lambda = 0.5$ and $m = 5$ hence the standard errors of parameter estimates using MI would be inated by about 5%.

RC and ML performed equally well for models 1 and 2 for the internal validation design. There was an approximate 10% increase in MSE for RC compared to ML for models 3 and 4. RC bias was small and comparable to ML bias for all parameters for Models 1 – 4 (Table II). The SD estimates using ML were consistently lower than RC for all parameters, leading to the modest albeit significant 10% efficiency gains (see footnote Table I) when estimating $\beta_1$ for Models 3 and 4. For Model 5, RC exhibited approximately 4% (100*0.122/3) bias when estimating the true $\beta_1 = 3$ compared to 3% bias for ML, and was much less efficient (RC SD=0.499 vs. 0.388 for ML). Thus, the ML estimator was considerably superior to the RC estimator for the extreme case of model 5, with a 28% gain in RMSE. In particular, using the approximate SE estimates for the relative RMSEs (see footnote to Table I), we conclude that for the internal design, ML significantly outperforms MI for Model 5 and RC for Models 3–5.

For external designs, the results were more extreme. In some cases there was substantial skewness and large outliers. Comparing ML performance in Models 1 – 4, parameter estimates were most biased for Models 1 and 2 compared to 3 and 4 (Table III). RC shows a similar pattern. Models 1 and 2 correspond to large measurement error situations, hence the poorer performance of ML and RC for these models is not surprising. Focusing on $\beta_1$, the primary parameter of interest, the two methods also had similar SDs for Model 1, leading to comparable MSEs (Table I). In particular, the mean squared error of RC when estimating $\beta_1$ was comparable to that of ML for the large measurement error case (Models 1 and 2) if $X_1$ was highly correlated with *Z*; (model 1) so that the prediction error of *X* given *W* and *Z* was relatively small, but not otherwise (model 2).

RC is known to perform well when $\beta' \sum_{x|w} \beta$ is small [2] in the external design. In our simulations, Models 3, and 4 had the lowest values for $\beta' \sum_{x|w} \beta$ (Table I) and the RC estimator worked best in these models. In these small measurement error cases (Models 3 and 4), the RC estimator performed better than ML, which was unexpected. For models 3 and 4 in the external design, RC had very little bias for all the parameters, whereas ML displayed moderate bias of 0.12–0.13 for $\beta_1$ (Table III), yielding estimates that were on average 12–13% higher than the true $\beta_1$ value of 1. ML was practically unbiased for the other β's. Furthermore, for Models 3 and 4, the SD of ML estimates of $\beta_1$ were larger than the corresponding estimates by RC leading to the apparent super-efficiency of RC (Table I). On closer inspection, the ML estimates were more skewed than the RC estimates, which degraded the performance of the ML estimator for Models 3 and 4. When we trimmed the extreme 10% of the ML estimates of $\beta_1$, the ratio of RMSE of RC to ML was 0.93 for Model 3, and 0.98 for Model 4 indicating that the two methods perform similarly after removing ML outliers. Further, with larger samples the asymptotic efficiency and normality of the ML estimator became evident. In particular, when we reran simulations for Model 3 in the external design with study sample-sizes of 1000 and validation sample-size of 500, ML outperformed RC with relative RMSE of RC to ML equal to 1.17.

For Models 2 and 5 in the external design, RC estimates were unstable as evidenced by the large variability in the estimates. In model 5, the RC estimates of β1 had a standard deviation

(SD) of 7.28 (Table III) and ranged from −40:29 to 171:12 compared to an SD of 1:32 for ML with a range of −1:57 to 5:80. Similarly the RC ranges for $\beta_2$ were −132:80 to 49:82 (compared to −3:82 to 4:07 for ML) and −31:49 to 20:36 (compared to −0:72 to 2:04 for ML) for $\beta_3$. The 5th and 95th percentiles of the RC estimate of $\beta_1$ were similar to the extreme values of the MLE, suggesting that RC estimates have extreme values in about 10% of cases. The poor performance of RC is attributable to the fact that models 2 and 5 are both scenarios with substantial measurement error and/or large $\beta_1$, so that $\beta^t \sum_{x|w} \beta$ is large (Table I). Hence the RC approximation, which amounts to using (10) followed by substituting β for $g(\beta)$; is not accurate. On the other hand, our implementation of ML maximizes the full likelihood (6) (with $l_1 = 0$ for the external design) using a numerical approximation which adapts to large values of $\beta^t \sum_{x|w} \beta$, and works even when approximation (9) does not hold. In this case, joint maximization of the likelihood equations prevents the extreme estimates of β which are sometimes seen in RC. Thus ML is recommended over RC in external design studies with large measurement error and strong anticipated associations between the true $X$ and disease $Y$ .

It is noteworthy that the poor performance of RC for Models 2 and 5 in the external design (Table III), is mitigated in the internal design (Table II), where the RC estimate for β is obtained by maximizing equation (8). The first term in this likelihood is the usual logistic regression likelihood for β given $X$ and $Y$ from the the validation study, in which $X$ and $Y$ are observed on half the sample. This term controls the estimate of β and reduces the variability that was observed in the external validation design especially for models 2 and 5. In fact, for model 2, the range of the RC estimate of $\beta_1$ was 0.29 to 2.09, comparable to the corresponding ML range of 0.27 to 1.96.

For external designs, the na|ve implementation of MI performed very poorly with downwards bias of 50% or more (Table III) especially when estimating $\beta_1$ (Table III). Hence naïve MI results are not summarized in Table I. In simulations 1 – 4 in which the true value of $\beta_1$ equals 1, the naïve MI estimate was between 0:05 (model 1) and 0:43 (model 4). The poor performance of naïve MI is not surprising since for the external validation design information on the outcome $Y$ is not available in the validation study. Hence the imputation step should sample missing $X$ *and* $Y$ from the distribution (16), which is not readily approximated in the standard MI package in SAS. Our na|ve implementation instead sampled the missing $X$'s from the conditional distribution of $X$ given the observed $W$ and $Z$ (ignoring $Y$ completely in the imputation step; see Figure III). The resulting MI estimates were severely biased (Table III) towards zero. A heuristic explanation might be that, because the imputed $X\hat{}$ ignores $Y$ , the imputation error $X - X\hat{}$ is likely to be large, particularly when $X$ and $Y$ are strongly associated, and this will bias the resulting MI estimate towards zero. Of note, the large biases in MI in the external design are easily corrected in the internal design in which $Y$ and $X$ are observed in the validation sample. Thus, in the internal design, the imputation step samples from the approximate conditional distribution of $X$ given $Y$ , $W$, and $Z$, and the imputed $X\hat{}$ is a reasonably good approximation of $X$, leading to minimal bias as observed in Table II. Thus the standard implementation of MI in SAS (or other software) should not be used for measurement error adjustment for an external validation design without further adaptation.

## 4. Conclusions

In this paper, we have compared maximum likelihood (ML), multiple imputation (MI) and regression calibration (RC) estimation methods in the setting of logistic regression when the primary covariates of interest are measured with error. We have shown how each method of estimation depends on a different numerical approximation to the likelihood, and that this is the major difference between them. We have adapted standard software to perform both maximum likelihood and multiple imputation estimation. We have compared the performance of these methods in both realistic and extreme settings using simulation, for internal and

external validation designs. In summary, all methods work well in the internal validation design, although ML has an expected small advantage in efficiency compared to RC. The MI and ML estimates are approximately equal in the internal design except for an extreme scenario with strong effects and large measurement error. For the external validation design, MI should not be used with the approach studied here because it does not properly impute the missing outcome data. RC can be used successfully for the external design unless the errors are large and the association between the exposure and disease is strong. Maximum likelihood works well under all circumstances, provided main study and validation sample-sizes are large enough.

A limitation of our simulation results is that we took 50% of the sample to be in the validation substudy, which may be a larger proportion than in many epidemiologic studies. However, the validation substudy sample size of 250 may not be unrealistic [4,8]. A further limitation is that it may be difficult to know whether ML or RC would be more efficient in practice in the external validation design. When $\beta' \sum_{x|w} \beta$ is larger than 1 (Models 2 and 5, Table I) ML appears to be as good or better than RC. However, for small $\beta' \sum_{x|w} \beta$ (Models 1, 3, and 4, Table I) either estimator may have the advantage for the sample sizes we used in the simulations. Interestingly, in all cases the relative advantage was determined by the relative variance of the estimators rather than bias. In those cases where ML was less efficient than RC (Models 3 and 4, external design) , the sampling distribution of the ML estimator was far from normal, indicating that large sample asymptotics did not yet obtain. Our simulations with larger sample sizes then showed the expected advantage for ML over RC. Finally it is important to note that our naïve implementation of MI for the external design would not be expected to work well. A MI approach that properly imputes both $Y$ and $X$ in the external design would likely perform well, but we do not investigate this here.

In our simulations the model used for estimation has always been the actual model used to generate the data. This is, of course, never the case in practice, where a statistical model is hoped to be a useful summary of the data but is never assumed to be an exact representation of reality. Thus an important question for future work, as pointed out by a reviewer, is how these methods would perform under model misspecification. Here, ML estimators again enjoy a theoretical advantage, in that under ML the estimated model will be closest to the data generating model in Kullback-Leibler distance, and it would be of interest to see whether this is a practical advantage. To the extent that MI is a stochastic form of the EM algorithm, it might share in this advantage, although not necessarily in the approximate form presented here. Questions of relative robustness to contamination by inuential observations or outliers may be equally important. We leave these questions to future work.

## Acknowledgements

Contract/grant sponsor: National Institutes of Health; contract/grant number: 5 R03 CA117292-02

Contract/grant sponsor: Tobacco Related Diseases Research Program, State of California; contract/grant number: 15RT023

## APPENDIX

## II. SAS Code: ML estimator

```
proc nlmixed data=<dataset>   itdetails;/*declare and initialize
parameters*/parms  beta0  -1   betaX1  3   betaX2    0.37  betaZ
0.37           gammaX10 0.0 gammaX11 0.09 gammaX12  0.06 gammaX1z
0.18           gammaX20 0.0 gammaX21 0.06 gammaX22  0.09 gammax2z
−0.12         sigmasq11 0.82 sigmasq22 0.85 sigmasq12 0.57;         /
```

```
* sigmasq11, sigmasq22, and sigmasq12 are cov(x1,x2|w1,w2,z) */X1hat
= gammaX10 + gammaX11*W1 + gammaX12*W2 + gammaX1z*Z;X2hat = gammaX20
+ gammaX21*W1 + gammax22*W2 + gammaX2z*Z;/*validation study sample
log-likelihood - W, X and possibly Y;*/if validation_sample then
do;           DET   = sigmasq11*sigmasq22 -sigmasq12**2;
TERM1 = -LOG (DET);           TERM2 = -( sigmasq22*(X1-X1hat)**2 +
sigmasq11*(X2-X2hat)**2                          -2 *sigmasq12*(X1-
X1hat)*(X2-X2hat) )/(DET) ;          LL = (TERM1 + TERM2)/2;
if Y ne . then do;            ETA     = beta0 + betaX1+ betaX2*X2 +
betaX*z  ;            LLBIN = y*ETA -log (1-exp(ETA));
LL    = LL + LLBIN;            end;           end;/*main study samply
log-likelihood - W and Y;*/else do;           ETA      = beta0 +
betaX1*X1hat+ betaX2*X2hat +betaZ*z + U ;           LLBIN   = Y*ETA -
log(1=exp(ETA)) ;           LL      = LLBIN;           end;/*specify
model and normal random effects parameters*/model  Y ~ general
(LL) ;random U ~ normal (0, (betaX1**2)*sigmasq11 + (betaX1*betaX2)
*sigmasq12                           + (betaX2**2)*sigmasq22)
subject=caseid;run;
```

## III. SAS Code: Multiple Imputation estimator, internal design

```
title1 "Multiple imputation: INTERNAL validation  ";/*COMPUTE NORMAL
IMPUTATION OF x1 X2 CONDITIONAL ON VALUE OF Y */;proc mi
data=<dataset>  seed=<seed> out=outmi nimpute = 5;  by y;  *data must
be sorted;  var w1 w2 z x1 x2 ;  MCMC nbiter = 500 NITER=200;  run;/
*RUN LOGISTIC REGRESSION BY IMPUTATION NUMBER*/;proc sort data=
outmi;  by _imputation_;  run;ods output ParameterEstimates =
betaImputed;proc logistic data=outmi desc;  by _imputation_;  model
y = x1 x2 z;    run;ods output close;/* GET MI ESTIMATE, PRINT AND
STORE IN BETAMI */;ods output summary=betaMI;proc means
data=betaImputed mean;  class variable;  var Estimate;  run;ods
output close;proc print data = betaMI;  run;
```

## REFERENCES


1. Fuller WA. Measurement Error Models. Wiley Series in Probability and Mathematical Statistics. 1987
2. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Monographs on Statistics and Applied Probability. 63. Chapman & Hall/CRC; 1995. Measurement Error in Nonlinear Models.
3. Prentice RL. Measurement error and results from analytic epidemiology Dietary fat and breast cancer. J. Natl. Cancer Inst 1996;88:1738–1747. [PubMed: 8944004]
4. Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano RP, Bingham S, Schoeller DA, Schatzkin A, Carroll R. Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study. American Journal of Epidemiology 2003;158:14–21. [PubMed: 12835281]
5. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med 1989;9:1051–1069. [PubMed: 2799131]
6. Spiegelman D, Schneeweiss S, McDermott A. Measurement Error Correction for Logistic Regression Models with an Alloyed Gold Standard. American Journal of Epidemiology 1997;145:184–196. [PubMed: 9006315]
7. 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. [PubMed: 11135353]

8. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error, in main study/ validation designs. J. Amer. Stat. Assoc 2000;95:51–61.
9. Stefanski LA, Cook J. Simulation Extrapolation: the measurement error jackknife. J. Amer. Stat. Assoc 1995;90:1247-156.
10. Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. Statistics in Medicine 1993;12:1703–1722. [PubMed: 8248663]
11. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. American Journal of Epidemiology 1993;15:430–442. [PubMed: 8213748]
12. Gustafson, P. Measurement Error and Misclassification in Statistics and Epidemiology. Chapman & Hall; 2003.
13. Thoresen M, Laake P. A simulation study of measurement error correction methods in logistic regression. Biometrics 2000;56:868–872. [PubMed: 10985228]
14. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol 2006;35(4):1074–1081. [PubMed: 16709616]
15. Jurek AM, Maldonado G, Greenland S, Church TR. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. Eur J Epidemiol 2006;21(12):871–876. [PubMed: 17186399]
16. Anderson, TW. Wiley Series in Probability and Statistics. Wiley-Interscience; 2003. An Introduction to Multivariate Statistical Analysis.
17. van der Vaart, AW. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press; 2000. Asymptotic Statistics.
18. Pinheiro, JC.; Bates, DM. Statistics and Computing. Springer; 2002. Mixed Effects Models in S and S-Plus.
19. Bickel PJ, Doksum KA. Mathematical Statistics. Prentice Hall. 1977
20. Statistical Analysis Software v. 9.1., NLMIXED documentation. Carey, North Carolina, USA: SAS Institute Inc.; 2003.
21. Golub GH, Welsh JH. Calculation of Gauss Quadrature Rules. Math. Comput 1969;23:221–230.
22. Monahan, J.; Stefanski, LA. Handbook of the Logistic Distribution. New York: Marcel-Dekker; 1992. Normal scale mixture approximations to *F***z* and computation of the logistic-normal integral.
23. Reeves GK, Cox DR, Darby SC, Whitley E. Some aspects of measurement error in explanatory variables for continuous and binary regression models. Stat Med 1998;17(19):2157–2177. [PubMed: 9802176]
24. Pinheiro J, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. J. Computational Stat. and Graphics 1995;4:12–35.
25. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalised estimating equation approach. Biometrics 1988;44:1049–1060. [PubMed: 3233245]
26. Little RA, Rubin D. Statistical Analysis with Missing Data. Wiley-Interscience. 2002
27. Schafer, JL. Monographs on Statistics & Applied Probability. 72. Chapman & Hall/CRC; 1997. Analysis of Incomplete Multivariate Data.
28. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: J Wiley and Sons; 1987.
29. Wang N, Robins JM. Large sample theory for parametric multiple imputation procedures. Biometrika 1998;85(4):935–948.
30. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J. of Clin. Epidemiology 2006;59:1092–1101.
31. Natarajan L, Flatt SW, Sun X, Gamst AC, Major JM, Rock CL, Al-Delaimy W, Thomson CA, Newman VA, Pierce JP. Womens Healthy Eating and Living Study Group. Validity and systematic error in measuring carotenoid consumption with dietary self-report instruments. Am J Epidemiol 2006;163(8):770–778. [PubMed: 16524958]

**Table I**

Simulation Characteristics and Relative Efficiency of Multiple Imputation (MI) and Regression Calibration (RC) compared to Maximum Likelihood (ML)[a]

| Simulation Parameters | | | | | | Relative Root Mean-Squared Error,[b] $\beta_1$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Internal Design | | External Design |
| **Model** | $\beta_1$ | $\sigma_{w\|x}$ | $\rho_{12}$ | $\rho_{1z}$ | $\beta'\Sigma_{x\|w}\beta$ | **MI**[c] | **RC**[c] | **RC**[c] |
| 1 | 1 | 3 | 0.2 | 0.7 | 0.73 | 0.99 | 1.01 | 1.03 |
| 2 | 1 | 3 | 0.7 | 0.2 | 1.37 | 0.99 | 1.01 | 1.48 |
| 3 | 1 | 1 | 0.2 | 0.7 | 0.45 | 1.02 | 1.10 | 0.72 |
| 4 | 1 | 1 | 0.7 | 0.2 | 0.63 | 1.00 | 1.11 | 0.71 |
| 5 | 3 | 3 | 0.7 | 0.2 | 8.84 | 1.15 | 1.28 | 3.89 |

[a] In all simulations, $\beta_0 = -1$, $\beta_2 = \beta_3 = 0{:}371$

[b] Relative RMSE is ratio of RMSE of RC (or MI) to RMSE of ML; RMSE = Root Mean-Squared Error

[c] SE estimates for the ratio of RMSE were approximated using a δ-method. The covariance between ML, MI, and RC estimates was ignored in the SE estimation which would likely introduce a conservative bias (i.e. larger estimated SEs) in our calculations, since the methods would be expected to yield positively correlated MSEs. In the internal design, the SE of the ratio of RMSE of MI to ML was approximately 0.03 in all 5 models; the SE of the ratio of RMSE of RC to ML was approximately 0.04 for models 1 – 4 (for model 5 it was 0.05). For the external design the SE of ratio of RMSE of RC to ML were approximately 0.06, 0.19,0.03,0.04, and 1.10 for Models 1 – 5 respectively.

**Table II**

Internal Design Simulation[a] Results: Bias (SE)[b], Standard Deviation (SE)[b] of Estimates

| Model | Coefficient | | Method | | | |
|---|---|---|---|---|---|---|
| | | | Complete[c] | ML[c] | RC[c] | MI[c] |
| 1 | $\beta_1$ | Bias | 0.012 (0.006) | 0.037 (0.008) | 0.03 (0.008) | 0.001(0.008) |
| 1 | $\beta_1$ | SD | 0.18 (0.003) | 0.261 (0.004) | 0.265 (0.004) | 0.262 (0.004) |
| 1 | $\beta_2$ | Bias | 0.012 (0.004) | 0.021 (0.005) | 0.022 (0.006) | 0.007 (0.005) |
| 1 | $\beta_2$ | SD | 0.119 (0.002) | 0.169 (0.003) | 0.175 (0.003) | 0.167 (0.003) |
| 1 | $\beta_3$ | Bias | 0.007 (0.005) | 0 (0.006) | 0.018 (0.008) | 0.016 (0.006) |
| 1 | $\beta_3$ | SD | 0.166 (0.003) | 0.199 (0.003) | 0.242 (0.004) | 0.2 (0.003) |
| 2 | $\beta_1$ | Bias | 0.016 (0.006) | 0.03 (0.008) | 0.027 (0.008) | −0.002 (0.008) |
| 2 | $\beta_1$ | SD | 0.179 (0.003) | 0.248 (0.004) | 0.252 (0.004) | 0.248 (0.004) |
| 2 | $\beta_2$ | Bias | 0.008 (0.005) | 0.016 (0.007) | 0.016 (0.007) | 0.003 (0.007) |
| 2 | $\beta_2$ | SD | 0.158 (0.002) | 0.231 (0.004) | 0.234 (0.004) | 0.233 (0.004) |
| 2 | $\beta_3$ | Bias | 0.007 (0.004) | 0.01 (0.004) | 0.015 (0.006) | 0.008 (0.004) |
| 2 | $\beta_3$ | SD | 0.118 (0.002) | 0.134 (0.002) | 0.178 (0.003) | 0.133 (0.002) |
| 3 | $\beta_1$ | Bias | 0.009(0.006 | 0.022(0.008) | 0.023(0.008) | 0(0.008) |
| 3 | $\beta_1$ | SD | 0.181(0.003) | 0.239(0.004) | 0.262(0.004) | 0.245(0.004) |
| 3 | $\beta_2$ | Bias | 0.009(0.004) | 0.016(0.005) | 0.018(0.006) | 0.009(0.005) |
| 3 | $\beta_2$ | SD | 0.118(0.002) | 0.145(0.002) | 0.176(0.003) | 0.151(0.002) |
| 3 | $\beta_3$ | Bias | 0.014(0.005) | 0.009(0.006) | 0.022(0.007) | 0.019(0.006) |
| 3 | $\beta_3$ | SD | 0.16(0.003) | 0.183(0.003) | 0.234(0.004) | 0.188(0.003) |
| 4 | $\beta_1$ | Bias | 0.032(0.006) | 0.038(0.007) | 0.039(0.008) | 0.015(0.008) |
| 4 | $\beta_1$ | SD | 0.184(0.003) | 0.236(0.004) | 0.262(0.004) | 0.239(0.004) |
| 4 | $\beta_2$ | Bias | −0.003(0.005) | 0(0.007) | 0.011(0.008) | −0.002(0.007) |
| 4 | $\beta_2$ | SD | 0.168(0.003) | 0.216(0.003) | 0.239(0.004) | 0.225(0.004) |
| 4 | $\beta_3$ | Bias | 0.001(0.004) | 0.001(0.004) | 0.008(0.006) | 0.002(0.004) |
| 4 | $\beta_3$ | SD | 0.122(0.002) | 0.131(0.002) | 0.177(0.003) | 0.131(0.002) |
| 5 | $\beta_1$ | Bias | 0.055 (0.01) | −0.096 (0.012) | 0.122(0.016) | −0.241(0.012) |
| 5 | $\beta_1$ | SD | 0.331(0.005) | 0.388(0.006) | 0.499(0.008) | 0.393(0.006) |
| 5 | $\beta_2$ | Bias | 0.016(0.007) | 0.038(0.01) | 0.022(0.01) | −0.023(0.009) |
| 5 | $\beta_2$ | SD | 0.209(0.003) | 0.309(0.005) | 0.303(0.005) | 0.276(0.004) |
| 5 | $\beta_3$ | Bias | 0.008(0.005) | 0.001(0.006) | 0.007(0.007) | −0.027(0.006) |
| 5 | $\beta_3$ | SD | 0.153(0.002) | 0.186(0.003) | 0.221(0.003) | 0.18(0.003) |

[a] 1000 datasets of size $n = 500$ each were generated

[b] SE: standard error of Bias and Standard Deviation (SD)

[c] Complete: Complete Data with $X$ and $Y$ observed on all subjects; ML: Maximum likelihood; RC: Regressio calibration; MI: Multiple imputation

**Table III**

External Design Simulation[a] Results: Bias(SE)[b], Standard Deviation (SE)[b] of Estimates

| Model | Coefficient | | Method | | | |
|---|---|---|---|---|---|---|
| | | | Complete[c] | ML[c] | RC[c] | Naïve MI[c] |
| 1 | $\beta_1$ | Bias | 0.013(0.006) | 0.24(0.04) | 0.08(0.05) | −0.948(0.003) |
| 1 | $\beta_1$ | SD | 0.184(0.003) | 1.37(0.02) | 1.43(0.02) | 0.085(0.001) |
| 1 | $\beta_2$ | Bias | 0.012(0.004) | 0.09(0.03) | −0.06(0.02) | −0.321(0.002) |
| 1 | $\beta_2$ | SD | 0.119(0.002) | 0.91(0.01) | 0.69(0.01) | 0.07(0.001) |
| 1 | $\beta_3$ | Bias | 0.006(0.005) | 0.08(0.03) | −0.16(0.03) | 0.552(0.006) |
| 1 | $\beta_3$ | SD | 0.162(0.003) | 0.82(0.01) | 1(0.02) | 0.181(0.003) |
| 2 | $\beta_1$ | Bias | 0.006(0.006) | 0.15(0.05) | −0.16(0.07) | −0.91(0.003) |
| 2 | $\beta_1$ | SD | 0.181(0.003) | 1.51(0.02) | 2.25(0.04) | 0.087(0.001) |
| 2 | $\beta_2$ | Bias | 0.008(0.005) | 0.15(0.05) | −0.07(0.07) | −0.296(0.003) |
| 2 | $\beta_2$ | SD | 0.164(0.003) | 1.56(0.02) | 2.16(0.03) | 0.085(0.001) |
| 2 | $\beta_3$ | Bias | 0.013(0.004) | 0.12(0.01) | −0.07(0.02) | 0.072(0.005) |
| 2 | $\beta_3$ | SD | 0.123(0.002) | 0.37(0.01) | 0.52(0.01) | 0.153(0.002) |
| 3 | $\beta_1$ | Bias | 0.015(0.006) | 0.13(0.02) | −0.02(0.01) | −0.667(0.005) |
| 3 | $\beta_1$ | SD | 0.187(0.003) | 0.6(0.01) | 0.45(0.01) | 0.145(0.002) |
| 3 | $\beta_2$ | Bias | 0.008(0.004) | 0.01(0.01) | −0.03(0.01) | −0.146(0.004) |
| 3 | $\beta_2$ | SD | 0.121(0.002) | 0.3(0.005) | 0.26(0.004) | 0.117(0.002) |
| 3 | $\beta_3$ | Bias | 0.007(0.005) | −0.02(0.01) | −0.04(0.01) | 0.381(0.006) |
| 3 | $\beta_3$ | SD | 0.164(0.003) | 0.37(0.01) | 0.34(0.01) | 0.193(0.003) |
| 4 | $\beta_1$ | Bias | 0.012(0.006) | 0.12(0.02) | −0.06(0.02) | −0.571(0.004) |
| 4 | $\beta_1$ | SD | 0.185(0.003) | 0.69(0.01) | 0.5(0.01) | 0.14(0.002) |
| 4 | $\beta_2$ | Bias | 0.004(0.005) | 0(0.02) | −0.04(0.02) | −0.058(0.004) |
| 4 | $\beta_2$ | SD | 0.162(0.003) | 0.58(0.01) | 0.48(0.01) | 0.135(0.002) |
| 4 | $\beta_3$ | Bias | 0.008(0.004) | 0.01(0.01) | −0.04(0.01) | 0.035(0.005) |
| 4 | $\beta_3$ | SD | 0.12(0.002) | 0.22(0.004) | 0.19(0.003) | 0.157(0.002) |
| 5 | $\beta_1$ | Bias | 0.055(0.01) | −1.36(0.04) | −1.22(0.23) | −2.861(0.003) |
| 5 | $\beta_1$ | SD | 0.33(0.005) | 1.32(0.02) | 7.28(0.12) | 0.098(0.002) |
| 5 | $\beta_2$ | Bias | −0.002(0.006) | 0.68(0.05) | −0.4(0.19) | −0.268(0.003) |
| 5 | $\beta_2$ | SD | 0.204(0.003) | 1.65(0.03) | 6.11(0.1) | 0.094(0.001) |
| 5 | $\beta_3$ | Bias | 0.008(0.005) | 0.18(0.01) | −0.25(0.05) | 0.045(0.005) |
| 5 | $\beta_3$ | SD | 0.152(0.002) | 0.43(0.01) | 1.53(0.02) | 0.147(0.002) |

[a] 1000 datasets of size $n = 500$ each were generated

[b] SE: standard error of Bias and Standard Deviation (SD)

[c] Complete: Complete Data with $X$ and $Y$ observed on all subjects; ML: Maximum likelihood; RC: Regression calibration; MI: Multiple imputation