Published in final edited form as:

Stat Med. 2014 February 20; 33(4): 675–692. doi:10.1002/sim.5966.

Expected Estimating Equation using Calibration Data for Generalized Linear Models with a Mixture of Berkson and Classical Errors in Covariates

Jean de Dieu Tapsoba^{1,*†}, Shen-Ming Lee², and Ching-Yun Wang¹

¹Division of Public Health, Fred Hutchinson Cancer Research Center, P. O. Box 19024, Seattle, Washington 98109, U.S.A.

²Department of Statistics, Feng Chia University, Taichung, Taiwan 40724, R.O.C.

Abstract

Data collected in many epidemiological or clinical research studies are often contaminated with measurement errors that may be of classical or Berkson error type. The measurement error may also be a combination of both classical and Berkson errors and failure to account for both errors could lead to unreliable inference in many situations. We consider regression analysis in generalized linear models when some covariates are prone to a mixture of Berkson and classical errors and calibration data are available only for some subjects in a subsample. We propose an expected estimating equation approach to accommodate both errors in generalized linear regression analyses. The proposed method can consistently estimate the classical and Berkson error variances based on the available data, without knowing the mixture percentage. Its finite-sample performance is investigated numerically. Our method is illustrated by an application to real data from an HIV vaccine study.

Keywords

Berkson error; calibration subsample; classical error; expected estimating equation; generalized linear model; instrumental variable

1 Introduction

Measurement error is a recurring issue in medical and epidemiological studies aiming to characterize the relationship between a response variable and a vector of covariates. The problem arises due to the fact that values of some covariates cannot be accurately obtained and are measured with errors instead. Some examples include the CD4 cell count in the AIDS clinical trial ACTG 175 [1] and dosimetry data [2]. Two different types of measurement errors exist in the literature, namely the classical and Berkson measurement errors [3]. A major difference between the two error types is that a classical-type error is independent of the unobservable covariates, while a Berkson-type error is positively correlated with the unobservable covariates. The CD4 cell counts are well-known to be prone to classical measurement error. Radiation dose measures on the other hand are often believed to be subject to a mixture of both Berkson and classical errors. For example, DS02 radiation dose estimates for atomic-bomb survivors, who were followed up by the Radiation

*Correspondence to: Jean de Dieu Tapsoba, 1100 Fairview Ave N., P. O. Box 19024, M2-B500, Fred Hutchinson Cancer Research Center, Seattle, WA. 98109-1024, U.S.A. .

†jtapsoba@fhcrc.org

Effects Research Foundation (RERF) are contaminated with both errors due to uncertainties related to averaging among survivors who generally shared the same location and survivors' individual recollections of location and shielding [4]. If ignored, measurement error could lead to loss of power for hypothesis testing and biased estimation in many situations [5, 6].

Substantial effort has been devoted in the literature to developing methods to accommodate covariate measurement error in linear and nonlinear models in the past decades, especially when the error is classified as Berkson error or purely classical error. A simple method adjusting for classical error in covariates is the regression calibration method, which is to replace the unobserved true covariates by their conditional expectations given observed covariates in regression [7]. Other important methods dealing with the problem of classical error in covariates include the instrumental variable approach, the conditional score approach [3], the corrected score method [8, 9] and the simulation-extrapolation (SIMEX) method [10]. For the problem of nonlinear regression with Berkson error in the covariates, Whitmore and Keller [11] suggested an approximation method that reduces the bias induced by the error. Also, a minimum distance method is developed in [12] and maximum likelihood based methods are discussed in [13, 14] for the same problem.

Methods simply accounting for classical or Berkson error may not be applicable to a situation where the covariates are subject to a combination of both classical and Berkson errors and the relationship between the response and true covariates is nonlinear. Recently, several methods have been proposed to simultaneously adjust for the presence of mixture of Berkson and classical errors in covariates for a few generalized regression models. For example, Reeves *et al.* [2] studied a mixture of Berkson and classical errors model and suggested a regression calibration method for logistic regression. Considering the same problem, Mallick *et al.* [15] proposed a Bayesian method using Markov chain Monte-Carlo (MCMC) techniques and Li *et al.* [16] developed a Monte Carlo expectation-maximization (MCEM) approach. Kukush *et al.* [17] investigated a different measurement error model that also incorporates errors of both types and provided maximum likelihood-based methods for logistic regression. The available methods accounting for the effect of both classical and Berkson measurement errors in regression analysis generally require that the variances of the errors be known or related through a known function. However, these assumptions are hardly justifiable and may not be appropriate in many situations.

We are concerned with the problem of parameter estimation in generalized linear models when the covariates are possibly subject to a mixture of classical and Berkson errors and there is no replication or validation data for the mismeasured covariates. We assume that in a subset of the study cohort, an instrumental variable and another surrogate for the unobserved covariates are available. This subset is called the calibration sample. To our knowledge, this problem has not been well-discussed yet in the literature. It is practically very important to address this issue as we illustrate with the VAX004 real data example. The VAX004 study is a double-blind randomized trial of a vaccine to protect against HIV-1 infection. It involved 5403 adults volunteers (5095 men and 308 women) and was conducted in the United States, Canada, Puerto Rico and the Netherlands between 1998 and 2002. The study participants were either men who have sex with men or women at high risk for heterosexual HIV-1 transmission. During the trial, data were collected on variables including the occurrences of sexually transmitted infections, the number of male partners, HIV sero-status of partners in unprotected anal, vaginal or oral sex acts through questionnaires. More details regarding the study can be found in Flynn *et al.* [18]. In a regression analysis to study the effect of the number of HIV positive male partners and vaccine treatment on HIV infection, a naive method is to use the reported number of HIV positive male partners as a true covariate value. A problem is that the reported number of HIV positive male partners is potentially subject to recall errors (classical type). Also,

misclassification in HIV sero-status of a male partner may lead to an error in the total number of HIV positive male partners. However, the potential misclassification error may be because the subject did not know his partners' HIV sero-status, and hence is likely to be independent of the reported number of HIV positive partners. This leads to the concern that in addition to classical error due to recall, Berkson error may be involved. To address the concern of Berkson error, the use of a mixture of errors in the analysis may serve as a tool to test if Berkson error is involved. Another complication is the fact that there are no replicates for the number of HIV positive male partners. The development of new methods to adjust for the measurement errors in this situation is then appealing. We treat the number of unprotected anal or oral sex with HIV positive male partners as an instrumental variable. This variable is likely to be correlated with the number of HIV positive male partners and could serve as an instrument for the true underlying number of HIV positive male partners. Data on the number of HIV positive male partners were not available for some participants because they did not respond to the questions regarding the number of times they had unprotected anal or oral sex with their male partners. We allow the measurement error to include features of both Berkson error and classical error and develop an expected estimating equation (EEE) approach to account for such a feature of the measurement error. The proposed method needs no assumption regarding the mixture proportion of the error variances.

The rest of this paper is structured as follows. Section 2 describes the model for the mixture of classical and Berkson errors and the general form of the primary regression model. Section 3 provides a brief review of the naive approach, the regression calibration and SIMEX methods for the estimation of the regression parameters. This section further presents our proposed approach to accommodating the mixture of errors in the covariates. Section 4 shows the results of a simulation study investigating the finite-sample performance of the proposed method. In Section 5, we illustrate our method with an application to the VAX004 data. Section 6 concludes this work with a summary and discussion.

2 Model formulations

Let n be the number of study individuals. For individual i , let Y_i denote the response variable, X_i be the primary covariate that cannot be measured precisely and $\mathbf{Z}_i$ represent a vector of error-free covariates in a generalized linear model. For notational simplicity, we consider the case when X_i is univariate. Letting W_i be the observed version of X_i , we assume that W_i and X_i are related through the following mixture of Berkson and classical errors model.

$$\begin{cases} X_i = L_i + U_{bi}, \\ W_i = L_i + U_{ci}, \end{cases} \quad (1)$$

where L_i is a latent variable with mean μ_l and variance σ_l^2 , U_{bi} and U_{ci} are independent zero-mean measurement errors with variances σ_b^2 and σ_c^2 , respectively. The errors U_{bi} and U_{ci} are assumed to be independent of L_i and $\mathbf{Z}_i$. Note that model (1), which was also studied by Mallick *et al.* [15], Li *et al.* [16], Carroll *et al.* [19] and Apanasovich *et al.* [20] can be written under the form $W_i = X_i + U_{ci} - U_{bi}$. It embodies features of both Berkson error and classical error structures and reduces to a Berkson error model when $\sigma_c^2 = 0$ and a classical error model when $\sigma_b^2 = 0$. Moreover, the relationship between the response variable Y_i and covariates X_i and $\mathbf{Z}_i$ is specified as follows.

$$E(Y_i|X_i, \mathbf{Z}_i) = \phi(\beta_0 + \beta_1 X_i + \beta_2' \mathbf{Z}_i), \quad (2)$$

where $\beta = (\beta_0, \beta_1, \beta_2')'$ is a vector of unknown parameters of interest and $\phi(\cdot)$ is a known function. In particular, $\phi(u) = u$ for linear regression, $\phi(u) = \exp(u)$ for Poisson regression and $\phi(u) = \{1 + \exp(-u)\}^{-1}$ for logistic regression. Furthermore, for each subject i in the calibration sample, let M_i denote an instrumental variable for X_i . As noted in [3, 21], an instrumental variable for X_i is essentially correlated with X_i and uncorrelated with the measurement error $U_{ci} - U_{bi}$. The instrumental variable M_i is modeled as follows.

$$M_i = \alpha_0 + \alpha_1 L_i + \alpha_2' \mathbf{Z}_i + V_i, \quad (3)$$

where $\alpha = (\alpha_0, \alpha_1, \alpha_2')'$ is a vector of unknown parameters and V_i is zero-mean random variable, which is independent of L_i , U_{bi} , U_{ci} and $\mathbf{Z}_i$, $i = 1, \dots, n$. In addition, we assume the availability in the calibration sample of another surrogate variable Q_i for X_i satisfying

$$Q_i = \gamma_0 + \gamma_1 X_i + \gamma_2' \mathbf{Z}_i + \epsilon_i, \quad (4)$$

where $\gamma = (\gamma_0, \gamma_1, \gamma_2')'$ is a vector of unknown parameters and ϵ_i has mean zero and is independent of L_i , U_{bi} , U_{ci} , $\mathbf{Z}_i$ and V_i , $i = 1, \dots, n$. Let η_i indicate whether subject i is in the calibration sample or not and $\theta = P(\eta_i = 1)$. Hence, $\eta_i = 1$ if M_i and Q_i are available and $\eta_i = 0$ otherwise. We assume that given X_i and $\mathbf{Z}_i$, the response Y_i is independent of W_i , M_i and Q_i , $i = 1, \dots, n$. It is further assumed that η_i is independent of $(Y_i, \mathbf{Z}_i, L_i, U_{bi}, U_{ci}, V_i, \epsilon_i)$ and that $(Y_i, \mathbf{Z}_i, L_i, U_{bi}, U_{ci}, V_i, \epsilon_i, \eta_i)$, $i = 1, \dots, n$, are independent. Our main interest lies in the

estimation of the vector of parameters $\beta = (\beta_0, \beta_1, \beta_2')'$ based on all the observed data in three common generalized linear models, which are linear, logistic and Poisson regression models. In the following, we denote the entire available data for the i th individual by $\mathbf{O}_{1i} = (Y_i, W_i, M_i, Q_i, \mathbf{Z}_i)$ if $\eta_i = 1$ and $\mathbf{O}_{2i} = (Y_i, W_i, \mathbf{Z}_i)$ if $\eta_i = 0$, $i = 1, \dots, n$.

3 Estimation methods

In this section we first review briefly the naive, regression calibration (RC) and simulation-extrapolation (SIMEX) methods for the estimation of the parameter of interest β .

Afterwards, we present our approach to accounting for the errors in the estimation of this parameter using all the observed data.

3.1 Naive regression

In a situation where the covariate X_i is observed, a consistent estimator of β solves the following equation:

$$\sum_{i=1}^n \begin{pmatrix} 1 \\ X_i \\ \mathbf{Z}_i \end{pmatrix} \left\{ Y_i - \phi(\beta_0 + \beta_1 X_i + \beta_2' \mathbf{Z}_i) \right\} = 0. \quad (5)$$

Since X_i is not observed, a naive approach to estimating β in our context is to substitute X_i with W_i in (5) and to solve the resulting equation for β . Let $\hat{\beta}_{na}$ denote the estimator of β obtained through the naive method. In linear regression, it is easily seen that

$E(Y_i|W_i, \mathbf{Z}_i) = \beta_0 + \beta_1 W_i + \beta_2' \mathbf{Z}_i$, if $\sigma_c^2 = 0$ in which case $\hat{\beta}_{na}$ is unbiased. Hence, pure

Berkson error is not a major concern in linear regression. However, $\hat{\beta}_{na}$ is biased whenever classical error is present in linear regression. For example, if Z_i is univariate and $(L_i, U_{bi}, U_{ci}, Z_i)$ is a multivariate normal random variable, it can be shown that

$$E(X_i|W_i, Z_i) = (1 - \lambda_1) \mu_l - \lambda_2 \mu_z + \lambda_1 W_i + \lambda_2 Z_i \text{ and}$$

$$\text{Var}(X_i|W_i, Z_i) = \sigma_b^2 + (1 - \lambda_1) \sigma_l^2 - \lambda_2 \sigma_{lz}, \text{ where}$$

$$\lambda_1 = (\sigma_l^2 \sigma_z^2 - \sigma_{lz}^2) / \{ \sigma_z^2 (\sigma_l^2 + \sigma_c^2) - \sigma_{lz}^2 \} \text{ and } \lambda_2 = \sigma_c^2 \sigma_{lz} / \{ \sigma_z^2 (\sigma_l^2 + \sigma_c^2) - \sigma_{lz}^2 \}. \text{ Here, } \sigma_{lz} = \text{cov}(L_i, Z_i), \text{ and } \mu_z \text{ and } \sigma_z^2 \text{ are the mean and variance of } Z_i, \text{ respectively. As a result,}$$

$$E(Y_i|W_i, Z_i) = \beta_0 + \beta_1 \{ (1 - \lambda_1) \mu_l - \lambda_2 \mu_z \} + \beta_1 \lambda_1 W_i + (\beta_2 + \lambda_2 \beta_1) Z_i, \text{ and it is clear that}$$

$\hat{\beta}_{na}$ is biased in linear regression if $\sigma_c^2 > 0$ under normality assumption. In logistic, it is well-known that $\hat{\beta}_{na}$ is not consistent in a Berkson error setting [11] or classical error situation [3, 22]. Hence, it is obvious that the naive estimation may not work in logistic regression when errors of both types are present in the covariates. Furthermore, it can be shown for Poisson

regression that $E(Y_i|W_i, Z_i) = \exp(\beta_0 + \sigma_b^2 \beta_1^2 / 2 + \beta_1 W_i + \beta_2' Z_i)$ in a normal Berkson error case, suggesting that the naive estimators for β_1 is unbiased in this case though, the estimation of the intercept is affected by the error. Also,

$E(Y_i|W_i, Z_i) = \exp(\beta_0 + \beta_2' Z_i) E\{\exp(\beta_1 X_i) | W_i, Z_i\}$ indicating that $\hat{\beta}_{na}$ is biased when the error is classical as noted in [21]. Therefore, the naive method could result in a biased estimator of β in Poisson regression when the covariates are subject to a mixture of errors of both types. In general, the combination of Berkson and classical errors in the covariates may accentuate the attenuation phenomenon in nonlinear regression.

3.2 RC and SIMEX methods

A simple alternative method that could be applied to reduce the attenuation bias in the estimation of β is the RC approach, which was investigated by Reeves *et al.* [2], Wang *et al.* [7] and Kuha [23]. The main idea of the RC method is to replace the unobserved covariates with their conditional expectations given the observed covariates in the estimating equations for β . The method is implemented in our problem by substituting $E(X_i|W_i, Z_i)$ for X_i in (5). It is here referred to as the RC1 method and the corresponding estimator is denoted by $\hat{\beta}_{rc1}$. Another RC approach that we termed as RC2 method consists to replace X_i in (5) with $E(X_i|W_i, M_i, Q_i, Z_i)$ if $\eta_i = 1$ and $E(X_i|W_i, Z_i)$ otherwise. Let $\hat{\beta}_{rc2}$ be the estimator, which is obtained based on the RC2 method. $\hat{\beta}_{rc2}$ is expected to perform better than $\hat{\beta}_{rc1}$ since it uses additional information provided by the available data in the calibration sample. When the primary regression model is linear, it is clear that

$E(Y_i|W_i, Z_i) = \beta_0 + \beta_1 E(X_i|W_i, Z_i) + \beta_2' Z_i$. This shows that both $\hat{\beta}_{rc1}$ and $\hat{\beta}_{rc2}$ are consistent estimators for β in linear regression when the covariates are subject to both classical and Berkson errors. However, they may be biased in logistic or Poisson regression.

In fact, both estimators rely on an approximation to $E\{\phi(\beta_0 + \beta_1 X_i + \beta_2' Z_i) | W_i, Z_i\}$ based on a first order Taylor series expansion of this expression about $E(X_i|W_i, Z_i)$, where $\phi(\cdot)$ was defined in (2). The biasedness of $\hat{\beta}_{rc1}$ or $\hat{\beta}_{rc2}$ in nonlinear regression becomes obvious when a second order Taylor series expansion is considered. For example, it can be obtained from (2) that

$E(Y_i|W_i, Z_i) \approx \phi\{\beta_0 + \beta_1 E(X_i|W_i, Z_i) + \beta_2' Z_i\} + 0.5 \beta_1^2 \left[\frac{\partial^2 \phi\{\beta_0 + \beta_1 E(X_i|W_i, Z_i) + \beta_2' Z_i\}}{\partial X^2} \right] \text{Var}(X_i|W_i, Z_i)$. Therefore, it is not difficult to see that the approximation, which is based on a first order Taylor series expansion may not be satisfactory when $\beta_1^2 \text{Var}(X_i|W_i, Z_i)$ is large. Also, it can be noted that this term is an increasing function of $|\beta_1|$, σ_b^2 and σ_c^2 under normality

assumption for all the variables. Therefore, $\hat{\beta}_{rc1}$ and $\hat{\beta}_{rc2}$ are not consistent in logistic or Poisson regression when the covariates are prone to a mixture of Berkson and classical errors.

Another approximate bias-correction method that has gained substantial attention in the literature is the SIMEX approach, which was first suggested by Cook and Stefanski [10] to deal with the problem of classical measurement error in covariates for linear and nonlinear regression models. It was further discussed in [3, 24, 25]. Moreover, Apanasovich et al. [20] adapted the method to a situation where the mismeasured covariates are modeled purely nonparametrically, purely parametrically or have components that are modeled both parametrically and nonparametrically. They used a kernel-based method to estimate the model parameters. Generally speaking, SIMEX is a simulation-based method that adjusts for the measurement error through the use of a two-step procedure consisting of a simulation step followed by an extrapolation step. The simulation step involves the specification of a naive estimation procedure that would lead to a consistent estimator of the model parameter in the absence of measurement error and the construction of a large number of naive estimates based on additionally simulated data sets of errors with gradually increasing variance. The extrapolation step consists to extrapolate back to the situation of no measurement error using an extrapolant function. The application of the method to our situation is described as follows. Assuming that the conditional distribution of X_i given L_i and $\mathbf{Z}_i$ does not involve and that U_b is normally distributed, it can be noted that if L_i in (1) could be observed, the maximum likelihood estimation of β would be to solve

$$\sum_{i=1}^n \frac{\partial}{\partial \beta} \log \{ \mathcal{L}(Y_i | L_i, \mathbf{Z}_i, \beta) \} = 0, \quad (6)$$

where $\mathcal{L}(Y_i | L_i, \mathbf{Z}_i, \beta) = \int \mathcal{L}(Y_i | X_i = x, \mathbf{Z}_i, \beta) \mathcal{L}(X_i = x | L_i, \mathbf{Z}_i)$ is the conditional distribution of Y_i given $L_i, \mathbf{Z}_i$, and $\mathcal{L}(\cdot)$ denotes likelihood function. We recall that L_i is not observable and replacing it by W_i in (6) will lead to a biased estimation of β . In the simulation step of the SIMEX procedure to reduce the bias, for a non-negative real value, we generate new data $W_{\zeta, r, i} = W_i + \sigma_c \sqrt{\zeta} U_{r, i}, i = 1, \dots, n, r = 1, \dots, R$, where $U_{r, i}$ are generated as independent and identically distributed random variables following the standard normal distribution. Here R represents the number of simulated data sets. Let $\hat{\beta}_r(\zeta_j)$ denote the estimate of β obtained by replacing L_i with $W_{\zeta, r, i}$ in (6) and denote

$\hat{\beta}(\zeta_j) = R^{-1} \sum_{r=1}^R \hat{\beta}_r(\zeta_j)$, where $0 = \zeta_0 < \zeta_1 < \dots < \zeta_J$ and $J > 1$. The extrapolation step consists to fit a regression model of $\hat{\beta}(\zeta_j)$ on $\zeta_j, j = 1, \dots, J$, using the ordinary least squares estimation method. The SIMEX estimate of β , denoted by $\hat{\beta}_{simex}$ is then obtained by extrapolating back to the case when $\zeta = -1$, which represents the situation of no measurement error. A routinely used extrapolant function is the quadratic function. The covariance matrix of $\hat{\beta}_{simex}$ can be estimated by means of standard bootstrap method. A problem associated with the SIMEX procedure is that it is an approximate method, subject to the choice of the extrapolant function. Furthermore, similar to RC, the procedure may not work well in nonlinear regression when the variance of the measurement error is large or when $|\beta_1|$ is large [25]. We pursue a more reliable estimation approach, which is based on all the observed data in the following subsection.

3.3 Expected estimating equation approach

The expected estimating equation (EEE) method was investigated by Wang *et al.* [26] in a situation where the response variable and some covariates may be missing, misclassified or

subject to classical measurement error and multiple unbiased surrogates for the error-prone variables are available. In our problem, the covariates are possibly subject to both Berkson and classical errors and there is only one unbiased surrogate. Moreover, a calibration subsample data is available only for some individuals. We propose the EEE estimator for β that solves the following estimating equation:

$$\sum_{i=1}^n \eta_i E \{ \mathcal{S}(Y_i, X_i, \mathbf{Z}_i) | \mathbf{O}_{1i} \} + (1 - \eta_i) E \{ \mathcal{S}(Y_i, X_i, \mathbf{Z}_i) | \mathbf{O}_{2i} \} = 0, \quad (7)$$

where $\mathcal{S}(Y_i, X_i, \mathbf{Z}_i) = (1, X_i, \mathbf{Z}_i')' \{ Y_i - \phi(\beta_0 + \beta_1 X_i + \beta_2' \mathbf{Z}_i) \}$. We denote the solution to (7) by $\hat{\beta}_{eee}$ and refer to it as the EEE estimator for β . The calculations of conditional expectations involved in (7) and the evaluations of $E(X_i | W_i, \mathbf{Z}_i)$ and $E(X_i | W_i, M_i, Q_i, \mathbf{Z}_i)$ for RC1 and RC2 methods require specifications of the distribution functions of L_i, U_{bi}, U_{ci}, V_i and ε_i which do not have to be necessarily normal. Moreover, letting $\mathbf{O}_i = \mathbf{O}_{1i}$ if $\eta_i = 1$ and $\mathbf{O}_i = \mathbf{O}_{2i}$ otherwise, the conditional expectation $E \{ \mathcal{S}(Y_i, X_i, \mathbf{Z}_i) | \mathbf{O}_i \}$ can be evaluated as

$$\frac{\int_L \int_U \mathcal{S}(Y_i, X_i = L + U, \mathbf{Z}_i) \mathcal{L}(Y_i | X_i = L + U, \mathbf{Z}_i) \mathcal{L}(W_i | L) \mathcal{F}_i(L, U)^{\eta_i} \mathcal{L}(L, \mathbf{Z}_i) \mathcal{L}(U) dL dU}{\int_L \int_U \mathcal{L}(Y_i | X_i = L + U, \mathbf{Z}_i) \mathcal{L}(W_i | L) \mathcal{F}_i(L, U)^{\eta_i} \mathcal{L}(L, \mathbf{Z}_i) \mathcal{L}(U) dL dU},$$

where $\mathcal{L}(\cdot)$ denotes likelihood function and

$\mathcal{F}_i(L, U) = \mathcal{L}(Q_i | X_i = L + U, \mathbf{Z}_i) \mathcal{L}(M_i | L, \mathbf{Z}_i)$. The expectations $E(X_i | W_i, \mathbf{Z}_i)$ and $E(X_i | W_i, M_i, Q_i, \mathbf{Z}_i)$ can be calculated similarly. The integrals in the above expression can be evaluated by means of numerical integration techniques including the Gauss-Hermite quadrature rule and the trapezoidal integration rule, which involves uniformly partitioning acceptable ranges of the L and U_b axes into a specified number of intervals and applying Riemann summation techniques.

A tacit assumption that has been made so far is the knowledge of the nuisance parameters including $\mu_l, \alpha, \gamma, \sigma_l^2, \sigma_b^2, \sigma_c^2, \sigma_v^2$ and σ_ε^2 which are needed for the calculation of conditional expectations. In practice, these parameters may need to be estimated from the data. If σ_b^2 is known or is linked to σ_c^2 through a known bijective function, then the assumption of the availability of data on Q_i can be relaxed. In this case $\mathbf{O}_{1i} = (Y_i, W_i, M_i, \mathbf{Z}_i)$ and all the nuisance parameters can be identified based on the observations $\mathbf{O}_{1i}$ such that $\eta_i = 1, i = 1, \dots, n$. Moreover, conditional expectations for the estimation of β by the RC2 method will not involve Q_i . In a more general situation when σ_b^2 is unknown and there is no assumed relationship between the error variances, the data on Q_i will serve as an additional information for the estimation of the nuisance parameters. In such a situation, estimating equations for the vector of nuisance parameters, denoted by ν , can be obtained based on moment calculations using the observed data on Y_i, M_i, Q_i, W_i and $\mathbf{Z}_i$ for subjects in the calibration subsample. For example, estimating equations based on moment calculations are given in Appendix B when the primary regression model is assumed to be linear and does not involve $\mathbf{Z}$. Let $\Omega = (\beta, \nu)'$, write (7) as $\sum_{i=1}^n \Psi_{1i}(\beta) = 0$ and define $\Psi_i(\Omega) \equiv \{ \Psi'_{1i}(\beta), \Psi'_{2i}(\nu) \}'$, where $\sum_{i=1}^n \Psi_{2i}(\nu) = 0$ is an estimating equation for ν . Furthermore, let Ω_0 be the true value of Ω and $\hat{\Omega}$ denote the solution to $\sum_{i=1}^n \Psi_i(\Omega) = 0$. The asymptotic properties of $\sqrt{n}(\hat{\Omega} - \Omega_0)$ are given by the following result.

Proposition: Under the regular conditions (C1)-(C6) in Appendix A, $\sqrt{n}(\hat{\Omega} - \Omega_0)$ is asymptotically normally distributed with mean zero and covariance matrix

$\mathcal{A}^{-1}(\Omega_0) \mathcal{B}(\Omega_0) \left\{ \mathcal{A}^{-1}(\Omega_0) \right\}'$, where $\mathcal{A}(\Omega_0)$ and $\mathcal{B}(\Omega_0)$ are the limits in probability of $\widehat{\mathcal{A}}_n(\Omega_0) = n^{-1} \sum_{i=1}^n \partial \Psi_i(\Omega_0) / \partial \Omega$ and $\widehat{\mathcal{B}}_n(\Omega_0) = n^{-1} \sum_{i=1}^n \Psi_i(\Omega_0) \Psi_i'(\Omega_0)$, respectively.

The proof of the proposition is sketched in Appendix A. The asymptotic covariance matrix of $\sqrt{n}(\hat{\Omega} - \Omega_0)$ can be consistently estimated by $\widehat{\mathcal{A}}_n^{-1}(\hat{\Omega}) \widehat{\mathcal{B}}_n(\hat{\Omega}) \left\{ \widehat{\mathcal{A}}_n^{-1}(\hat{\Omega}) \right\}'$.

4 Simulation study

We investigated the finite-sample performances of the methods discussed in the previous section through a simulation study. We considered linear regression, logistic regression and Poisson regression models for the response Y with a single explanatory variable X , which is scalar. In the simulations, the latent variable L was generated from $N(\mu_l, \sigma_l^2)$ with $\mu_l = 0.5$ and $\sigma_l^2 = 1$. The Berkson error U_b and the classical error U_c followed zero-mean normal distributions with variances σ_b^2 and σ_c^2 , respectively. Also, X was simulated from the model $X = L + U_b$ and the observed version of X was drawn as $W = L + U_c$. We set $\sigma_b^2 = 0.2$ or 0.4 , and $\sigma_c^2 = 0.3$ or 0.5 to study the separate and combined effects of the errors on the performances of the methods. Moreover, the proportion of the calibration data was chosen as $\theta = 0.5$ or 0.7 to investigate how the results evolve with the size of the available data in the calibration sample. The instrumental variable was simulated following the model $M = \alpha_0 + \alpha_1 L + V$, where V was normal with mean 0 and variance $\sigma_v^2 = 1$. We took $\alpha_0 = 1$ and $\alpha_1 = -2$. Furthermore, the additional variable in the calibration sample was simulated as $Q = \gamma_0 + \gamma_1 X + \varepsilon$, where $\gamma_0 = -1$, $\gamma_1 = 2$ and ε was normally distributed with mean zero and variance $\sigma_\varepsilon^2 = 1$. The variable η , indicating whether M and Q are available or not was generated from the Bernoulli distribution with probability of success $P(\eta = 1) = \theta$. For the linear regression, the response variable followed the model $Y = \beta_0 + \beta_1 X + e$, where e is normal with mean zero and variance $\sigma_e^2 = 1$. In the logistic regression case, Y was simulated from the Bernoulli trial with probability of success $\{1 + \exp(-\beta_0 - \beta_1 X)\}^{-1}$. We generated Y as Poisson random variable with mean $\exp(\beta_0 + \beta_1 X)$ for Poisson regression.

A total of 400 Monte Carlo samples of size $n = 500$ were generated in the simulations for each case. We estimated the parameter of interest $\beta = (\beta_0, \beta_1)'$ based on the naive method, the RC methods (RC1 and RC2), the SIMEX method and the EEE approach. The naive estimator simply replaces X by W in (5) with no covariates Z . The RC1 estimator uses the conditional expectation $E(X|W)$ in place of X in (5) without Z . The RC2 substitutes $E(X|W, M, Q)$ for X in (5) if $\eta = 1$ and replaces X by $E(X|W)$ in (5) when there is no calibration data ($\eta = 0$). The EEE estimator, which solves (7) without Z involved in the equation uses all observed data and accounts for the mixture of Berkson and classical errors. For the SIMEX method, we created $R = 50$ additional data sets of measurement error at each point $\zeta \in \{0, 0.5, 1, 1.5, 2\}$ in the simulation step and used the quadratic model in the extrapolation step. The estimators were evaluated with regard to their biases (Bias), sample standard deviation of the estimates (SD), average of the estimated standard errors (ASE) of the estimators and coverage probabilities (CP) of their 95% Wald confidence intervals. The standard errors of the RC1, RC2 and EEE estimators were computed using the sandwich standard error estimation approach. The estimation of the standard error for the SIMEX estimator was based on the bootstrap method, resampling 30 times. Integrals involved in the evaluation of

condition expectations for the implementation of RC2, SIMEX or EEE method were computed using Gauss-Hermite integration techniques with 10 quadrature points.

Table 1 displays the simulation results for the estimation of β in the linear regression case, where we set $\beta = (-2, 2)'$. It can be seen from this table that the naive estimator for β performs poorly as expected. It shows large biases and very low coverage probabilities of less than 50%. The biases of the naive estimator increase as the magnitude of the classical error increases. However, these biases appear unaffected to some extent by the Berkson error, indicating that the naive estimator would be unbiased if the classical errors were absent. The RC1, RC2, SIMEX and EEE estimators on the other hand work well, showing small biases and coverage probabilities that are close to the nominal level of 95%. Larger variance of Berkson or classical error have the effect of increasing the standard errors of the RC1, RC2, SIMEX and EEE estimators. Furthermore, the performances of these estimators in terms of efficiencies are better when the proportion of calibration data is larger ($\theta = 0.7$). The SIMEX estimator appears to be more efficient than the regression calibration and EEE estimators in these simulations. It can also be seen that the RC2 performs better than the RC1 with regards to biases and standard errors. This is probably due to the fact that RC2 estimator uses more information from the data than RC1 estimator does. Moreover, the EEE estimator for β_1 is more efficient than the RC2 estimator when the proportion of calibration data is small ($\theta = 0.5$) and the magnitude of the classical error is greater than that of the Berkson error. The advantage seems to vanish as θ gets larger ($\theta = 0.7$). A plausible explanation is that the EEE estimator makes more efficient use of the data over RC2 when using the non-calibration data, while they perform equally well in the calibration data. In Table S1 in the Supplementary materials we have demonstrated the phenomenon that in general when the proportion of the calibration sample is smaller the efficiency gain of the EEE over RC2 is larger. The simulation settings for this table were similar to those for Table 1 with the difference that we set $\sigma_b^2 = 0.2$ and $\theta = 0.3, 0.5, 0.7$ or 0.9 to examine the effect of the proportion of the calibration data on the parameter estimation. The results in this table indicated that the efficiency gain of EEE over RC2 decreases as the proportion of calibration data gets larger.

Table 2 shows the results for the estimation of the vector of nuisance parameters

$\nu = (\mu_1, \alpha_0, \alpha_1, \gamma_0, \gamma_1, \sigma_b^2, \sigma_c^2, \sigma_e^2, \sigma_v^2, \sigma_\epsilon^2)'$ by the method of moments in the linear regression case using data on Y_i, W_i, M_i and Q_i for individuals in the calibration sample ($\eta_i = 1$). The estimators for all components of ν show reasonable biases and coverage probabilities close to the nominal level. In addition, their standard errors are generally increasing functions of the variances of both Berkson and classical errors. Also, the efficiencies of the estimators improve as the proportion of the calibration data gets larger.

Table 3 shows the simulation results for the estimation of β in the logistic regression case, where the regression parameter was set $(\beta_0, \beta_1)' = (-1, \ln(5))'$. It appears that the naive estimator performs unsatisfactorily, showing large biases and coverage probabilities well below 50%. Its performance worsens as σ_b^2 or σ_c^2 becomes large. The effect of the classical error on the bias from the naive estimator seems to be more pronounced than that of the Berkson error. Also, the RC1, RC2 and SIMEX estimators are less biased than the naive one, but still exhibit unacceptable biases and coverage probabilities of less than 90%. It can be seen that their biases increase as σ_b^2 or σ_c^2 gets large, indicating that the RC1, RC2 and SIMEX methods may not work well in logistic regression when the covariates are subject to Berkson error, classical error or a mixture of errors of both type. The EEE estimator on the other hand performs satisfactorily. It has reasonable biases and coverage

probabilities close to 95%. Its standard errors get larger as σ_b^2 or σ_c^2 increases. Moreover, its performance improves as the proportion of calibration subsample increases.

In Table 4, we considered the case when the regression model is Poisson with parameter $\beta = (0.5, \ln(2))'$. From this table, it can be observed that the naive estimator performs very poorly as in Tables 1 and 3. Moreover, the RC1 and RC2 estimators for β_1 show acceptable biases and coverage probabilities. The RC2 estimator for β_1 appears to be more efficient than the RC1 estimator. Note however that both RC1 and RC2 methods lead to substantially biased estimation for β_0 . Also, the coverage probabilities of RC1 and RC2 estimators for β_0 are below 90%. The SIMEX estimator for β_0 has smaller bias when compared with the RC1 and RC2 counterparts. Furthermore, the SIMEX estimator for β_1 appears to work well when σ_c^2 is small ($\sigma_c^2=0.3$). It has a similar performance with the EEE estimator when $\sigma_b^2=0.2$, $\sigma_c^2=0.3$ and $\theta=0.7$. However, it shows considerable biases and low coverage probabilities (less than 90%) when σ_b^2 is small ($\sigma_b^2=0.2$) and σ_c^2 is large ($\sigma_c^2=0.5$). In contrast, the EEE method provides acceptable results for the estimation of both β_0 and β_1 with small biases and coverage probabilities not far from the nominal level of 95%. It is generally more efficient than the RC2 estimator, but the efficiency gain was small when $\sigma_b^2=0.2$.

In Table 5, we investigated the performances of the methods in logistic regression when the true model for the measurement error was a mixture of both Berkson error and classical error, but the estimation of the regression coefficients ignored a component of the mixture and incorrectly assumed a classical or Berkson error model. The simulation settings were same as in Table 3 and we set $\theta=0.7$ for simplicity. It can be noted that when the estimation wrongly assumes a classical error model, the naive, RC1, RC2 and SIMEX estimators perform poorly as in Table 3 for $\theta=0.7$. They have large biases and low coverage probabilities. The biases from the RC1 estimation for β_1 are same as in Table 3 while those from the RC2 method appear to be smaller. Moreover, the bias problem of the SIMEX estimator for β_1 is more severe than in Table 3. The EEE estimator surprisingly still works acceptably under the simulation settings. Its standard errors are smaller than those reported in Table 3 with $\theta=0.7$. In contrast, all the four bias-adjusting approaches show considerable biases and very low coverage probabilities under the wrong assumption of a Berkson error model. Similarly, we examined the sensitivity of the methods to the misspecification of the measurement error model in linear regression and Poisson regression in simulations. The sensitivity results for linear regression are presented in Table S2 and those for Poisson regression are shown in Table S3 in the Supplementary materials. The poor performances of all methods can also be noted in both Tables S2 and S3 when the true error model is a mixture of errors of both types but the estimation is carried out under a misspecified Berkson error model. Furthermore, incorrectly assuming a classical error model when the covariates are contaminated by both Berkson and classical errors in linear regression may lead to a biased estimation by all the methods except RC1, which shows a good performance while the RC2, SIMEX and EEE methods result in large biases and low coverage probabilities in Table S2. In addition, the RC1 estimator for the slope has small biases although the intercept estimation is biased in Poisson regression (Table S3) when the estimation incorrectly assumes a classical error model rather than the true mixture model for the measurement error. The bias problem for the RC2, SIMEX or EEE method is also observed in Table S3 under the misspecified measurement error models.

The simulation results indicate that both the Berkson error and classical error need to be properly adjusted in regression analysis when they are both present in the covariates. Moreover, when the measurement error is modeled correctly, the RC1, RC2, SIMEX and EEE methods satisfactorily accommodate the presence of both Berkson and classical errors in the covariates in linear regression. In logistic regression, the RC1, RC2 and SIMEX

methods may not perform well especially when $|\beta_1|$ is large or the variance of the Berkson or classical error is substantial. But, their performance may be acceptable when $|\beta_1|$ and the error variances are not large as can be seen in Table S4 of the Supplementary materials. In Poisson regression, the RC1 and RC2 methods may yield acceptable results for the estimation of the slope when all variables involved are normal. However, estimators of the intercept based on the regression calibration methods are generally inconsistent. Furthermore, the SIMEX method may produce acceptable results in Poisson regression when the variance of the classical error is small. But it may lead to a severely biased estimation in some cases when the variance of the classical error is large. The EEE approach in contrast works well not only in linear regression but also in logistic regression and Poisson regression. Its performance improves when the proportion of calibration subsample or the cohort sample size is large. All the four error-adjusting methods are sensitive to misspecification of the model for the measurement error. Incorrectly assuming a Berkson or classical error model could lead to biased estimations by the RC1, RC2, SIMEX and EEE methods in linear, logistic or Poisson regression when the true error model is a mixture of both Berkson and classical errors. Hence, it is important to incorporate both the Berkson error and classical error in the measurement error model (e.g. by using the mixture model) in regression analysis when it is suspected that the covariates are contaminated by errors of both types.

5 Application

We applied the proposed method to data from VAX004 study, which was briefly introduced in Section 1. In this application, we are interested in evaluating the effect of the number of HIV positive male partners and the vaccine treatment on HIV infection. There were totally 5403 participants and 368 cases of HIV infections in the study period. The number of HIV positive male partners was self-reported and hence, was potentially subject to measurement errors. The measurement errors possibly include recall error, which is common in self-reported data [27]. They may also be due to the difficulty for participants to accurately determine the HIV positive status of their male sex partners because sex partners may not tell that they are HIV positive. As noted in [28, 29], HIV positive sex partners are reluctant to disclose their true sero-status in many situations. Such an error could affect the true number of HIV positive male partners. Two variables that may be related to the true number of HIV positive male partners are the number of unprotected anal or oral sex with HIV positive male partners and the total number of sex acts with male partners. Totally, there were 5081 participants who reported the number of times they engaged in unprotected anal or oral sex activities with HIV positive male partners.

We modeled the relationship between occurrence of HIV infection (Y) and the covariates of interest, which are the logarithmic transformation of the true number of HIV positive male partners (X) and the indicator variable for vaccine treatment (Z), by $E(Y|X, Z) = \{1 + \exp(-\beta_0 - \beta_1 X - \beta_2 Z)\}^{-1}$. The measurement error was modeled as in (1) as we allowed the measurement error to include both classical error and Berkson error features. An unbiased surrogate (W) for X was the logarithmic transformation of the reported number of HIV positive male partners. Furthermore, it is highly probable that the logarithmic transformation of the reported number of unprotected anal or oral sex with HIV positive male partners (M) is correlated with the logarithmic transformation of the number of HIV positive male partners. M was assumed to be independent of the measurement error and treated as an instrumental variable for X . In addition, the logarithmic transformation of the reported total number of sex acts with male partners was considered as another surrogate (Q) for X . The total number of sex acts with all male partners was thought to encompass the error involved in the true number of HIV positive male partners. There were 5081 individuals in the calibration subsample. We ran a logistic regression with outcome Z and

covariate M to examine whether Z and M are associated, based on the data in the calibration subsample. The resulting estimate of the coefficient of M was -0.024 with standard deviation 0.028 . Similarly, we conducted a logistic regression of Z on Q and the estimate of the coefficient of Q was 0.01 with standard deviation 0.02 . It appeared from these results that vaccine treatment was not significantly associated with the log-transformed number of anal or oral sex acts with HIV positive male partners, nor with the log-transformed total number of sex acts with male partners. In the analysis, the variables M and Q were modeled as in (3) and (4), respectively, but independently of the indicator variable for vaccine treatment.

The analysis results in Table 6 indicate that the estimate of the variance of the classical error ($\sigma_c^2=0.144$, $SE = 0.026$) is significant (p -value < 0.001) at the 5% significance level, while that of the Berkson error ($\sigma_b^2=0.027$, $SE = 0.029$) appears to be statistically not significant (p -value $= 0.363$). This suggests that the log-transformed number of HIV positive male partners is affected by some measurement error, which is purely classical. The results of the estimation of $\beta = (\beta_0, \beta_1, \beta_2)'$ by the RC1, RC2, SIMEX and EEE methods assuming a classical error model are reported in Table 6. For comparison purpose, we also presented the results of the analysis with the mixture of Berkson and classical error model. The results of the analysis assuming a Berkson error model were omitted here because the estimate of σ_c^2 was significantly different from zero (p -value < 0.001) and it was very likely that the self-reported number of HIV positive male partners was contaminated with classical error. The SIMEX estimates were based on a quadratic extrapolant function with $\zeta = 0, 0.5, 1, 1.5, 2$, and $R = 500$. The bootstrap procedure resampling 50 times was used to obtain the standard errors of the SIMEX estimates of the regression coefficients while the standard errors for the RC1, RC2 and EEE estimates were estimated using the sandwich method.

The results of the analysis that assumed a mixture of Berkson and classical errors model were similar to those of the analysis with a classical error model probably because the variance of the Berkson error was small although not statistically significant. Furthermore, the standard errors of the RC2, SIMEX and EEE estimates for β_1 from the analysis assuming a classical error model were smaller than those from the analysis using the mixture of Berkson and classical errors model to some extent. Since σ_b^2 was not significant, we later relied on the results of the analysis assuming a classical error model for interpretations. The estimates by the five methods of β_1 , which is the coefficient of the log-transformed number of HIV-positive male partners were significant (p -value < 0.001) and positive, suggesting that the log-transformed number of HIV positive male partners significantly affects the probability of getting HIV infection. Also, the risk of HIV infection increases as the number of HIV positive male partners gets large. The RC1, RC2, SIMEX and EEE methods showed stronger effect of the log-transformed number of HIV positive male partners on HIV infection than the naive approach. Moreover, RC2 and EEE estimates of β_1 were very close and both larger than the SIMEX estimate. They were less efficient than the RC1 estimate. Also, it appeared that all five methods did not show evidence of significant effect of the trial vaccine on the risk of HIV infection. This was consistent with findings that were reported in Flynn *et al.* [18]. A plausible explanation of the fact that all the methods including the naive one produced similar results is that the variances of the measurement errors were small.

6 Conclusion

We have proposed a method to deal with the problem of generalized linear regression analysis when some covariates are possibly subject to classical errors, Berkson errors or a combination of both types of errors. The method does not require replicates for the error-prone covariates in the situation under consideration. The proposed approach is based on

expected estimating equation techniques and uses data that are available only for some subjects in a calibration sample to adjust for the combination of the classical and Berkson measurement errors. It requires no assumption about the mixture percentage of the error variances. Simulation studies have revealed a good performance of the method in handling the presence of both errors in the covariates in linear, logistic or Poisson regression. It is more reliable than the RC and SIMEX methods when the variance of the Berkson or classical error is large in logistic or Poisson regression. It is also superior to the RC methods in linear regression when the proportion of the calibration subsample is small. The approach presented in this work can be extended to survival analysis models when some covariates are subject to both Berkson and classical errors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the National Institute of Health grants P01CA53996 (Wang), R01ES017030 (Wang and Tapsoba), a travel award from the Mathematics Research Promotion Center of National Science Council of Taiwan (Wang) and the National Science Foundation grant 101-2118-M035-004-MY2 (Lee).

Appendix A: Regularity conditions and proposition proof

Regularity conditions

- (C1) ω_0 lies in the interior of a compact set $\mathcal{C}$.
- (C2) $\Psi(\cdot)$ is continuously differentiable with respect to, $\Omega, i = 1, \dots, n$.
- (C3) $E(L^2) < \infty, E(U_b^2) < \infty, E(U_c^2) < \infty$ and $E(\mathbf{Z}'\mathbf{Z}) < \infty$.
- (C4) $E(\eta_i) > 0$.
- (C5) $E\{\Psi_i(\Omega)\beta_i(\Omega)\} < \infty, E\{\Psi_i(\Omega)/\Omega\} < \infty$ for each $\Omega \in \mathcal{C}$.
- (C6) $E\{\Psi_i(\Omega_0)/\Omega\}$ is positive definite.

Proof of Proposition

If X_i were observed β would be estimated by solving (5), which is a maximum likelihood score estimating equation and can also be written as $\sum_{i=1}^n \partial \log \{\mathcal{L}(Y_i|X_i, \mathbf{Z}_i)\} / \partial \beta = 0$. Because X_i is not observable in our context, the estimation is based on the likelihood of all

observed data which are $(Y_i, W_i, \mathbf{Z}_i', \eta_i M_i, \eta_i Q_i)', i = 1, \dots, n$. The contribution to the likelihood of the observed data for subject i can be written as

$\mathcal{L}(Y_i, W_i, \mathbf{Z}_i, \eta_i M_i, \eta_i Q_i) = \mathcal{L}(Y_i|W_i, \mathbf{Z}_i, M_i, Q_i)^{\eta_i} \mathcal{L}(Y_i|W_i, \mathbf{Z}_i)^{1-\eta_i} \mathcal{L}(W_i, \mathbf{Z}_i, M_i, Q_i)^{\eta_i} \mathcal{L}(W_i, \mathbf{Z}_i)^{1-\eta_i}$. Assuming that both $\mathcal{L}(W_i, \mathbf{Z}_i, M_i, Q_i)$ and $\mathcal{L}(W_i, \mathbf{Z}_i)$ do not involve β , the estimator of β based on the likelihood of all observed data solves

$$\sum_{i=1}^n \eta_i \frac{\partial}{\partial \beta} \log \{\mathcal{L}(Y_i|W_i, \mathbf{Z}_i, M_i, Q_i)\} + (1 - \eta_i) \frac{\partial}{\partial \beta} \log \{\mathcal{L}(Y_i|W_i, \mathbf{Z}_i)\} = 0.$$

Note that $\partial \log \{ \mathcal{L}(Y_i|W_i, \mathbf{Z}_i, M_i, Q_i) \} / \partial \beta$ can be written as

$$\frac{\frac{\partial}{\partial \beta} \int_L \int_U \mathcal{L}(Y_i|X_i=L+U, \mathbf{Z}_i) \mathcal{L}(\mathbf{W}_i|L) \{ \mathcal{L}(Q_i|X_i=L+U, \mathbf{Z}_i) \mathcal{L}(M_i|L, \mathbf{Z}_i)^{\eta_i} \mathcal{L}(L, \mathbf{Z}_i) \mathcal{L}(U) dL dU \}}{\mathcal{L}(Y_i, W_i, \mathbf{Z}_i, M_i, Q_i)},$$

which can also be expressed as $E[\partial \log \{ \mathcal{L}(Y_i|X_i, \mathbf{Z}_i) \} / \partial \beta | \mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i, \mathbf{M}_i, \mathbf{Q}_i]$.

Similarly, it can be obtained that

$\partial \log \{ \mathcal{L}(Y_i|W_i, \mathbf{Z}_i) \} / \partial \beta = E[\partial \log \{ \mathcal{L}(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i) \} / \partial \beta | \mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i]$. It follows that under conditions (C1)-(C4), equation (6) is an unbiased estimating equation for β given the

nuisance parameter and that $\sum_{i=1}^n \Psi_i(\Omega) = 0$ is an unbiased estimating equation for Ω . This result leads to the consistency of $\hat{\Omega}$.

Let $U_n(\Omega) \equiv n^{-1} \sum_{i=1}^n \Psi_i(\Omega)$. The asymptotic distribution of $\sqrt{n}(\hat{\Omega} - \Omega_0)$ is derived as follows. A first order Taylor series expansion of $U_n(\Omega)$ can be applied to obtain that

$$\sqrt{n}(\hat{\Omega} - \Omega_0) = \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \Omega} \Psi_i(\Omega_0) \right]^{-1} \sqrt{n} U_n(\Omega_0) + o_p(1).$$

Note that $n^{-1} \sum_{i=1}^n \partial \Psi_i(\Omega_0) / \partial \beta$ converges to $\mathcal{A}(\Omega_0) = E \{ \partial \Psi_i(\Omega_0) / \partial \Omega \}$, which is positive definite under assumptions (C4)-(C6). Furthermore, $\sqrt{n} U_n(\Omega_0)$ is a normalized sum of independent and identically distributed random variables with mean zero. The application of the central limit theorem under assumptions (C4)-(C6) leads to the fact that $\sqrt{n} U_n(\Omega_0)$ is normally distributed with zero and variance $\mathcal{B}(\Omega_0)$, which is the limit of $n^{-1} \sum_{i=1}^n \Psi_i(\Omega_0) \Psi_i(\Omega_0)'$. It follows that $\sqrt{n}(\hat{\Omega} - \Omega_0)$ is asymptotically normally distributed with mean zero and covariance matrix $\mathcal{A}^{-1}(\Omega_0) \mathcal{B}(\Omega_0) \{ \mathcal{A}^{-1}(\Omega_0) \}'$.

Appendix B: Estimation of the nuisance parameters

Assuming that the primary regression model is linear with a single error-prone covariate X , the response variable Y and X are linked through the model $Y = \beta_0 + \beta_1 X + e$, where e has mean 0 and variance σ_e^2 . The unknown parameters are $\mu_l, \alpha_0, \alpha_1, \gamma_0, \gamma_1, \beta_0, \beta_1, \sigma_l^2, \sigma_b^2, \sigma_c^2, \sigma_v^2, \sigma_e^2$ and σ_ϵ^2 . Based on moment calculations, unbiased estimating equations for all the parameters are given in the following where $\bar{T} = \sum_{i=1}^n \eta_i T_i / \sum_{i=1}^n \eta_i$ for any variable T .

$$\sum_{i=1}^n \eta_i (W_i - \mu_l) = 0$$

$$\sum_{i=1}^n \eta_i (Y_i - \beta_0 - \beta_1 \mu_l) = 0$$

$$\sum_{i=1}^n \eta_i (M_i - \alpha_0 - \alpha_1 \mu_i) = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (M_i - \bar{M}) (Y_i - \bar{Y}) - \beta_1 (W_i - \bar{W}) (M_i - \bar{M}) \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (W_i - \bar{W}) (Y_i - \bar{Y}) - \beta_1 \sigma_l^2 \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (W_i - \bar{W}) (M_i - \bar{M}) - \alpha_1 \sigma_l^2 \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (W_i - \bar{W})^2 - \sigma_l^2 - \sigma_c^2 \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (Y_i - \bar{Y})^2 - \beta_1^2 (\sigma_l^2 + \sigma_b^2) - \sigma_e^2 \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (M_i - \bar{M})^2 - \alpha_1^2 \sigma_l^2 - \sigma_v^2 \right\} = 0$$

$$\sum_{i=1}^n \eta_i (Q_i - \gamma_0 - \gamma_1 \mu_i) = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (M_i - \bar{M}) (Q_i - \bar{Q}) - \gamma_1 (W_i - \bar{W}) (M_i - \bar{M}) \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ (Y_i - \bar{Y}) (Q_i - \bar{Q}) - \gamma_1 \beta_1 (\sigma_l^2 + \sigma_b^2) \right\} = 0$$

$$\sum_{i=1}^n \eta_i \left\{ \left(Q_i - \bar{Q} \right)^2 - \gamma_1^2 \left(\sigma_l^2 + \sigma_b^2 \right) - \sigma_e^2 \right\} = 0,$$

References

1. Hammer SM, Katzenstein DA, Huges MD, Gundacker H, Schooley RT, Haubrich MR, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine*. 1996; 335:1081–1090. [PubMed: 8813038]
2. Reeves GK, Cox DR, Darby SC, Whitley E. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*. 1998; 17:2157–2177. DOI: 10.1002/(SICI)1097-0258(19981015)17:19<2157::AID-SIM 916>3.0.CO;2-F. [PubMed: 9802176]
3. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, CM. *Nonlinear Measurement Error Models, A modern Perspective*. second edition. Chapman and Hall; London: 2006.
4. Schafer DW, Gilbert ES. Some Statistical implications of dose uncertainty in radiation dose response analyses. *Radiation Research*. 2006; 166:303–312. [PubMed: 16808615]
5. Prentice RL. Covariate measurement errors and parameter estimates in failure time regression. *Biometrika*. 1982; 69:331–342.
6. Batistatou E, McNamee R. Performance of bias-correction methods for exposure measurement error using repeated measurements with and without missing data. *Statistics in Medicine*. 2012; 31:3467–3480. DOI: 10.1002/sim.5422. [PubMed: 22733598]
7. Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. *Biometrics*. 1997; 53:131–145. [PubMed: 9147589]
8. Huang Y, Wang CY. Cox regression with accurate covariates unascertainable: a non-parametric correction approach. *Journal of the American Statistical Association*. 2000; 95:1209–1219.
9. Wang CY. Corrected score estimator for joint modeling of longitudinal and failure time data. *Statistica Sinica*. 2006; 16:235–253.
10. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*. 1995; 89:1314–1328.
11. Whitmore AS, Keller JB. Approximations for regression with covariate measurement error. *Journal of the American Statistical Association*. 1988; 83:1057–1066.
12. Wang L. Estimation of nonlinear models with Berkson measurement errors. *The Annals of Statistics*. 2004; 32:2559–2579. DOI: 10.1214/009053604000000670.
13. Fuller, WA. *Measurement Error Models*. John Wiley & Sons; New York: 1987.
14. Guolo A. Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*. 2008; 17:555–580. DOI:10.1177/0962280207081318. [PubMed: 18375458]
15. Mallick B, Hoffman FO, Carroll RJ. Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada test site. *Biometrics*. 2002; 58:13–20. [PubMed: 11890308]
16. Li Y, Guolo A, Hoffman FO, Carroll RJ. Shared uncertainty in measurement error problems, with application to Nevada Test Site fallout data. *Biometrics*. 2007; 63:1226–1236. DOI: 10.1111/j.1541-0420.2007.00810.x. [PubMed: 18078484]
17. Kukush A, Shklyar S, Masiuk S, Likhtarov I, Kovgan L, Carroll RJ, Bouville A. Method for estimation of radiation risk in epidemiological studies accounting for classical and Berkson errors in doses. *The International Journal of Biostatistics*. 2011; 7(1):15. DOI: 10.2202/1557-4679.1281. [PubMed: 21423564]
18. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF, the rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases*. 2005; 191:654–665. [PubMed: 15688278]

19. Carroll RJ, Delaigle A, Hall P. Non-parametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *Journal of the Royal Statistical Society, Series B*. 2007; 69(5):859–878. DOI: 10.1111/j.1467-9868.2007.00614.x.
20. Apanasovich TV, Carroll RJ, Maity A. SIMEX and standard error estimation in semi-parametric measurement error models. *Electronic Journal of Statistics*. 2009; 3:318–348. DOI: 10.1214/08-EJS341. [PubMed: 19609371]
21. Wang CY. Robust best linear estimation for regression analysis using surrogate and instrumental variables. *Biostatistics*. 2012; 13(2):326–340. DOI:10.1093/biostatistics/kxr051. [PubMed: 22285992]
22. Huang Y, Wang CY. Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*. 2001; 96:1469–1482.
23. Kuha J. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine*. 1994; 13(11):1135–1148. DOI: 10.1002/sim.4780131105. [PubMed: 8091040]
24. Stefanski LA, Cook RJ. Simulation-extrapolation: the measurement error Jackknife. *Journal of the American Statistical Association*. 1995; 90:1247–1256.
25. Wang CY, Huang Y. Error in timing regression with observed longitudinal measurements. *Statistics in Medicine*. 2003; 22:2577–2590. DOI: 10.1002/sim.1435. [PubMed: 12898545]
26. Wang CY, Huang Y, Chao EC, Jeffcoat MK. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics*. 2008; 64:85–95. DOI: 10.1111/j.1541-0420.2007.00839.x. [PubMed: 17608787]
27. Fenton KA, Johnson AM, McManus S, Erens B. Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections*. 2001; 77(2):84–92. DOI: 10.1136/sti.77.2.84. [PubMed: 11287683]
28. McKay T, Mutchler MG. The effect of partner sex: nondisclosure of HIV status to male and female partners among men who have sex with men and women (MSMW). *AIDS and Behavior*. 2011; 15(6):1140–1152. DOI: 10.1007/s10461-010-9851-4. [PubMed: 21082339]
29. Serovich JM. A test of two HIV disclosure theories among men who have sex with men and women (MSMW). *AIDS Education and Prevention*. 2001; 13(4):355–364. DOI: 10.1521/aeap.13.4.355.21424. [PubMed: 11565594]

Table 1

Simulation results for linear regression: $Y = \beta_0 + \beta_1 X + e$, where $X = L + U_b$; $W = L + U_c$; $\beta = (-2, 2)'$; $M = 1 - 2L + V$; $Q = -1 + 2X + \epsilon$; θ is the proportion of the calibration data; $L \sim N(0.5, 1)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $U_c \sim N(0, \sigma_c^2)$ is the classical error; $V \sim N(0, 1)$; $e \sim N(0, 1)$; $n = 500$; Naive, "naive" regression replacing X by W ; RC1, Regression calibration approach replacing X by its conditional expectation given W ; RC2, Regression calibration approach replacing X by its conditional expectation given W and Q , M in the calibration sample; SIMEX, simulation extrapolation procedure; EEE, Expected estimating equation method.

θ	σ_b^2	β	$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$					
			Naive	RC1	RC2	SIMEX	EEE	Naive	RC1	RC2	SIMEX	EEE	
0.5	0.2	β_0	Bias	0.237	0.006	0.006	0.009	0.003	0.330	-0.001	-0.002	0.010	-0.007
			SD	0.082	0.104	0.104	0.094	0.107	0.083	0.112	0.110	0.096	0.111
		ASE		0.080	0.099	0.098	0.091	0.099	0.085	0.117	0.115	0.102	0.116
			CP	0.182	0.948	0.950	0.945	0.950	0.025	0.958	0.960	0.962	0.960
	β_1	Bias	-0.464	0.003	0.000	-0.007	0.005	-0.666	-0.003	-0.003	-0.028	0.006	
		SD	0.061	0.114	0.110	0.096	0.108	0.064	0.137	0.131	0.102	0.122	
		ASE	0.065	0.112	0.108	0.097	0.105	0.065	0.137	0.130	0.108	0.122	
		CP	0.000	0.965	0.955	0.950	0.942	0.000	0.960	0.948	0.948	0.958	
	0.4	β_0	Bias	0.236	0.002	0.002	0.001	-0.001	0.340	-0.001	0.000	-0.002	-0.004
			SD	0.089	0.109	0.108	0.106	0.114	0.093	0.129	0.126	0.116	0.132
		ASE		0.092	0.110	0.109	0.106	0.114	0.096	0.128	0.126	0.116	0.129
			CP	0.292	0.960	0.958	0.940	0.948	0.052	0.952	0.950	0.938	0.960
0.7	β_1	Bias	-0.466	0.001	0.001	0.003	0.008	-0.672	0.007	0.006	0.008	0.013	
		SD	0.072	0.126	0.122	0.118	0.126	0.072	0.154	0.145	0.130	0.142	
		ASE	0.073	0.125	0.121	0.118	0.123	0.072	0.151	0.142	0.129	0.138	
		CP	0.000	0.955	0.948	0.952	0.955	0.000	0.955	0.948	0.952	0.955	
	0.2	β_0	Bias	0.241	0.010	0.010	0.014	0.010	0.334	0.001	0.002	0.014	0.003
			SD	0.079	0.092	0.092	0.089	0.094	0.085	0.108	0.107	0.102	0.109
		ASE		0.081	0.093	0.092	0.089	0.094	0.085	0.108	0.106	0.100	0.108
			CP	0.160	0.962	0.960	0.938	0.970	0.030	0.962	0.962	0.948	0.960
	β_1	Bias	-0.463	0.005	0.001	-0.007	0.003	-0.662	0.004	0.003	-0.023	0.004	
		SD	0.065	0.098	0.097	0.090	0.097	0.067	0.119	0.115	0.101	0.112	

θ	σ_b^2	β	$\sigma_c^2 = 0.3$						$\sigma_c^2 = 0.5$					
			Naïve	RC1	RC2	SIMEX	EEE	Naïve	RC1	RC2	SIMEX	EEE	Naïve	RC1
0.4	β_0	ASE	0.064	0.098	0.095	0.090	0.095	0.064	0.116	0.111	0.099	0.108	0.064	0.116
		CP	0.000	0.970	0.955	0.942	0.952	0.000	0.955	0.945	0.922	0.952	0.000	0.955
		Bias	0.236	0.001	0.002	0.000	0.003	0.340	0.000	0.002	-0.003	0.004	0.340	0.000
		SD	0.089	0.102	0.101	0.100	0.103	0.093	0.116	0.114	0.111	0.116	0.093	0.116
		ASE	0.092	0.105	0.104	0.102	0.107	0.096	0.118	0.117	0.113	0.120	0.096	0.118
	β_1	CP	0.355	0.968	0.958	0.945	0.960	0.052	0.962	0.960	0.945	0.952	0.052	0.962
		Bias	-0.466	0.004	0.001	0.004	0.002	-0.672	0.007	0.003	0.009	0.003	-0.672	0.007
		SD	0.072	0.106	0.102	0.103	0.103	0.072	0.125	0.117	0.114	0.115	0.072	0.125
		ASE	0.073	0.110	0.107	0.106	0.108	0.072	0.128	0.122	0.118	0.121	0.072	0.128
		CP	0.000	0.960	0.960	0.958	0.965	0.000	0.955	0.962	0.952	0.965	0.000	0.955

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; CP represents the coverage probability of the 95% confidence intervals.

Table 2

Simulation results for the estimation of the nuisance parameters in the linear regression model: $Y = \beta_0 + \beta_1 X + e$, where $X = L + U_b$; $W = L + U_c$; $0 = (-2, 2)'$; $M = \alpha_0 + \alpha_1 L + V$; $Q = \gamma_0 + \gamma_1 X + \epsilon$; θ is the proportion of the calibration data; $L \sim N(\mu_1, \sigma_l^2)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $U_c \sim N(0, \sigma_c^2)$ is the classical error; $V \sim N(0, \sigma_v^2)$; $e \sim N(0, \sigma_e^2)$; $\epsilon \sim N(0, \sigma_\epsilon^2)$.

V	$\theta = 0.5$				$\theta = 0.7$			
	Bias	SD	ASE	CP	Bias	SD	ASE	CP
$\sigma_b^2 = 0.2, \sigma_c^2 = 0.3$								
μ_1	-0.001	0.072	0.072	0.955	0.004	0.065	0.061	0.925
α_0	-0.010	0.106	0.110	0.962	-0.005	0.093	0.092	0.952
α_1	-0.005	0.119	0.113	0.942	-0.004	0.099	0.096	0.958
γ_0	0.008	0.130	0.125	0.932	0.011	0.105	0.106	0.960
γ_1	0.006	0.129	0.122	0.935	0.001	0.111	0.103	0.955
σ_l^2	-0.004	0.115	0.117	0.942	-0.002	0.102	0.099	0.940
σ_b^2	-0.002	0.056	0.057	0.955	0.000	0.051	0.048	0.940
σ_c^2	-0.004	0.045	0.044	0.948	-0.002	0.038	0.038	0.950
σ_e^2	-0.007	0.171	0.175	0.962	0.001	0.151	0.149	0.942
σ_v^2	-0.009	0.168	0.168	0.955	-0.009	0.141	0.141	0.962
σ_ϵ^2	-0.012	0.177	0.175	0.950	-0.010	0.156	0.147	0.932
$\sigma_b^2 = 0.4, \sigma_c^2 = 0.3$								
μ_1	0.003	0.074	0.072	0.962	0.001	0.058	0.061	0.960
α_0	-0.003	0.122	0.112	0.935	-0.006	0.097	0.095	0.948
α_1	-0.003	0.128	0.122	0.935	-0.009	0.108	0.102	0.945
γ_0	-0.003	0.146	0.142	0.942	0.012	0.116	0.119	0.955
γ_1	0.010	0.146	0.139	0.945	-0.004	0.109	0.115	0.965
σ_l^2	-0.007	0.126	0.119	0.930	-0.002	0.104	0.101	0.938
σ_b^2	0.002	0.096	0.090	0.942	0.003	0.073	0.076	0.942
σ_c^2	-0.003	0.052	0.049	0.938	0.002	0.043	0.042	0.942
σ_e^2	-0.009	0.213	0.213	0.948	-0.018	0.178	0.179	0.945
σ_v^2	0.003	0.189	0.188	0.955	-0.024	0.163	0.158	0.945
σ_ϵ^2	0.000	0.218	0.215	0.958	0.005	0.174	0.179	0.968
$\sigma_b^2 = 0.4, \sigma_c^2 = 0.5$								

<i>V</i>	$\theta = 0.5$				$\theta = 0.7$			
	Bias	SD	ASE	CP	Bias	SD	ASE	CP
μ_1	0.003	0.080	0.077	0.955	0.000	0.063	0.065	0.962
α_0	0.000	0.142	0.132	0.938	-0.006	0.112	0.111	0.942
α_1	-0.007	0.149	0.143	0.942	-0.013	0.128	0.120	0.942
γ_0	-0.005	0.162	0.157	0.948	0.012	0.129	0.131	0.950
γ_1	0.015	0.162	0.153	0.935	-0.003	0.121	0.128	0.965
σ_l^2	-0.007	0.141	0.134	0.930	-0.003	0.117	0.114	0.942
σ_b^2	0.001	0.105	0.099	0.945	0.004	0.080	0.083	0.942
σ_c^2	-0.004	0.069	0.066	0.955	0.003	0.058	0.056	0.950
σ_e^2	-0.009	0.213	0.213	0.948	-0.018	0.178	0.179	0.945
σ_v^2	0.003	0.215	0.216	0.958	-0.028	0.188	0.183	0.948
σ_ϵ^2	0.000	0.218	0.215	0.958	0.005	0.174	0.179	0.968

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; CP represents the coverage probability of the 95% confidence intervals.

Table 3

Simulation results for Logistic regression: $E(Y|X) = \{1 + \exp(-\beta_0 - \beta_1 X)\}^{-1}$, where $X = L + U_b$; $W = L + U_c$; $\beta = (-1, \ln(5))'$; $M = 1 - 2L + V$; $Q = -1 + 2X + \epsilon$; θ is the proportion of the calibration data; $L \sim N(0, 0.5, 1)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $U_c \sim N(0, \sigma_c^2)$ is the classical error; $V \sim N(0, 1)$; $\epsilon \sim N(0, 1)$; $n = 500$; Naive, "naive" regression replacing X by W ; RC1, Regression calibration approach replacing X by its conditional expectation given W ; RC2, Regression calibration approach replacing X by its conditional expectation given W and Q , M in the calibration sample; SIMEX, simulation extrapolation procedure; EEE, Expected estimating equation method.

θ	σ_b^2	β	$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$				
			Naive	RC1	RC2	SIMEX	EEE	Naive	RC1	RC2	SIMEX	EEE
0.5	β_0	Bias	0.318	0.162	0.105	0.095	-0.015	0.417	0.202	0.129	0.188	-0.004
		SD	0.118	0.135	0.148	0.155	0.194	0.112	0.138	0.152	0.146	0.201
		ASE	0.119	0.133	0.145	0.156	0.188	0.114	0.135	0.150	0.145	0.199
		CP	0.276	0.760	0.889	0.889	0.959	0.066	0.660	0.827	0.734	0.949
	β_1	Bias	-0.575	-0.264	-0.169	-0.177	0.020	-0.741	-0.312	-0.191	-0.327	0.021
		SD	0.112	0.151	0.170	0.173	0.244	0.098	0.157	0.181	0.156	0.260
		ASE	0.111	0.156	0.174	0.179	0.249	0.099	0.159	0.184	0.153	0.266
		CP	0.003	0.613	0.825	0.794	0.966	0.000	0.503	0.763	0.418	0.963
	0.4	Bias	0.363	0.214	0.137	0.109	-0.021	0.449	0.240	0.147	0.209	-0.024
		SD	0.114	0.128	0.143	0.163	0.205	0.110	0.132	0.152	0.149	0.232
		ASE	0.117	0.130	0.145	0.162	0.205	0.112	0.133	0.151	0.148	0.218
		CP	0.156	0.622	0.833	0.870	0.966	0.031	0.536	0.830	0.688	0.969
0.7	β_1	Bias	-0.640	-0.337	-0.210	-0.192	0.042	-0.795	-0.378	-0.222	-0.369	0.049
		SD	0.105	0.147	0.172	0.193	0.281	0.094	0.146	0.175	0.155	0.301
		ASE	0.108	0.152	0.175	0.189	0.278	0.096	0.156	0.186	0.161	0.298
		CP	0.000	0.367	0.742	0.779	0.979	0.000	0.327	0.750	0.410	0.977
	β_0	Bias	0.315	0.159	0.081	0.089	-0.018	0.414	0.194	0.094	0.187	-0.012
		SD	0.115	0.125	0.141	0.146	0.175	0.108	0.125	0.143	0.133	0.179
		ASE	0.119	0.129	0.144	0.151	0.175	0.114	0.128	0.148	0.141	0.183
		CP	0.249	0.772	0.906	0.891	0.957	0.067	0.646	0.921	0.726	0.956
	β_1	Bias	-0.564	-0.254	-0.126	-0.160	0.032	-0.750	-0.315	-0.151	-0.340	0.018
		SD	0.110	0.146	0.172	0.169	0.230	0.095	0.141	0.170	0.143	0.228

θ	σ_b^2	β	$\sigma_\epsilon^2 = 0.3$					$\sigma_\epsilon^2 = 0.5$				
			Naïve	RC1	RC2	SIMEX	EEE	Naïve	RC1	RC2	SIMEX	EEE
0.4	β_0	ASE	0.112	0.143	0.166	0.168	0.221	0.098	0.140	0.172	0.144	0.231
		CP	0.008	0.551	0.858	0.807	0.962	0.000	0.403	0.841	0.341	0.956
		Bias	0.370	0.221	0.114	0.119	-0.010	0.449	0.245	0.119	0.210	-0.006
		SD	0.113	0.128	0.147	0.151	0.194	0.117	0.131	0.154	0.154	0.200
		ASE	0.117	0.126	0.146	0.156	0.190	0.112	0.126	0.150	0.146	0.196
		CP	0.113	0.562	0.876	0.869	0.956	0.049	0.491	0.859	0.678	0.962
	β_1	Bias	-0.634	-0.340	-0.162	-0.192	0.038	-0.805	-0.398	-0.192	-0.380	0.005
		SD	0.105	0.138	0.166	0.170	0.243	0.097	0.137	0.171	0.154	0.247
		ASE	0.109	0.139	0.168	0.175	0.245	0.096	0.138	0.174	0.151	0.253
		CP	0.000	0.340	0.830	0.768	0.956	0.000	0.207	0.785	0.302	0.969

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; CP represents the coverage probability of the 95% confidence intervals.

Table 4

Simulation results for Poisson regression: $E(Y|X) = \exp(\beta_0 + \beta_1 X)$, where $X = L + U_b$; $W = L + U_c$; $\beta = (0.5, \ln(2))'$; $M = 1 - 2L + V$; $Q = -1 + 2X + \epsilon$; θ is the proportion of the calibration data; $L \sim N(0.5, 1)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $U_c \sim N(0, \sigma_c^2)$ is the classical error; $V \sim N(0, 1)$; $\epsilon \sim N(0, 1)$; $n = 500$; Naive, "naive" regression replacing X by W ; RC1, Regression calibration approach replacing X by its conditional expectation given W ; RC2, Regression calibration approach replacing X by its conditional expectation given W and Q ; M in the calibration sample; SIMEX, simulation extrapolation procedure; EEE, Expected estimating equation method.

θ	σ_b^2	β	$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$				
			Naive	RC1	RC2	SIMEX	EEE	Naive	RC1	RC2	SIMEX	EEE
0.5	β_0	Bias	0.183	0.103	0.075	0.010	0.002	0.242	0.124	0.086	0.022	-0.001
		SD	0.048	0.054	0.053	0.058	0.056	0.046	0.053	0.053	0.058	0.053
		ASE	0.046	0.052	0.051	0.055	0.053	0.046	0.057	0.056	0.059	0.057
		CP	0.048	0.481	0.670	0.927	0.935	0.005	0.420	0.686	0.917	0.962
	β_1	Bias	-0.159	0.000	-0.006	-0.012	-0.001	-0.230	0.007	-0.004	-0.030	0.004
		SD	0.037	0.057	0.051	0.045	0.048	0.035	0.066	0.056	0.047	0.051
		ASE	0.033	0.055	0.049	0.043	0.045	0.032	0.065	0.056	0.046	0.051
		CP	0.023	0.942	0.929	0.917	0.932	0.003	0.957	0.940	0.877	0.942
	0.4	Bias	0.230	0.149	0.103	0.003	0.001	0.289	0.172	0.120	0.005	-0.001
		SD	0.048	0.054	0.054	0.055	0.054	0.051	0.063	0.061	0.061	0.061
		ASE	0.049	0.056	0.055	0.057	0.057	0.050	0.060	0.059	0.061	0.061
		CP	0.003	0.262	0.524	0.929	0.952	0.000	0.234	0.453	0.935	0.950
0.7	β_1	Bias	-0.161	-0.001	-0.010	-0.006	-0.001	-0.231	0.001	-0.015	-0.013	0.000
		SD	0.039	0.061	0.055	0.047	0.049	0.038	0.074	0.061	0.049	0.054
		ASE	0.037	0.061	0.054	0.048	0.050	0.036	0.070	0.059	0.050	0.056
		CP	0.038	0.942	0.935	0.935	0.957	0.003	0.952	0.919	0.922	0.950
	β_0	Bias	0.181	0.100	0.054	0.004	-0.002	0.242	0.127	0.069	0.021	0.003
		SD	0.046	0.050	0.050	0.054	0.052	0.044	0.053	0.052	0.057	0.054
		ASE	0.046	0.051	0.050	0.053	0.052	0.046	0.053	0.053	0.057	0.055
		CP	0.040	0.498	0.802	0.942	0.942	0.003	0.323	0.762	0.942	0.960
	β_1	Bias	-0.160	0.002	-0.002	-0.010	0.001	-0.230	0.001	-0.007	-0.030	-0.001
		SD	0.035	0.054	0.047	0.043	0.044	0.032	0.055	0.046	0.042	0.044

θ	σ_b^2	β	$\sigma_e^2 = 0.3$					$\sigma_e^2 = 0.5$				
			Naïve	RC1	RC2	SIMEX	EEE	Naïve	RC1	RC2	SIMEX	EEE
0.4	β_0	ASE	0.034	0.051	0.044	0.042	0.042	0.032	0.057	0.048	0.043	0.046
		CP	0.018	0.928	0.938	0.940	0.960	0.000	0.960	0.960	0.887	0.943
		Bias	0.226	0.148	0.078	-0.002	-0.002	0.293	0.177	0.093	0.006	0.003
		SD	0.049	0.056	0.056	0.060	0.059	0.051	0.061	0.058	0.058	0.059
		ASE	0.050	0.054	0.053	0.056	0.055	0.049	0.057	0.056	0.060	0.058
	β_1	CP	0.018	0.235	0.678	0.927	0.939	0.000	0.175	0.603	0.939	0.942
		Bias	-0.158	-0.002	-0.007	-0.007	0.001	-0.234	-0.001	-0.012	-0.012	-0.002
		SD	0.038	0.056	0.048	0.046	0.045	0.038	0.066	0.054	0.048	0.050
		ASE	0.037	0.054	0.048	0.045	0.045	0.036	0.061	0.052	0.048	0.050
		CP	0.041	0.937	0.942	0.934	0.947	0.000	0.934	0.929	0.924	0.939

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; CP represents the coverage probability of the 95% confidence intervals.

Table 5

Misspecification of the measurement error model in logistic regression: $E(Y|X) = \{1 + \exp(-\beta_0 - \beta_1 X)\}^{-1}$, where $X = L + U_b$; $W = L + U_c$; $\beta = (-1, \ln(5))'$; $M = 1 - 2L + V$; $Q = -1 + 2X + \epsilon$; $\theta = 0.7$; $L \sim N(0.5, 1)$; $U_b \sim N(0, \sigma_b^2)$; $U_c \sim N(0, \sigma_c^2)$ is the Berkson error; $V \sim N(0, 1)$; $\epsilon \sim N(0, 1)$; $n = 500$; Naive, "naive" regression replacing X by W ; RC1, Regression calibration replacing X by its conditional expectation given W ; RC2, Regression calibration replacing X by its conditional expectation given W and Q ; M in the calibration sample; SIMEX, simulation extrapolation procedure; EEE, Expected estimating equation method.

σ_b^2	β	$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$					
		Naive	RC1	RC2	SIMEX	EEE	Naive	RC1	RC2	SIMEX	EEE	
0.2	β_0	Bias	0.315	0.159	0.068	0.148	0.009	0.414	0.194	0.078	0.225	0.005
		SD	0.115	0.125	0.140	0.133	0.158	0.108	0.125	0.143	0.126	0.165
		ASE	0.119	0.129	0.143	0.137	0.161	0.114	0.128	0.148	0.133	0.171
		CP	0.249	0.772	0.916	0.802	0.949	0.067	0.646	0.931	0.582	0.959
	β_1	Bias	-0.564	-0.254	-0.096	-0.251	-0.001	-0.750	-0.315	-0.117	-0.399	0.001
		SD	0.110	0.146	0.172	0.152	0.206	0.095	0.141	0.170	0.135	0.208
		ASE	0.112	0.143	0.165	0.151	0.197	0.098	0.140	0.173	0.135	0.212
		CP	0.008	0.551	0.898	0.581	0.957	0.000	0.403	0.874	0.200	0.946
	β_0	Bias	0.370	0.221	0.096	0.217	0.039	0.449	0.245	0.119	0.210	-0.006
		SD	0.113	0.128	0.143	0.131	0.161	0.117	0.131	0.154	0.154	0.200
ASE		0.117	0.126	0.144	0.133	0.162	0.112	0.126	0.150	0.146	0.196	
CP		0.113	0.562	0.902	0.613	0.938	0.049	0.491	0.859	0.678	0.962	
β_1	Bias	-0.634	-0.340	-0.119	-0.346	-0.026	-0.805	-0.398	-0.139	-0.481	-0.022	
	SD	0.105	0.138	0.162	0.139	0.194	0.097	0.137	0.169	0.133	0.207	
	ASE	0.109	0.139	0.166	0.144	0.200	0.096	0.138	0.176	0.132	0.219	
	CP	0.000	0.340	0.884	0.358	0.946	0.000	0.207	0.870	0.105	0.957	
Incorrectly assuming a Berkson error model												
0.2	β_0	Bias	0.315	0.315	0.200	0.282	0.161	0.414	0.278	0.395	0.248	
		SD	0.115	0.115	0.127	0.122	0.139	0.108	0.108	0.119	0.112	0.129
		ASE	0.119	0.119	0.130	0.126	0.142	0.114	0.114	0.129	0.117	0.138
		CP	0.249	0.249	0.670	0.398	0.777	0.067	0.067	0.415	0.113	0.574

σ_b^2	β	$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$				
		Naïve	RC1	RC2	SIMEX	EEE	Naïve	RC1	RC2	SIMEX	EEE
β_1	Bias	-0.564	-0.564	-0.355	-0.514	-0.293	-0.750	-0.750	-0.502	-0.722	-0.455
	SD	0.110	0.110	0.132	0.118	0.152	0.095	0.095	0.122	0.099	0.140
	ASE	0.112	0.112	0.131	0.120	0.151	0.098	0.098	0.127	0.101	0.145
	CP	0.008	0.008	0.259	0.030	0.497	0.000	0.000	0.041	0.000	0.159
0.4 β_0	Bias	0.370	0.370	0.208	0.315	0.136	0.449	0.449	0.256	0.415	0.195
	SD	0.113	0.113	0.128	0.124	0.146	0.117	0.117	0.136	0.127	0.153
	ASE	0.117	0.117	0.133	0.129	0.152	0.112	0.112	0.131	0.120	0.147
	CP	0.113	0.113	0.642	0.325	0.853	0.049	0.049	0.501	0.118	0.673
β_1	Bias	-0.634	-0.634	-0.341	-0.548	-0.225	-0.805	-0.805	-0.455	-0.755	-0.357
	SD	0.105	0.105	0.127	0.122	0.155	0.097	0.097	0.126	0.109	0.153
	ASE	0.109	0.109	0.133	0.126	0.165	0.096	0.096	0.125	0.105	0.153
	CP	0.000	0.000	0.289	0.018	0.711	0.000	0.000	0.084	0.000	0.353

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; CP represents the coverage probability of the 95% confidence intervals.

Table 6

Results of the analysis from fitting logistic regression to the VAX004 data when the outcome is occurrence of HIV infection; Naive, “naive” regression; RC1, Regression calibration approach not using the data in the calibration sample; RC2, Regression calibration using the data in the calibration sample; EEE, Expected estimating equation method; SIMEX, simulation extrapolation procedure.

β	β_0			β_1			β_2		
	Est	SE	<i>p</i> -value	Est	SE	<i>p</i> -value	Est	SE	<i>p</i> -value
<i>Mixture of classical and Berkson errors</i>									
Naive	-2.8357	0.1024	< 0.0001	0.5704	0.0822	< 0.0001	-0.0608	0.1145	0.5954
RC1	-3.0674	0.1237	< 0.0001	1.1725	0.2086	< 0.0001	-0.0608	0.1145	0.5954
RC2	-3.2202	0.1217	< 0.0001	1.4256	0.1690	< 0.0001	-0.0557	0.1150	0.6285
SIMEX	-2.9235	0.1021	< 0.0001	0.7640	0.0921	< 0.0001	-0.0629	0.1244	0.6132
EEE	-3.2360	0.1301	< 0.0001	1.4282	0.1743	< 0.0001	-0.0946	0.1176	0.4213
<i>Classical error only</i>									
Naive	-2.8357	0.1024	< 0.0001	0.5704	0.0822	< 0.0001	-0.0608	0.1145	0.5954
RC1	-3.0674	0.1237	< 0.0001	1.1725	0.2086	< 0.0001	-0.0608	0.1145	0.5954
RC2	-3.2200	0.1189	< 0.0001	1.4411	0.1583	< 0.0001	-0.0544	0.1150	0.6364
SIMEX	-2.9169	0.1049	< 0.0001	0.7634	0.0859	< 0.0001	-0.0633	0.1240	0.6099
EEE	-3.2331	0.1256	< 0.0001	1.4442	0.1607	< 0.0001	-0.0839	0.1186	0.4792
<i>Nuisance parameter</i>									
ν	Est	SE	<i>p</i> -value						
μ_1	0.3849	0.0074	< 0.0001						
α_0	-0.2839	0.1367	0.0378						
α_1	1.8873	0.3591	< 0.0001						
γ_0	0.5347	0.0528	< 0.0001						
γ_1	2.8325	0.1144	< 0.0001						
σ_l^2	0.1364	0.0264	< 0.0001						
σ_b^2	0.0267	0.0294	0.3632						
σ_c^2	0.1439	0.0262	< 0.0001						
σ_v^2	0.5752	0.0940	< 0.0001						
σ_ϵ^2	0.8827	0.2113	< 0.0001						

$\beta = (\beta_0, \beta_1, \beta_2)'$ is the vector of primary parameters; β_0 is the intercept, β_1 is the coefficient of the log-transformed number of HIV positive male partners and β_2 is the coefficient of the indicator variable for vaccine; ν is the vector of nuisance parameters; Est denotes estimate; SE means standard error and *p*-value is the Wald-test-based *p*-value.