Efficient semiparametric inference for two-phase studies with outcome and covariate measurement errors

Ran Tao[1,2], **Sarah C. Lotspeich**[1], **Gustavo Amorim**[1], **Pamela A. Shaw**[3], **Bryan E. Shepherd**[1]

[1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee

[2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee

[3]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

Abstract

In modern observational studies using electronic health records or other routinely collected data, both the outcome and covariates of interest can be error-prone and their errors often correlated. A cost-effective solution is the two-phase design, under which the error-prone outcome and covariates are observed for all subjects during the first phase and that information is used to select a validation subsample for accurate measurements of these variables in the second phase. Previous research on two-phase measurement error problems largely focused on scenarios where there are errors in covariates only or the validation sample is a simple random sample of study subjects. Herein, we propose a semiparametric approach to general two-phase measurement error problems with a quantitative outcome, allowing for correlated errors in the outcome and covariates and arbitrary second-phase selection. We devise a computationally efficient and numerically stable expectation-maximization algorithm to maximize the nonparametric likelihood function. The resulting estimators possess desired statistical properties. We demonstrate the superiority of the proposed methods over existing approaches through extensive simulation studies, and we illustrate their use in an observational HIV study.

Keywords

data audits; electronic health records; EM algorithm; HIV/AIDS; missing data; sieve approximation

1 | INTRODUCTION

In modern observational studies using electronic health records or other routinely collected data, a multitude of variables are collected on a large number of subjects. These databases

---

generate abundant opportunities for researchers to investigate associations of scientific and clinical interest. Due to the observational or retrospective nature of their data collection, however, the outcome and covariates of interest in these databases are often error-prone. Common errors include misclassification and/or incorrect dates of clinical diagnoses, incorrectly recorded measurements potentially due to data entry errors or wrong units, incorrect types and/or dates of medications, and missing pertinent information in free text notes by only extracting based on structured data (eg, insurance billing codes). In addition, errors in the outcome and covariates are frequently correlated, particularly when analysis variables are derived from other variables in the electronic health record. For example, if the date of treatment initiation are incorrectly recorded, then lab values at the time of treatment initiation are also likely to be incorrect. It is important to check the validity of these data records before incorporating them in analysis so as to avoid biased and misleading results.[1]

Data audits, which are commonly used in clinical trials,[2] have also been implemented in observational studies to ensure data quality.[3–6] A data audit typically involves a group of external auditors comparing the data in the research database to those in the patients' clinical charts and reporting any discrepancies between the two data sources. An audit of the entire database is generally prohibitively time-consuming and expensive for large databases. A cost-effective solution often applied in practice is the two-phase design, under which the error-prone outcome and covariates are obtained for all subjects during the first phase (eg, extraction of data from electronic health records) and a subsample of these error-prone variables are subsequently audited in the second phase. This type of design greatly reduces the cost associated with data validation and thus has been used in several large-scale studies, including the Caribbean, Central, and South America Network for HIV Research (CCASAnet).[7]

There is extensive research on analyzing data from two-phase studies with measurement errors in covariates only.[8–10] Measurement error in a quantitative outcome is generally ignored in regression analysis because it can be absorbed into the residual error provided that it is homoscedastic.[11] This conclusion no longer holds when both the outcome and covariates are error-prone and their errors are correlated.[12] To deal with this general measurement error problem, Chen and Chen[13] proposed a "unified approach based on estimating equations" (abbreviated as CCE hereafter). This approach requires the second-phase validation sample to be selected completely at random. Shepherd and Yu[12] proposed a moment-based estimator (MBE). They allow the selection of the second-phase validation sample to be stratified on an error-free covariate (eg, study site). Both the CCE and MBE are computationally simple but statistically inefficient. Shepherd et al[14] proposed a multiple imputation approach, which requires specifying the conditional distribution of measurement errors given the true outcome and covariates. This approach is sometimes more efficient, but may yield biased estimators when the error models are misspecified.

Classical two-phase studies with error-prone covariates are closely related to those with expensive covariates, where the outcome and inexpensive covariates are accurately measured for all subjects in the first phase and that information is used to select subjects for measurements of expensive covariates during the second phase. Efficient semiparametric estimation theory for two-phase studies with expensive covariates was established by Robins

et al.[15] For continuous outcomes, their augmented inverse-probability weighting estimator is difficult to implement in practice because it requires numerical solution of an infinite-dimensional integral equation. Efficient estimators that are computationally feasible were developed by Breslow et al,[16] Song et al,[17] Lin et al,[18] and Tao et al.[19] These approaches are more appealing than parametric ones (eg, multiple imputation) because they allow for an arbitrary covariate distribution. In addition, they accommodate outcome- or residual-dependent sampling designs, or both, which tend to be more efficient than (stratified) simple random sampling.[18,20] Efficient estimators that are computationally feasible have not been developed for two-phase studies with error-prone outcome and covariates.

The primary contributions of this article are: (a) to adapt the semiparametric efficient method of Tao et al[19] developed for two-phase studies with expensive covariates to settings with errors in covariates and (b) to extend this method to handle the often-encountered situation where there are also errors in a quantitative outcome, which may be correlated with the errors in covariates. Specifically, we relate the quantitative outcome and covariates of interest through linear regression models while leaving the distribution of errors unspecified. We consider additive errors in the outcome while accommodating both additive and multiplicative errors in covariates. We allow the existence and magnitude of measurement errors to be correlated with error-free covariates (if there are any). Dealing with this general framework is very challenging because the likelihood function involves the conditional density functions of measurement errors given quantitative covariates. We address this challenge by approximating the conditional density functions with B-spline sieves.[21,22] We maximize the sieve nonparametric likelihood function through a computationally efficient and numerically stable expectation-maximization (EM) algorithm. We show that the resulting estimators are consistent, asymptotically normal, and asymptotically efficient. We demonstrate the superiority of the proposed methods over existing approaches through extensive simulation studies. We illustrate their use in an observational HIV/AIDS study using the data from CCASAnet.

2 | METHODS

2.1 | Model and Data

Let Y and X denote the true outcome of interest and vector of covariates, respectively. We relate Y and X through the linear model

$$Y = \alpha + \beta^{\mathrm{T}} X + \epsilon,$$

where ϵ is normally distributed with mean zero and variance σ^2. In the database, Y^* and X^* are recorded instead of Y and X, where

$$Y^* = Y + W, \tag{1}$$

$$X^* = X + U, \tag{2}$$

W and U are the discrepancies between the true values of the outcome and covariates in the patient's clinical chart and those in the database, respectively. We assume that W and U are independent of ϵ.

The observation (Y^*, X^*, W, U) is assumed to be generated from the joint density

$$p(Y^*, X^*, W, U) = p(Y^* | X^*, W, U)p(W, U | X^*)p(X^*)$$
$$= p_\theta(Y | X)p(W, U | X^*)p(X^*),$$

(3)

where $p(Y^*|X^*, W, U)$ is the conditional density of Y^* given X^*, W, and U, $p(W, U|X^*)$ is the joint conditional density of W and U given X^*, $p(X^*)$ is the marginal density of X^*, and $p_\theta(Y|X)$ is a linear regression model indexed by parameter θ, that is,

$$p_\theta(Y | X) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{(Y - \alpha - \beta^T X)^2 / (2\sigma^2)\right\},$$

and $\theta = (\alpha, \beta^T, \sigma^2)^T$. In Equation (3), the equivalence of $p(Y^*|X^*, W, U)$ and $p_\theta(Y|X)$ follows from the additive error models (1) and (2) and the assumption that W and U are independent of ϵ. Our main interest lies in the inference of θ.

Remark 1. In classical measurement error problems, the outcome is accurately measured and only the covariates are error-prone. In this situation, the distribution of W has a point mass at zero, and Equation (3) reduces to

$$p(Y, X^*, U) = p_\theta(Y | X)p(U | X^*)p(X^*).$$

Alternatively, there may be a subset of the covariates that are error-prone, and the others are error-free. In this situation, we represent X and X^* as $\left(X_a^T, X_b^T\right)^T$ and $\left(X_a^{*T}, X_b^T\right)^T$, respectively, where the subscripts a and b correspond to error-prone and error-free covariates, respectively. Then, Equation (3) can be rewritten as

$$p(Y^*, X^*, W, U) = p_\theta(Y | X_a, X_b)p(W, U_a | X_a^*, X_b)p(X_a^*, X_b),$$

(4)

where U_a is the corresponding part of X_a in U. We observe from Equation (4) that the outcome and covariate measurement errors are allowed to depend on both error-prone and error-free covariates.

Remark 2. We do not assume a specific form for $p(W, U|X^*)$ in Equation (3). Therefore, we accommodate both unbiased errors that are centered around zero and biased errors that are not centered around zero. In addition, we allow for the measurement errors in X^* to be multiplicative. To see this, suppose that

$$X^* = X \circ \widetilde{U},$$

(5)

where $\widetilde{U}$ denotes the multiplicative errors in X^* that are independent of ϵ, and "$\circ$" denotes component-wise product. Equation (5) can be rewritten as $X^* = X + (X \circ \widetilde{U} - X)$, which has the same form as Equation (2) because $X \circ \widetilde{U} - X$ is independent of ϵ. As a side note, we cannot accommodate multiplicative errors $\widetilde{W}$ in Y^* $\left(\text{ie}, Y^* = Y\widetilde{W}\right)$ because $Y\widetilde{W} - Y$ is not independent of ϵ.

2.2 | Sieve maximum likelihood estimation

If data audits are performed for all n subjects in the study, then the inference on $\boldsymbol{\theta}$ is typically based on the likelihood $\prod_{i=1}^{n} p_\theta(Y_i | X_i)$. Under the two-phase design, however, only (Y^*, X^*) is recorded for all n subjects in the database, and (Y, X) is validated for a subsample of size n_2 in the second phase. Let V denote the indicator of a subject being selected for data auditing in the second phase. The two-phase design requires that the joint distribution of $(V_1, \ldots, V_n)$ depends on $(Y_i, X_i, Y_i^*, X_i^*, W_i, U_i)$ $(i = 1, \ldots, n)$ only through the first-phase data $(Y_i^*, X_i^*,)$ $(i = 1, \ldots, n)$; this is equivalent to assuming that the variables (Y, X, W, U) are missing at random. Thus, the observed-data log-likelihood takes the form

$$
\sum_{i=1}^{n} V_i \left\{ \log p_\theta(Y_i | X_i) + \log p(W_i, U_i | X_i^*) \right\}
$$
$$
+ \sum_{i=1}^{n} (1 - V_i) \log \int p_\theta(Y_i^* - w | X_i^* - \boldsymbol{u}) p(w, \boldsymbol{u} | X_i^*) dw d\boldsymbol{u} .
$$
(6)

We wish to maximize Expression (6) using nonparametric maximum likelihood estimation (NPMLE). That is, for each distinct observed $\boldsymbol{x}^*$, we wish to estimate $p(W, U | \boldsymbol{x}^*)$ by a discrete probability function on the distinct observed values of (W, U), denoted by $(w_1, \boldsymbol{u}_1)$, $\ldots, (w_m, \boldsymbol{u}_m)$ $(m \; n_2)$, where m is the total number of distinct values (ie, m increases with n_2). Unfortunately, this NPMLE approach is not feasible because only a small number of observations on (W, U) are associated with each $\boldsymbol{x}^*$.

To address this difficulty, we extend the sieve maximum likelihood estimation (SMLE) of Tao et al[19] for two-phase studies with expensive-covariates to studies with error-prone outcome and covariates. Specifically, we approximate $p(w, \boldsymbol{u} | X_i^*)$ and $\log p(w, \boldsymbol{u} | X_i^*)$ in Expression (6) by

$$
\sum_{k=1}^{m} I(w = w_k, \boldsymbol{u} = \boldsymbol{u}_k) \sum_{j=1}^{s_n} B_j^q(X_i^*) p_{kj},
$$
(7)

and

$$
\sum_{k=1}^{m} I(w = w_k, \boldsymbol{u} = \boldsymbol{u}_k) \sum_{j=1}^{s_n} B_j^q(X_i^*) \log p_{kj},
$$
(8)

respectively, where $B_j^q(\cdot)$ is the jth B-spline basis function of order q, s_n is the total number of functions in the B-spline basis, and p_{kj} is the coefficient of $B_j^q(X_i^*)$ at (w_k, u_k) ($k = 1, \ldots, m; j = 1, \ldots, s_n$) in the B-spline approximation of $p(w, u | X_i^*)$. Details about the construction of the B-spline basis $\left\{ B_j^q(\cdot) \right\}_{j=1}^{s_n}$ and guidelines about the choices of q and s_n can be found in Schumaker[22] or Tao et al.[19] In practice, q is typically chosen to be less than or equal to four, which corresponds to cubic splines, and s_n is determined by the first-phase sample size n. We note that by the approximation theory of B-splines,[22] both the log of Expression (7) and Expression (8) converge to $\log p(w, u | X_i^*)$ as $s_n \to \infty$. We choose Expression (8) over the log of Expression (7) because it is easier to compute. We standardize the B-spline basis such that $\sum_{j=1}^{s_n} B_j^q p(X_i^*) = I$ Consequently, p_{kj} needs to satisfy the constraints

$$\sum_{k=1}^{m} p_{kj} = 1 \text{ and } p_{kj} \geq 0 \ (k = 1, \ldots, m; j = 1, \ldots s_n), \tag{9}$$

because $p(w, u | X_i^*)$ is a conditional probability function. Given Expressions (7) and (8), the observed-data log-likelihood (6) can be rewritten as

$$
\begin{aligned}
l_n\left(\theta, \left\{ p_{kj} \right\}\right) \sum_{i=1}^{n} V_i & \left\{ \log p_\theta(Y_i | X_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(W_i = w_k, U_i = u_k) B_j^q(X_i^*) \log p_{kj} \right\} \\
& + \sum_{i=1}^{n} (1 - V_i) \log \left\{ \sum_{k=1}^{m} p_\theta(Y_i^* - w_k | X_i^* - u_k) \sum_{j=1}^{s_n} B_j^q(X_i^*) p_{kj} \right\}
\end{aligned}
\tag{10}
$$

.

We aim to maximize Expression (10) under the two constraints in Expression (9).

It is difficult to maximize Expression (10) directly because of the intractable form of the second term. Following Tao et al,[19] we solve this maximization problem by artificially creating a latent variable Z for subjects with $V = 0$ such that Z takes values on $1/s_n, 2/s_n, \ldots, 1$ and satisfies the equations

$$
\begin{aligned}
p(Z = j/s_n | X^*) &= B_j^q(X^*), \\
p(W = w_k, U = u_k | X^*, Z = j/s_n) = p(W = w_k, U = u_k | Z = j/s_n) &= p_{kj}, \\
p(Y^* | X^*, W, U, Z) = p(Y^* | X^*, W, U) &= p_\theta(Y | X).
\end{aligned}
$$

This step is essential because it enables us to interpret $\sum_{j=1}^{s_n} B_j^q(X^*) p_{kj}$ as $p(W = w_k, U = u_k | X^*)$ for subjects with $V = 0$. Hence, the second term in Expression (10) is equivalent to the log-likelihood of (Y_i^*, X_i^*), assuming that the complete data consist of $(Y_i^*, X_i^*, W_i, U_i, Z_i)$ but with $W_i, U_i,$ and Z_i missing.

The maximization of Expression (10) is carried out through an EM-algorithm, where (W, U, Z) for subjects with $V = 0$ are treated as missing data. The complete-data log-likelihood is

$$
\sum_{i=1}^{n} V_i \left\{ \log p_\theta(Y_i | X_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(W_i = w_k, U_i = u_k) B_j^q(X_i^*) \log p_{kj} \right\}
$$

$$
+ \sum_{i=1}^{n} (1 - V_i) \left\{ \log p_\theta(Y_i^* - W_i | X_i^* - U_i) + \log p(W_i, U_i | Z_i) + \log p(Z_i | X_i^*) \right\}
$$

$$
= \sum_{i=1}^{n} V_i \left\{ \log p_\theta(Y_i | X_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(W_i = w_k, U_i = u_k) B_j^q(X_i^*) \log p_{kj} \right\}
$$

$$
+ \sum_{i=1}^{n} (1 - V_i) \left\{ \sum_{k=1}^{m} I(W_i = w_k, U_i = u_k) \log p_\theta(Y_i^* - w_k | X_i^* - u_k) \right.
$$

$$
\left. + \sum_{k=1}^{m} \sum_{j=1}^{s_n} I(W_i = w_k, U_i = u_k, Z_i = j/s_n) \log p_{kj} + \sum_{j=1}^{s_n} I(Z_i = j/s_n) \log B_j^q(X_i^*) \right\}.
$$

We start with the following initial values: $\hat{\alpha}^{(0)} = 0$, $\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{0}$, $\hat{\sigma}^{2(0)}$ being the sample variance of Y^*, and $\hat{p}_{kj}^{(0)} = \sum_{i=1}^{n} V_i I(W_i = w_k, U_i = u_k) B_j^q(X_i^*) / \sum_{i=1}^{n} V_i B_j^q(X_i^*)$. We iterate between the following E-step and M-step until convergence.

In the E-step of the $(t+1)$th iteration, we calculate the conditional expectations of $I(W_i = w_k, U_i = u_k, Z_i = j/s_n)$ and $I(W_i = w_k, U_i = u_k)$ given $((Y_i^*, X_i^*), (w_1, u_1), \ldots, (w_m, u_m))$, evaluated at $\hat{\theta}^{(t)}, \hat{p}_{11}^{(t)}, \ldots, \hat{p}_{ms_n}^{(t)}$ denoted as $\hat{\psi}_{kji}^{(t+1)}$ and $\hat{q}_{ik}^{(t+1)}$, respectively. That is,

$$
\hat{\psi}_{kji}^{(t+1)} = \frac{p_{\hat{\theta}^{(t)}}(Y_i^* - w_k | X_i^* - u_k) B_j^q(X_i^*) \hat{p}_{kj}^{(t)}}{\sum_{k'=1}^{n} p_{\hat{\theta}^{(t)}}(Y_i^* - w_{k'} | X_i^* - u_{k'}) \sum_{j'=1}^{s_n} B_{j'}^q(X_i^*) \hat{p}_{k'j'}^{(t)}}, V_i = 0,
$$

$$
\hat{q}_{ik}^{(t+1)} = \begin{cases} I(W_i = w_k, U_i = u_k), \quad V_i = 1, \\ \dfrac{p_{\hat{\theta}^{(t)}}(Y_i^* - w_k | X_i^* - u_k) \sum_{j'=1}^{s_n} B_{j'}^q(X_i^*) \hat{p}_{kj'}^{(t)}}{\sum_{k'=1}^{m} p_{\hat{\theta}^{(t)}}(Y_i^* - w_{k'} | X_i^* - u_{k'}) \sum_{j'=1}^{s_n} B_{j'}^q(X_i^*) \hat{p}_{k'j'}^{(t)}}, V_i = 0. \end{cases}
$$

In the M-step of the $(t+1)$th iteration, we update $\hat{\theta}^{(t+1)}$ by maximizing

$$
\sum_{i=1}^{n} \sum_{k=1}^{m} \hat{q}_{ik}^{(t+1)} \log p_\theta(Y_i^* - w_k | X_i^* - u_k), \tag{11}
$$

such that

$$
\begin{pmatrix} \hat{\alpha}^{(t+1)} \\ \hat{\beta}^{(t+1)} \end{pmatrix} = \left(\sum_{i=1}^{n} \sum_{k=1}^{m} \hat{q}_{ik}^{(t+1)} \begin{bmatrix} 1 & (X_i^* - u_k)^{\mathrm{T}} \\ (X_i^* - u_k) & (X_i^* - u_k)(X_i^* - u_k)^{\mathrm{T}} \end{bmatrix} \right)^{-1}
$$

$$
\cdot \left(\sum_{i=1}^{n} \sum_{k=1}^{m} \hat{q}_{ik}^{(t+1)} \begin{bmatrix} Y_i^* - w_k \\ (X_i^* - u_k)(Y_i^* - w_k) \end{bmatrix} \right),
$$

$$
\hat{\sigma^2}^{(t+1)} = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m} \hat{q}_{ik}^{(t+1)} \left\{ Y_i^* - w_k - \hat{\alpha}^{(t+1)} - (\hat{\beta}^{(t+1)})^{\mathrm{T}}(X_i^* - u_k) \right\}^2.
$$

Then, we update $\hat{p}_{kj}^{(t+1)}(k = 1, \ldots, m; j = 1, \ldots, s_n)$ by maximizing

$$
\sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{j=1}^{s_n} \left\{ V_i \hat{q}_{ik}^{(t+1)} B_j^q(X_i^*) + (1 - V_i)\hat{\psi}_{kji}^{(t+1)} \right\} \log p_{kj},
$$

such that

$$
\hat{p}_{kj}^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{ V_i \hat{q}_{ik}^{(t+1)} B_j^q(X_i^*) + (1 - V_i)\hat{\psi}_{kji}^{(t+1)} \right\}}{\sum_{k'=1}^{m} \sum_{i=1}^{n} \left\{ V_i \hat{q}_{ik}^{(t+1)} B_j^q(X_i^*) + (1 - V_i)\hat{\psi}_{k'ji}^{(t+1)} \right\}}.
$$

We observe that $\hat{p}_{kj}^{(t+1)}$ satisfies the two constraints in Expression (9).

At convergence, we obtain the SMLEs $\hat{\theta}$ and $\hat{p}_{kj}(k = 1, \ldots, m; j = 1, \ldots, s_n)$. It follows from theorems S.1 and S.2 of Tao et al[19] that $\hat{\theta}$ is consistent, asymptotically normal, and asymptotically efficient as $n \to \infty$ and $n_2/n \to \Pr(V = 1) > 0$. To see this, we can redefine Y^* and X^* as the "outcome of interest" and "inexpensive covariates," respectively, which are available for all subjects in the first phase, and (W, U) as the "expensive covariates," which are available for subjects selected in the second phase only. Then, maximizing Expression (6) is equivalent to maximizing Expression (1) of Tao et al[19] under the constraints that the regression coefficient for W is fixed at one, and the regression coefficients for U and X^* are opposite to each other.

To obtain the variance estimate of $\hat{\theta}$, we use the profile likelihood method proposed by Murphy and van der Vaart.[23] By verifying the smoothness conditions of theorem 1 in Murphy and van der Vaart,[23] it can be shown that the negative inverse of the Hessian matrix of the profile likelihood function $pl(\theta) = \max_{\{p_{kj}\}} l_n(\theta, \{p_{kj}\})$ is a consistent estimator for the limiting covariance matrix of $n^{1/2}(\hat{\theta} - \theta)$. In practice, we obtain the value of $pl(\theta)$ by holding θ fixed in the EM algorithm and obtaining the value of $l_n(\theta, \{p_{kj}\})$ at convergence. We estimate the covariance matrix of $\hat{\theta}$ by the negative inverse of the matrix whose (k, l)th element is $h_n^{-2}\left\{ pl(\hat{\theta} + e_k h_n + e_l h_n) - pl(\hat{\theta} + e_k h_n) - pl(\hat{\theta} + e_l h_n) + pl(\hat{\theta}) \right\}$ where e_k is the kth canonical vector, and h_n is a constant of the order $n^{-1/2}$.

3 | SIMULATION STUDIES

We conducted extensive simulation studies to compare the performance of the SMLE, MBE, and CCE mimicking the settings in Shepherd and Yu.[12] In the first set of studies, we set X to be standard normal, and generated the outcome from the linear model: $Y = 0.3 + 0.4X + \epsilon$, where ϵ is a standard normal random variable independent of X. We generated $(W, U)^{\mathrm{T}}$ from a mixture distribution of a point mass at $(0,0)^{\mathrm{T}}$ and a bivariate normal distribution, that is,

$$\binom{W}{U} \sim \begin{cases} (0, 0)^{\mathrm{T}} \text{ with probability } 1 - p, \\ \mathrm{N}\left(\binom{0}{0}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right) \text{ with probability } p, \end{cases}$$

where p is a parameter controlling the proportion of subjects with measurement errors in Y^* and X^*, and r is a parameter controlling the correlation between W and U when both are not equal to zero. We varied p and r from 0.1 to 1 and −0.5 to 0.5, respectively. We generated Y^* and X^* from Equations (1) and (2), respectively. We set $n = 1000$ and selected $n_2 = 400$ subjects randomly in the second phase. For subjects selected in the second phase, the data consist of (Y, X, Y^*, X^*, W, U); for those not selected in the second phase, the data consist of (Y^*, X^*). When implementing the SMLE method, we estimated $p(W, U|X^*)$ using cubic splines. We partitioned the domain of X^* using evenly-spaced quantiles and varied s_n from 15 to 25 to assess its effects on model-fitting. The results with different s_n are very similar; the maximum difference in the coverage probability of the 95% confidence interval for $\hat{\beta}$ is only 0.6%. Therefore, we only report the results for $s_n = 20$. We estimated the covariance matrix of $\hat{\theta}$ by the profile likelihood method with step size of $0.1n^{-1/2}$.

The results for the first set of simulations are shown in Table 1. The SMLE, MBE, and CCE were virtually unbiased. Their variance estimators reflected the true variations, and their corresponding confidence intervals had reasonable coverage probabilities. The SMLE was more efficient than the CCE, which tended to be more efficient than the MBE. The efficiency gain of the SMLE over MBE and CCE increased and decreased, respectively, as the proportion of subjects with measurement errors in Y^* and X^* increased. The efficiency gain was larger when the correlation between W and U was negative as compared to that when the correlation was positive. For benchmark comparison, we also used standard linear regression based on least squares estimation (LSE) to analyze the validation sample only. The results are summarized in Table S1 of the Supporting Information. The LSE was less efficient than the CCE or SMLE, although it was more efficient than the MBE in cases with a large proportion of subjects having measurement errors. For $p = 0.6$ and $r = 0.3$, we also considered smaller n_2 and reported the results in Table S2. The SMLE, MBE, and CCE performed reasonably well when n_2 ≥ 50, but tended to underestimate the variance when $n_2 = 25$.

In a second set of simulations, we considered both error-prone and error-free covariates. Specifically, we set X_a and X_b to be standard normal and Bern(0.25), respectively. We generated the outcome from the linear model: $Y = 0.3 + 0.4X_a + 0.5X_b + \epsilon$, where ϵ is a

standard normal random variable independent of (X_a, X_b). We generated $(W, U_a)^\mathrm{T}$ from the following mixture distribution

$$\binom{W}{U_a} \sim \begin{cases} \text{when } X_b = 0 \begin{cases} (0,0)^\mathrm{T} \text{ with probability } 0.4, \\ N\left(\binom{0}{0}, 0.5\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right) \text{with probability} \quad 0.6, \end{cases} \\ \text{when } X_b = 1 \begin{cases} (0,0)^\mathrm{T} \text{ with probability } 1-p, \\ N\left(\binom{0}{0}, \tau\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right) \text{with probability } p, \end{cases} \end{cases}$$

where τ is a parameter controlling the magnitude of the measurement errors when $X_b = 1$. We varied p, r, and τ from 0.6 to 1, 0.3 to 0.5, and 0.5 to 1, respectively. We generated Y^* and X_a^* from Equation (1) and the linear model $X_a^* = X_a + U_a$, respectively. We set $n = 1000$ and considered two sampling strategies in the second phase: simple random sampling selects 400 subjects randomly; stratified simple random sampling selects 200 subjects from each stratum of X_b randomly. When implementing the SMLE method, we estimated $p(W, U_a | X_a^*, X_b = 0)$ and $p(W, U_a | X_a^*, X_b = 1)$ using separate cubic splines. The results under simple random sampling and stratified simple random sampling are shown in Tables S3 and S4 of the Supporting Information, respectively. Under simple random sampling, the SMLE was more efficient than the MBE and CCE for X_a. The efficiency gain was larger when the proportion of subjects with measurement errors or the magnitude of errors were heterogeneous across the strata as compared to when they were homogeneous. The MBE and CCE were as efficient as the SMLE for X_b. Under stratified simple random sampling, the SMLE and MBE continued to perform well. The variance estimator of CCE underestimated the true variation, and its confidence interval undercovered.

To assess the performance of the SMLE and the robustness of the MBE and CCE under biased errors that are not centered around zero, we generated (W, U) from the bivariate normal distribution

$$\binom{W}{U} \sim N\left(\binom{\mu_W}{\mu_U}, \binom{1 \ \ r}{r \ \ 1}\right),$$

where μ_W and μ_U denote the mean of W and U, respectively. We varied μ_W and μ_U from 0 to 0.5. We generated (Y, X, Y^*, X^*) in the same manner as in the first set of simulations. The results are summarized in Table 2. The SMLE and CCE performed well in all scenarios. The MBE performed well only when at most one of W or U was not centered around zero, but was severely biased when both W and U were not centered around zero. These conclusions held no matter whether W and U were correlated or not.

To assess the performance of the SMLE and the robustness of the MBE and CCE under multiplicative errors in X^*, we generated X^* from the model: $X^* = X\{1 + \exp(-U)\}^{-1}$. We generated (Y, X, Y^*, W, U) in the same manner as in the first set of simulations. The results are summarized in Table 3. The SMLE and CCE continued to perform well. The MBE was

severely biased even when the proportion of subjects with measurement errors was as low as 10%.

To assess the robustness of the SMLE, MBE, and CCE to the normality assumption, we generated data in the same manner as in the first set of studies but let ϵ follow t-distributions with 3 to 30 degrees of freedom or the Uniform$(-c, c)$ distribution, where $c = 1$ or 2. We fixed p and r at 0.6 and 0.3, respectively. The results are summarized in Table S5 of the Supporting Information. The SMLE, MBE, and CCE performed well in these situations.

Next, we considered residual-dependent sampling rather than simple or stratified simple random sampling in the second phase. We generated (Y, X, Y^*, X^*, W, U) for 1000 subjects in the same manner as in the first set of studies. We calculated the residuals from the linear model relating Y^* to X^* for all subjects. We then selected 200 subjects with the highest and 200 subjects with the lowest values of residuals in the second phase. The results are summarized in Tables 4 and S6 of the Supporting Information. The SMLE continued to perform well under residual-dependent sampling. The LSE, MBE, and CCE incorrectly applied to this setting were severely biased, yielding poor coverage probabilities for their confidence intervals. The bias of the LSE and CCE tended to be larger than that of the MBE. We observe from the last column of Table 4 that residual-dependent sampling could be more efficient than simple random sampling for two-phase studies with measurement errors.

Finally, we evaluated the performance of the SMLE with more than one error-prone covariate. Specifically, we set $X = (X_1, X_2)$, where X_1 and X_2 are standard normal. We generated the outcome from the linear model: $Y = 0.3 + 0.4X_1 + 0.4X_2 + \epsilon$, where ϵ is a standard normal random variable independent of X. We generated $(W, U)^{\mathrm{T}}$ from a mixture distribution of a point mass at $(0,0,0)^{\mathrm{T}}$ and a trivariate normal distribution, that is,

$$\binom{W}{U} \sim \begin{cases} (0, 0, 0)^{\mathrm{T}} \text{ with probability } 1 - p, \\ \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{pmatrix} \right) \text{ with probability } p. \end{cases}$$

We varied p and r from 0.1 to 1 and 0 to 0.5, respectively. We generated Y^* and X^* from Equations (1) and (2), respectively. We set $n = 1000$ and selected $n_2 = 400$ subjects randomly in the second phase. When implementing the SMLE method, we estimated $p(W, U|X^*)$ using the tensor product of two one-dimensional cubic-spline bases for X_1 and X_2, each with six evenly spaced interior knots. Simulation results are shown in Table S7. The SMLE continued to perform well, with bias close to zero and coverage near the nominal level for regression coefficients for both covariates.

4 | CCASANET STUDY

CCASANet is a multi-site cohort designed to address questions about the HIV epidemic in Latin America using existing clinical databases. CCASANet data include patient characteristics, date of HIV diagnosis, dates of clinic visits, longitudinal laboratory measurements, ART medications and dates, clinical events, follow-up information, and vital

status. Study sites submit datasets to the CCASAnet data coordinating center at Vanderbilt University, which then merges the data for analyses. The CCASAnet data coordinating center periodically performs data audits, where auditors visit the study site and compare data sent to the coordinating center with data in the patients' clinical charts. Detailed descriptions of the CCASAnet cohort and data audit procedures are given by McGowan et al[7] and Duda et al,[6] respectively.

Shepherd and Yu[12] illustrated the MBE method with CCASAnet data, evaluating the association between ART initiation date and CD4 at ART initiation. Here we apply our new SMLE method to the exact same CCASAnet data to contrast methods. A total of 2815 HIV-positive patients starting ART from 1996 to 2007 at sites in Argentina, Brazil, Chile, Honduras, Mexico, and Peru were included in this analysis. To preserve anonymity, sites were randomly labeled as Sites A-F. The data coordinating center audited a total of 234 patients, randomly sampled at each study site, between April 2007 and March 2008. CD4 count at ART initiation was defined as the CD4 measurement taken closest to, but no more than seven days after or 180 days before, the ART initiation date.

The data audits found that 16% of the ART initiation dates in the CCASAnet databases were different from those in the clinical charts. Although CD4 count was generally correct, when the ART initiation date was incorrect in the database, the CD4 count at the incorrect date was sometimes not the same as that at the true ART initiation date. Consequently, 4.3% of the CD4 counts at ART initiation were incorrect, and some of the differences between unvalidated and validated CD4 were quite large (as big as 12.4 $(\text{cells/mm}^3)^{1/2}$); see table 3 and figure 1 of Shepherd and Yu[12] for more details, including a scatterplot of the errors. Square-root transformed CD4 count at ART initiation and ART initiation date were the outcome and covariate of interest, respectively. In addition, we included gender and study site as error-free covariates. There appeared to be no correlation between gender and the errors in CD4 count or ART initiation date conditioning on study site. On the other hand, the error rates and magnitude of CD4 count and ART initiation date varied across the study sites; see Table 3 of Shepherd and Yu.[12] When implementing the SMLE method, we used separate linear splines for the study sites. We chose Site F as the reference site and used two, two, five, zero, and three evenly spaced interior knots for Sites A, B, C, D, and E, respectively. We used more interior knots for Site C because it had the largest number of errors (ie, 16). We did not use any interior knots for Site D because the data audits identified only one erroneous record. In this situation, the B-spline basis reduces to a constant function.

Table 5 shows the results for the SMLE and MBE methods, a naive analysis that ignored the database errors, and the LSE method using validation data only. The estimates of the LSE method appeared to be quite different from the other methods, with its 95% confidence intervals much wider because of the small validation sample size. The SMLE and MBE methods yielded similar effect size estimates for ART initiation date, which were larger than the naive estimate. The corresponding 95% confidence intervals of the SMLE and MBE did not include zero, while that of the naive estimate did. The 95% confidence interval of the SMLE for ART initiation date was 27.2% narrower than that of the MBE. These results were consistent with the theoretical and simulation results. The positive association between ART

initiation date and CD4 count suggested that the HIV-positive patients in CCASAnet started their medications in less advanced stages of HIV-disease in later years of the study. This trend was consistent with guidelines encouraging patients to initiate ART at higher CD4 counts.[24] We did not apply the CCE method here because the selection of audited records was stratified by study site, which violated the assumption that the second-phase validation sample is a simple random sample from the first-phase sample.

5 | DISCUSSION

We have developed valid and efficient semiparametric inference procedures for general two-phase studies with an error-prone quantitative outcome and error-prone covariates. The proposed method requires minimal assumptions for the error models. It can be applied to any two-phase design in which, conditional on the first phase data, the second-phase sample selection is independent of the true values of the outcome and covariates. Therefore, the SMLE approach can be applied to efficient designs that existing methods cannot address (eg, outcome-dependent sampling and residual-dependent sampling). Even with simple random sampling, however, the efficiency gains of the SMLE over the LSE, the MBE of Shepherd and Yu,[12] and CCE of Chen and Chen[13] are substantial. The proposed EM algorithm is numerically stable and not sensitive to the choice of initial values (results not shown). In our simulation studies, the algorithm converged in all replicates in each scenario.

As mentioned in Section 1, the proposed SMLE approach is a novel extension of the method of Tao et al,[19] which was developed for two-phase studies with expensive covariates. The method of Tao et al[19] assumes that Y is observed for everyone; therefore, it cannot accommodate outcome measurement error. The proposed method can simultaneously accommodate outcome and covariate errors. The EM algorithm presented in Section 2.2, and the corresponding software implementation are all novel developments.

In our simulation studies, the number of B-spline basis functions s_n had little impact on the parameter estimates. In principle, one could use the Akaike information criterion or Bayesian information criterion to select the "optimal" s_n. Alternatively, one could choose s_n through cross-validation.

In our sieve approximation to $p(W, U|X^*)$, X^* cannot contain too many continuous components. This is because the multivariate B-spline basis is built by the tensor-product of one-dimensional B-spline bases.[19,22] Consequently, it suffers from the curse of dimensionality. If there is prior knowledge that W and U are independent of some error-free covariates, then these covariates can be omitted from X^* when estimating $p(W, U|X^*)$. Alternatively, one could assume a parametric transformation $h : \mathbb{R}^d \to \mathbb{R}^{d_1}$, where d is the dimension of X^*, and $d_1 < d$, such that W and U depend on X^* only through $h(X^*)$. There are numerous choices for h (eg, the top components in a principal component analysis), each with potentially different robustness properties that warrant further study.

We considered classical measurement error models (1) and (2), where the observed value in the database equals the true value plus measurement error. Alternatively, one may consider Berkson measurement error models, where the true value equals the observed value plus

measurement error, that is, $Y = Y^* + W$, and $X = X^* + U$.[25] Practical guidance on determining whether the data follow classical or Berkson measurement error models can be found in Carroll et al.[9] Our framework can be easily modified to accommodate Berkson errors, where W and U are independent of X^* and ϵ. In this situation, $p(W, U|X^*) = p(W, U)$, where $p(W, U)$ is the joint distribution of W and U. We can estimate $p(W, U)$ by a discrete probability function on the distinct observed values of (W, U). Consequently, Equation (10) can be rewritten as

$$l_n(\theta, \{r_k\}) = \sum_{i=1}^{n} V_i \left\{ \log p_\theta(Y_i | X_i) + \sum_{k=1}^{m} I(W_i = w_k, U_i = u_k) \log r_k \right\}$$
$$+ \sum_{i=1}^{n} (1 - V_i) \log \left\{ \sum_{k=1}^{m} p_\theta(Y_i^* + w_k | X_i^* + u_k) r_k \right\},$$

where $r_k = \Pr(W = w_k, U = u_k)$. The maximization of $l_n(\theta, \{r_k\})$ is simpler than that of $l_n(\theta, \{p_{kj}\})$ because the former does not involve B-spline sieves.

We have focused on linear regression models with quantitative measurement errors. Our framework can be extended to generalized linear models with categorical data subject to classification errors or proportional hazards models with time-to-event errors. In our EM algorithm, the E-step and the M-step for updating p_{kj} ($k = 1, \dots, m; j = 1, \dots, s_n$) are generic for any regression model. The M-step for updating θ involves the maximization of Expression (11), which is a weighted sum of the log-likelihood functions for the regression model. Consequently, we can use existing algorithms for weighted regression to maximize Expression (11). We are currently working on these extensions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

DATA AVAILABILITY STATEMENT

The SMLE approach has been implemented in the R package TwoPhaseReg, which is freely available on GitHub at https://github.com/dragontaoran/TwoPhaseReg. All simulation and analysis code is available on GitHub at https://github.com/dragontaoran/proj_two_phase_mexy.

REFERENCES

1. Mullooly JP. The effects of data entry error: an analysis of partial verification. Comput Biomed Res. 1990;23(3):259–267. [PubMed: 2350961]

2. Weiss RB. Systems of protocol review, quality assurance and data audit. Cancer Chemother Pharmacol. 2002;42:S88–S92.

3. Chaulagai CN, Moyo CM, Koot J, et al. Design and implementation of a health management information system in Malawi: issues innovations and results. Health Policy Plan. 2005;20(6):375–384. [PubMed: 16143590]

4. Kiragga AN, Castelnuovo B, Schaefer P, Muwonge T, Easterbrook PJ. Quality of data collection in a large HIV observational clinic database in sub-Saharan Africa: implications for clinical research and audit of care. J Int AIDS Soc. 2011;14(1):3–3. [PubMed: 21251327]

5. Mphatswe W, Mate KS, Bennett B, et al. Improving public health information: a data quality intervention in KwaZulu-Natal South Africa. Bull World Health Organ. 2012;90(3):176–182. [PubMed: 22461712]

6. Duda SN, Shepherd BE, Gadd CS, Masys DR, McGowan CC. Measuring the quality of observational study data in an international HIV research network. PLoS One. 2012;7(4):e33908. [PubMed: 22493676]

7. McGowan CC, Cahn P, Gotuzzo E, et al. Cohort profile: Caribbean, Central and South America Network for HIV research (CCASAnet) collaboration within the international epidemiologic databases to evaluate AIDS (IeDEA) programme.IntJEpidemiol. 2007;36(5):969–976.

8. Fuller WA. Measurement Error Models. New York, NY: John Wiley & Sons; 1987.

9. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models: A Modern Perspective. Boca Raton, FL: Chapman & Hall/CRC Press; 2006.

10. Xu Y, Kim JK, Li Y. Semiparametric estimation for measurement error models with validation data. Can J Stat. 2017;45(2):185–201.

11. Abrevaya J, Hausman JA. Response error in a transformation model with an application to earnings-equation estimation. Econ J. 2004;7(2):366–388.

12. Shepherd BE, Yu C. Accounting for data errors discovered from an audit in multiple linear regression. Biometrics. 2011;67(3):1083–1091. [PubMed: 21281274]

13. Chen YH, Chen H. A unified approach to regression analysis under double-sampling designs. J Royal Stat Soc B. 2000;62(3):449–460.

14. Shepherd BE, Shaw PA, Dodd LE. Using audit information to adjust parameter estimates for data errors in clinical trials. Clin Trials. 2012;9(6):721–729. [PubMed: 22848072]

15. Robins JM, Hsieh F, Newey W. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. J Royal Stat Soc B. 1995;57(2):409–424.

16. Breslow NE, McNeney B, Wellner JA. Large sample theory for semiparametric regression models with two-phase outcome dependent sampling. Ann Stat. 2003;31(4):1110–1139.

17. Song R, Zhou H, Kosorok MR. A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. Biometrika. 2009;96(1):221–228. [PubMed: 20107493]

18. Lin DY, Zeng D, Tang ZZ. Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proc Natl Acad Sci. 2013;110(30):12247–12252. [PubMed: 23847208]

19. Tao R, Zeng D, Lin DY. Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. J Am Stat Assoc. 2017;112(520):1468–1476. [PubMed: 29479125]

20. Tao R, Zeng D, Lin DY. Optimal designs of two-phase studies. J Am Stat Assoc. 2020; in press. 10.1080/01621459.2019.1671200.

21. Grenander U Abstract Inference. New York, NY: Wiley; 1981.

22. Schumaker L Spline Functions: Basic Theory. New York, NY: Wiley-Interscience; 1981.

23. Murphy SA, van der Vaart AW. On profile likelihood. J Am Stat Assoc. 2000;95(450):449–465.

24. Richardson ET, Grant PM, Zolopa AR. Evolution of HIV treatment guidelines in high- and low-income countries: converging recommendations. Antivir Res. 2014;103:88–93. [PubMed: 24374148]

25. Berkson J Are there two regressions? J Am Stat Assoc. 1950;45(25):164–180.

TABLE 1

Simulation results under additive errors in Y^* and X^* when simple random sampling is used in the second phase

		MBE					CCE					SMLE			
r	*p*	Bias	SE	SEE	CP	RE	Bias	SE	SEE	CP	RE	Bias	SE	SEE	CP
−0.5	0.1	0.001	0.041	0.040	0.950	0.748	−0.002	0.039	0.038	0.942	0.839	−0.002	0.035	0.035	0.946
	0.3	0.002	0.051	0.051	0.952	0.559	−0.001	0.043	0.043	0.948	0.806	−0.004	0.038	0.038	0.945
	0.6	0.004	0.062	0.062	0.950	0.451	0.000	0.045	0.045	0.946	0.856	−0.007	0.042	0.042	0.946
	1.0	0.004	0.072	0.072	0.951	0.393	0.001	0.046	0.046	0.949	0.958	−0.008	0.045	0.044	0.943
−0.3	0.1	0.000	0.039	0.039	0.949	0.800	−0.001	0.038	0.038	0.944	0.869	−0.002	0.035	0.035	0.946
	0.3	0.001	0.050	0.049	0.952	0.610	−0.001	0.042	0.042	0.947	0.833	−0.003	0.039	0.038	0.946
	0.6	0.002	0.061	0.060	0.950	0.484	0.000	0.045	0.045	0.947	0.879	−0.005	0.042	0.042	0.947
	1.0	0.002	0.072	0.071	0.950	0.410	0.001	0.046	0.046	0.949	0.984	−0.006	0.046	0.044	0.942
0.0	0.1	−0.001	0.038	0.038	0.949	0.871	−0.001	0.037	0.037	0.949	0.913	−0.001	0.035	0.035	0.942
	0.3	0.000	0.047	0.047	0.949	0.676	0.000	0.042	0.041	0.947	0.864	−0.002	0.039	0.038	0.947
	0.6	0.000	0.058	0.058	0.948	0.547	0.000	0.045	0.044	0.944	0.914	−0.002	0.043	0.042	0.946
	1.0	0.000	0.070	0.069	0.948	0.440	0.001	0.046	0.046	0.947	1.004	−0.003	0.046	0.045	0.940
0.3	0.1	−0.002	0.037	0.036	0.946	0.901	0.000	0.036	0.036	0.951	0.928	0.000	0.035	0.034	0.944
	0.3	−0.001	0.045	0.045	0.950	0.741	0.000	0.041	0.041	0.948	0.882	0.000	0.039	0.038	0.946
	0.6	−0.001	0.054	0.055	0.952	0.618	0.000	0.044	0.044	0.945	0.913	0.000	0.042	0.042	0.947
	1.0	−0.001	0.067	0.066	0.946	0.475	0.001	0.046	0.046	0.949	0.996	−0.001	0.046	0.045	0.943
0.5	0.1	−0.002	0.036	0.036	0.947	0.913	0.000	0.036	0.036	0.952	0.935	0.000	0.035	0.034	0.946
	0.3	−0.002	0.044	0.043	0.950	0.769	0.000	0.041	0.040	0.948	0.883	0.001	0.038	0.038	0.944
	0.6	−0.002	0.052	0.052	0.952	0.650	0.000	0.044	0.043	0.944	0.908	0.001	0.042	0.041	0.948
	1.0	−0.002	0.064	0.063	0.946	0.507	0.001	0.046	0.045	0.948	0.986	0.001	0.045	0.044	0.943

Notes: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the empirical mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the efficiency relative to that of the SMLE. Each entry is based on 10 000 replicates.

TABLE 2

Simulation results under errors in Y^* and X^* that are not centered around zero

			MBE				CCE					SMLE			
r	μ_U	μ_W	Bias	SE	SEE	CP	Bias	SE	SEE	CP	RE	Bias	SE	SEE	CP
−0.5	0.0	0.0	0.004	0.072	0.072	0.951	0.001	0.046	0.046	0.949	0.958	−0.008	0.045	0.044	0.943
		0.5	0.003	0.077	0.076	0.949	0.001	0.046	0.046	0.949	0.958	−0.008	0.045	0.044	0.943
	0.5	0.0	0.004	0.076	0.076	0.952	0.001	0.046	0.046	0.949	0.958	−0.008	0.045	0.044	0.943
		0.5	−0.250	0.074	0.074	0.088	0.001	0.046	0.046	0.949	0.958	−0.008	0.045	0.044	0.943
0.0	0.0	0.0	0.000	0.070	0.069	0.948	0.001	0.046	0.046	0.947	1.004	−0.003	0.046	0.045	0.940
		0.5	0.000	0.074	0.074	0.951	0.001	0.046	0.046	0.947	1.004	−0.003	0.046	0.045	0.940
	0.5	0.0	0.001	0.074	0.074	0.946	0.001	0.046	0.046	0.947	1.004	−0.003	0.046	0.045	0.940
		0.5	−0.252	0.078	0.075	0.087	0.001	0.046	0.046	0.947	1.004	−0.003	0.046	0.045	0.940
0.5	0.0	0.0	−0.002	0.064	0.063	0.946	0.001	0.046	0.045	0.948	0.986	0.001	0.045	0.044	0.943
		0.5	−0.002	0.068	0.067	0.947	0.001	0.046	0.045	0.948	0.986	0.001	0.045	0.044	0.943
	0.5	0.0	−0.001	0.069	0.068	0.946	0.001	0.046	0.045	0.948	0.986	0.001	0.045	0.044	0.943
		0.5	−0.254	0.080	0.071	0.062	0.001	0.046	0.045	0.948	0.986	0.001	0.045	0.044	0.943

Notes: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the empirical mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the efficiency relative to that of the SMLE. Each entry is based on 10 000 replicates.

TABLE 3

Simulation results under multiplicative errors in X^* and additive errors in Y^*

r	p	MBE				CCE					SMLE			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP	RE	Bias	SE	SEE	CP
−0.5	0.1	0.587	0.086	0.085	0.000	0.000	0.034	0.034	0.947	0.991	0.004	0.034	0.034	0.947
	0.3	0.562	0.097	0.096	0.000	0.000	0.038	0.038	0.948	0.945	0.003	0.037	0.036	0.947
	0.6	0.524	0.109	0.108	0.002	0.000	0.041	0.041	0.947	0.917	0.002	0.040	0.039	0.950
	1.0	0.482	0.122	0.121	0.018	0.000	0.043	0.043	0.950	0.992	0.001	0.043	0.042	0.943
−0.3	0.1	0.587	0.086	0.085	0.000	0.000	0.034	0.034	0.948	0.995	0.004	0.034	0.034	0.946
	0.3	0.562	0.097	0.096	0.000	0.000	0.038	0.038	0.949	0.958	0.003	0.037	0.037	0.946
	0.6	0.524	0.109	0.108	0.002	0.000	0.041	0.041	0.945	0.936	0.002	0.040	0.040	0.950
	1.0	0.482	0.121	0.121	0.018	0.000	0.043	0.043	0.949	1.016	0.001	0.044	0.042	0.943
0.0	0.1	0.587	0.086	0.085	0.000	0.000	0.034	0.034	0.947	0.991	0.003	0.034	0.034	0.946
	0.3	0.562	0.097	0.096	0.000	0.000	0.038	0.038	0.945	0.943	0.003	0.037	0.037	0.946
	0.6	0.526	0.108	0.108	0.002	0.000	0.041	0.041	0.943	0.940	0.002	0.040	0.040	0.949
	1.0	0.482	0.120	0.121	0.018	0.000	0.043	0.043	0.947	1.005	0.001	0.043	0.043	0.942
0.3	0.1	0.587	0.086	0.085	0.000	0.000	0.034	0.034	0.950	0.985	0.003	0.034	0.034	0.947
	0.3	0.562	0.096	0.096	0.000	0.000	0.038	0.038	0.947	0.921	0.003	0.037	0.037	0.948
	0.6	0.527	0.108	0.108	0.001	0.000	0.041	0.041	0.947	0.936	0.002	0.040	0.040	0.946
	1.0	0.483	0.121	0.121	0.020	0.001	0.043	0.043	0.948	1.010	0.001	0.043	0.042	0.941
0.5	0.1	0.587	0.086	0.085	0.000	0.000	0.034	0.034	0.949	0.976	0.003	0.034	0.034	0.947
	0.3	0.561	0.096	0.096	0.000	0.000	0.038	0.038	0.947	0.907	0.003	0.037	0.036	0.949
	0.6	0.527	0.108	0.109	0.002	0.000	0.042	0.041	0.948	0.917	0.002	0.040	0.039	0.945
	1.0	0.483	0.121	0.121	0.021	0.000	0.043	0.043	0.948	0.989	0.001	0.043	0.042	0.941

Notes: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the empirical mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the efficiency relative to that of the SMLE. Each entry is based on 10 000 replicates.

TABLE 4

Simulation results under additive errors in Y^* and X^* when residual-dependent sampling is used in the second phase

		MBE				CCE				SMLE				
r	*p*	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP	RE
−0.5	0.1	0.083	0.045	0.042	0.499	0.102	0.045	0.055	0.568	−0.002	0.032	0.032	0.951	1.089
	0.3	0.170	0.057	0.050	0.089	0.205	0.050	0.055	0.026	−0.007	0.035	0.035	0.946	1.096
	0.6	0.198	0.065	0.055	0.065	0.242	0.050	0.052	0.003	−0.012	0.040	0.040	0.938	1.039
	1.0	0.170	0.071	0.058	0.202	0.222	0.049	0.049	0.007	−0.014	0.046	0.046	0.939	0.977
−0.3	0.1	0.063	0.043	0.040	0.660	0.080	0.044	0.055	0.746	−0.001	0.032	0.032	0.952	1.089
	0.3	0.136	0.055	0.049	0.216	0.166	0.051	0.057	0.138	−0.006	0.035	0.035	0.948	1.117
	0.6	0.174	0.065	0.055	0.142	0.200	0.051	0.054	0.033	−0.010	0.039	0.040	0.943	1.071
	1.0	0.166	0.073	0.060	0.241	0.180	0.051	0.052	0.062	−0.011	0.046	0.046	0.945	0.994
0.0	0.1	0.033	0.040	0.038	0.859	0.047	0.043	0.054	0.923	−0.001	0.032	0.032	0.949	1.091
	0.3	0.077	0.052	0.047	0.616	0.099	0.051	0.057	0.614	−0.003	0.034	0.034	0.949	1.135
	0.6	0.107	0.062	0.054	0.503	0.124	0.053	0.057	0.408	−0.006	0.038	0.038	0.949	1.129
	1.0	0.112	0.073	0.061	0.542	0.111	0.053	0.054	0.466	−0.007	0.046	0.046	0.950	1.010
0.3	0.1	0.006	0.038	0.037	0.939	0.012	0.041	0.053	0.986	0.000	0.032	0.032	0.952	1.084
	0.3	0.017	0.048	0.044	0.913	0.026	0.050	0.058	0.956	0.000	0.034	0.034	0.952	1.150
	0.6	0.025	0.059	0.052	0.887	0.033	0.054	0.058	0.931	−0.002	0.037	0.037	0.950	1.142
	1.0	0.027	0.070	0.059	0.879	0.030	0.054	0.056	0.926	−0.003	0.044	0.044	0.951	1.043
0.5	0.1	−0.009	0.037	0.036	0.933	−0.011	0.041	0.052	0.986	0.001	0.032	0.032	0.952	1.077
	0.3	−0.018	0.046	0.043	0.908	−0.025	0.049	0.057	0.958	0.002	0.033	0.034	0.952	1.142
	0.6	−0.025	0.056	0.050	0.888	−0.034	0.053	0.058	0.928	0.002	0.037	0.037	0.952	1.145
	1.0	−0.028	0.066	0.055	0.875	−0.034	0.054	0.057	0.919	0.001	0.042	0.043	0.953	1.068

Notes: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the empirical mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the empirical variance of the SMLE under simple random sampling over that under residual-dependent sampling. Each entry is based on 10 000 replicates.

TABLE 5

Effect size estimates and 95% confidence intervals from the analysis of the CCASAnet data

Covariate	LSE		Naive		MBE		SMLE	
	Est	**(95% CI)**	**Est**	**(95% CI)**	**Est**	**(95% CI)**	**Est**	**(95% CI)**
ART initiation date (per year)	−0.248	(−0.688, 0.191)	0.117	(−0.009, 0.243)	0.187	(0.006, 0.368)	0.174	(0.042, 0.305)
Male	0.177	(−1.513, 1.867)	−0.736	(−1.182, −0.290)	−0.740	(−1.192, −0.288)	−0.725	(−1.17, −0.280)
Site A	−0.093	(−2.695, 2.509)	1.281	(0.540, 2.021)	1.310	(0.663, 1.956)	1.110	(0.341, 1.879)
Site B	−1.340	(−3.904, 1.224)	0.932	(0.268, 1.597)	1.063	(0.393, 1.734)	0.904	(0.231, 1.577)
Site C	2.452	(0.102, 4.803)	2.759	(2.203, 3.315)	2.821	(2.232, 3.411)	2.434	(1.835, 3.032)
Site D	1.209	(−2.445, 4.862)	2.389	(1.602, 3.176)	2.614	(1.736, 3.491)	2.494	(1.710, 3.278)
Site E	0.705	(−1.686, 3.095)	0.576	(−0.075, 1.227)	0.636	(0.003, 1.269)	0.627	(−0.035, 1.289)

Notes: Est and CI stand for effect size estimate and confidence interval, respectively.