Jean de Dieu Tapsoba[1] / Edward C. Chao[2] / Ching-Yun Wang[3]

Simulation Extrapolation Method for Cox Regression Model with a Mixture of Berkson and Classical Errors in the Covariates using Calibration Data

[1] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A., E-mail: jtapsoba@fredhutch.org

[2] Data Numerica Institute, Bellevue, Washington 98006, U.S.A.

[3] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.

Abstract:
Many biomedical or epidemiological studies often aim to assess the association between the time to an event of interest and some covariates under the Cox proportional hazards model. However, a problem is that the covariate data routinely involve measurement error, which may be of classical type, Berkson type or a combination of both types. The issue of Cox regression with error-prone covariates has been well-discussed in the statistical literature, which has focused mainly on classical error so far. This paper considers Cox regression analysis when some covariates are possibly contaminated with a mixture of Berkson and classical errors. We propose a simulation extrapolation-based method to address this problem when two replicates of the mismeasured covariates are available along with calibration data for some subjects in a subsample only. The proposed method places no assumption on the mixture percentage. Its finite-sample performance is assessed through a simulation study. It is applied to the analysis of data from an AIDS clinical trial study.

Keywords: Berkson error, classical error, instrumental variable, proportional hazards model, simulation extrapolation

DOI: 10.1515/ijb-2018-0028

Received: March 5, 2018; **Revised:** January 13, 2019; **Accepted:** March 20, 2019

1 Introduction

Many epidemiological or biomedical studies often aim to investigate the association between the time to an event of interest and some covariates. The examination of the association is usually performed under the popular Cox proportional hazards model [1] that stipulates proportional covariate effects on the hazard function for the time-to-event. A commonly encountered problem in such an examination is that some covariates potentially involve measurement error. This may be due to imprecision of the data measuring instrument, high costs to obtain precise measurements or the nature of the covariates themselves. For example, CD4 counts in HIV studies [2], dietary protein intake in nutrition studies and radiation dose in radiation epidemiology studies [3] are well-known to be subject to measurement error. Naively omitting the measurement error and simply using the error-affected covariate data in the survival analysis could lead to invalid inference [4].

There are two different types of measurement errors, namely the classical type and the Berkson type [3]. The main difference between these two types of errors lies in their relationship with the unobserved covariates. The classical error is assumed to be independent of the unobserved covariates whereas the Berkson error is positively correlated with the unobserved covariates. Self-reported data, collected through questionnaire typically involve recall error that is usually assumed to be of the classical type [5]. Also, grouped-exposure data obtained by assigning the same value of the unobserved exposure variable (e. g. air pollution or radiation dose) to individuals sharing some characteristics are often assumed to be contaminated with Berkson error [6]. In many other situations, the covariate data may be affected by a mixture of errors of both classical and Berkson types. For example, DS02 radiation dose estimates for atomic-bomb survivors, who were followed up by the Radiation Effects Research Foundation (RERF) are contaminated with both errors [7]. The classical error is thought

Jean de Dieu Tapsoba is the corresponding author.

to originate from the survivors' individual recollections of location and shielding at the time of the bombings and the Berkson error is believed to be associated with averaging radiation dose estimates among survivors with common location and shielding type. Taken individually and if omitted, measurement error of each of these two types could affect the estimation of the survival model parameters. The impact of the classical error in the covariates of a Cox regression model has been well-documented and includes the attenuation of the covariate effect [8, 9]. Moreover, a few references have reported that the Berkson error may also have a similar effect [6, 10]. However, the effect of a mixture of classical and Berkson errors in the covariates of a proportional hazards model still remains to be well investigated. The presence of errors of both types in the covariates of the regression model is expected to worsen the attenuation effect as will become clear in the simulation study section. It is therefore important to simultaneously account for both types of errors when there are present in the covariates of a Cox proportional hazards model.

Several methods have been developed to deal with covariate measurement error in a Cox proportional hazards model and extensive attention has been given to the classical error. Popular methods adjusting for classical error in survival analysis are the regression calibration (RC) method [11, 12], simulation extrapolation (SIMEX) method [3, 13, 14], parametric corrected score [15], non-parametric corrected score [8, 16], conditional score [17, 18], maximum likelihood method [19, 20] and non-parametric maximum likelihood method [21]. On the other hand, methods addressing the issue of Berkson error in the covariates in Cox regression model are very scarce and this problem still has to be well-discussed in the literature. Among the very limited number of references, [22] discussed an extension to a Berkson error of a corrected profile likelihood method for Cox model with classical error in the covariates. It is important to note that methods that simply deal with classical error or Berkson error exclusively may not work well when errors of both types are involved in the covariates. Moreover, the existing methods accounting for a mixture of errors of both types are developed for generalized linear models [23, 24] and may not be directly applicable in the context of survival analysis. Therefore, new statistical methods that jointly adjust for both classical and Berkson errors need to be developed for a reliable inference in this situation.

In this paper, we address the problem of parameter estimation in Cox regression analysis when some covariates are subject to a combination of both Berkson and classical errors. We are not aware of an existing reference in the statistical literature that has properly dealt with this problem. We assume that two replicates of the error-prone covariates are available for all individuals. Also, we assume that data on an instrumental variable for the mis-measured covariate are available only for some individuals in a subset of the study cohort that is here referred to as calibration subsample. For illustration, we consider the AIDS Clinical Trials Group (ACTG) 175 study, a randomized double-blind clinical trial that was conducted to compare four antiretroviral treatments with either a single nucleoside or two nucleosides in (HIV-1) infected adult participants [25]. In addition to sociodemographic data, the study also collected data on survival time and virologic variables such as CD4 cell count and CD8 cell count. An important interest is in assessing the effect of the true baseline CD4 cell count on time to AIDS disease development or death under a Cox proportional hazards model adjusting for the treatment effect. A complication related to the use of standard survival analysis methods for this assessment is the possible contamination of the CD4 cell count data with measurement error, which has been assumed to be of the classical type in previous applications [8, 16]. Here we allow the measurement error to incorporate features of both classical and Berkson error types. We use a mixture of classical and Berkson errors model, which is more general than the classical model. Moreover, we treat the two CD4 cell counts taken before the start of the treatment and within three weeks of randomization as replicates. Also, the CD8 cell count measured at randomization is likely correlated with the baseline CD4 cell count and may be used as instrumental variable for the true baseline CD4 count. There were 2441 participants with two replicates for the true baseline CD4 cell count and only 1459 of them had CD8 cell count measured at randomization. We propose a SIMEX-based new approach to adjusting for both classical error and Berkson error under these settings. It places no assumption on the mixture proportion of the error variances.

In the next section, we provide the model formulations for the survival data and the mixture of classical and Berkson errors. In Section 3, we present the naive and RC methods and develop the SIMEX method for this problem in Section 4. We present a simulation study in Section 5 and apply the developed method to the data from the ACTG 175 study in Section 6. Finally, we conclude with some discussions in Section 7.

2 Model formulations

Suppose that there are n study subjects, which are followed-up over a study period of length τ. For subject i, let T_i^0 denote the unobserved survival time, which is assumed to be right-censored and C_i be the censoring time. The observed survival data are (T_i, δ_i), where $T_i = \min(T_i^0, C_i)$ is the observed failure time and $\delta_i = I(T_i^0 \leq$

C_i) is the non-censoring indicator, $i = 1,..., n$. Also, let $\mathbf{X}_i$ denote a vector of time-independent covariates that are subject to measurement error, and $\mathbf{Z}_i$ represent a vector of other time-independent covariates that can be measured accurately. It is assumed that replicates $\mathbf{W}_{ij}, j = 1, \dots, k_i (\geq 2)$ of $\mathbf{X}_i$ are available for all subjects and data on an instrumental variable Q_i for $\mathbf{X}_i$ are available only for some subjects in a calibration subsample. Letting η_i indicate whether Q_i is observed ($\eta_i = 1$) or not ($\eta_i = 0$) and $\mathbf{W}_i = (\mathbf{W}_{i1}^{'}, \dots, \mathbf{W}_{ik_i}^{'})^{'}$, the observed data are $(T_i, \delta_i, \mathbf{W}_i, \mathbf{Z}_i, \eta_i Q_i), i = 1,..., n$, which are assumed to be independent and identically distributed random vectors. In the ACTG 175 data example, δ_i is the indicator of the event of interest, which is AIDS disease progression or death during the study follow-up and T_i is the time to the event if $\delta_i = 1$ or end of follow-up (due to lost to follow up or end of the study observation period) if $\delta_i = 0$. Also, X_i is a single covariate representing the log-transformation of baseline CD4 counts, $\mathbf{Z}_i$ is a vector of indicator variables of randomized treatment groups, W_{ij} ($j = 1,2$) are the log-transformation of two CD4 count measurements taken before treatment initiation and Q_i is the log-transformation of baseline CD8.

We consider the following mixture of Berkson and classical errors model for the relation between $\mathbf{W}_{ij}$ and $\mathbf{X}_i$:

$$\begin{cases} \mathbf{X}_i = \mathbf{L}_i + \mathbf{U}_{b,i}, \\ \mathbf{W}_{ij} = \mathbf{L}_i + \mathbf{U}_{c,ij}, \end{cases} \tag{1}$$

where $\mathbf{L}_i$ is a latent variable with mean $\boldsymbol{\mu}_l$ and variance Σ_l, and $\mathbf{U}_{b,i}$ is a zero-mean error with variance Σ_b. Also, $\mathbf{U}_{c,ij}, j = 1, \dots, k_i$, are independent and identically normally distributed with mean 0 and variance Σ_c. Moreover, $\mathbf{U}_{c,ij}$ is independent of $\mathbf{U}_{b,i}, j = 1, \dots, k_i$, $(\mathbf{U}_{c,i1}, \dots, \mathbf{U}_{c,ik_i})$ and $\mathbf{U}_{b,i}$ are independent of $(T_i^0, C_i, \mathbf{L}_i, \mathbf{Z}_i)$, $i = 1,..., n$. Model (1), which was also studied by [23, 26] and [27] in different contexts encompasses features of both classical and Berkson error models. It reduces to a Berkson measurement error model when $\Sigma_c = 0$ and a classical error model when $\Sigma_b = 0$. Furthermore, Q_i is assumed to be associated with $\mathbf{X}_i$ according to the following model:

$$Q_i = \alpha_0 + \boldsymbol{\alpha}_1^{'} \mathbf{X}_i + \epsilon_i, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1^{'})^{'}$ is a vector of unknown parameters and ϵ_i is a zero mean normally distributed random variable with variance σ_ϵ^2 and independent of $(T_i^0, C_i, \mathbf{L}_i, \mathbf{Z}_i, \mathbf{U}_{b,i}, \mathbf{U}_{c,i1}, \dots, \mathbf{U}_{c,ik_i})$, $i = 1,..., n$. The time-to-event is associated with $\mathbf{X}_i$ and $\mathbf{Z}_i$ via the Cox proportional hazards model, which is a commonly used tool for survival data analysis. Assuming that T_i^0 is conditionally independent of C_i, $\mathbf{W}_i$ and Q_i given $\mathbf{X}_i$ and $\mathbf{Z}_i$, the model formulates the hazard function as follows.

$$\lambda(u|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, Q_i) = \lambda_0(u) \exp(\boldsymbol{\beta}_x^{'} \mathbf{X}_i + \boldsymbol{\beta}_z^{'} \mathbf{Z}_i), \tag{3}$$

where $\lambda_0(.)$ is an unspecified baseline hazards function and $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^{'}, \boldsymbol{\beta}_z^{'})^{'}$ is a vector of unknown parameters. Our interest lies in the estimation of $\boldsymbol{\beta}$ based on the observed data. In what follows, we let $Y_i(u) \equiv I(T_i \geq u)$ denote the at-risk process and define the counting process $N_i(u) \equiv I(T_i \leq u, \delta_i = 1)$, where $I(.)$ is the indicator function.

3 Naive and RC methods

The naive and RC estimations are two simple imputation-based methods that could be used to estimate the parameter $\boldsymbol{\beta}$ although, they might not handle correctly the measurement errors, especially when the error magnitudes are not small.

3.1 Naive method

The estimation of $\boldsymbol{\beta}$ would be based on the usual Cox partial likelihood score if $\mathbf{X}_i$ were observed. In such a situation, the partial likelihood estimator for $\boldsymbol{\beta}$ would solve $U_n(\boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}) = 0$, where

$$U_n(\boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left\{ \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} - \frac{S_1(u, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z})}{S_0(u, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z})} \right\} dN_i(u), \tag{4}$$

and $S_k(u, \boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}) = n^{-1} \sum_{i=1}^{n} Y_i(u) \{(\mathbf{X}_i^{'}, \mathbf{Z}_i^{'})^{'}\}^k \exp(\boldsymbol{\beta}_x^{'} \mathbf{X}_i + \boldsymbol{\beta}_z^{'} \mathbf{Z}_i)$, $k = 0,1$. Since $\mathbf{X}_i$ cannot be observed in our context, a naive estimator of $\boldsymbol{\beta}$ is obtained by replacing $\mathbf{X}_i$ by $\bar{\mathbf{W}}_i = k_i^{-1} \sum_{j=1}^{k_i} \mathbf{W}_{ij}$ in (4) and solving $U_n(\boldsymbol{\beta}, \bar{\mathbf{W}}, \mathbf{Z}) =$

0. It has been well documented in the literature pertaining to Cox regression with covariates affected by classical measurement error that the naive estimator is often downwardly biased [9, 28]. Also, it has been reported in [10] that the naive approach could lead to inconsistent estimation of the baseline hazard as well as incorrect tests for the proportional hazards assumption when the error is of Berkson type. Hence, it is clear that ignoring the mixture of Berkson and classical errors in our context will also result in a biased estimation and expectedly a more severe attenuation effect compared with the situation when the error is simply of Berkson or classical type.

3.2 Regression calibration method

The RC method is a simple and well-known method that can be applied to reduce the bias caused by the naive estimation. It has been extensively studied by [11, 29] and [12] in the context of classical error in the covariates of a Cox model. One implementation (RC1) of RC for our problem is to use $\mathbb{E}(\mathbf{X}_i|\bar{\mathbf{W}}_i)$ in place of $\mathbf{X}_i$ in (4). Another RC approach (RC2) that makes use of the calibration data consists to replace $\mathbf{X}_i$ in (4) with $\mathbb{E}(\mathbf{X}_i|\bar{\mathbf{W}}_i, \eta_i Q_i)$ which is $\mathbb{E}(\mathbf{X}_i|\bar{\mathbf{W}}_i, Q_i)$ if i is in the calibration sample ($\eta_i = 1$) and $\mathbb{E}(\mathbf{X}_i|\bar{\mathbf{W}}_i)$ otherwise ($\eta_i = 0$), $i = 1, \ldots, n$. In general, RC methods perform well in linear regression when the error is classical, Berkson or a mixture of Berkson and classical errors. They may also work satisfactorily in nonlinear regression when the measurement error or the covariate effect is small. However, they do not lead to consistent estimation for generalized linear models when the link function is different from the identity function and the error is of classical type [3], Chapter 6], Berkson type [30] or a combination of both classical and Berkson types [24]. In Cox regression with covariates affected by a classical error, [11, 28] noted that the bias of RC estimator is large when the covariate effect or the error variance is large. As will be seen in our simulation study, RC may not perform well in Cox regression with a mixture of classical and Berkson errors in the covariates. A conditional expectation of the form $\mathbb{E}\{h(\mathbf{X}_i)|\bar{\mathbf{W}}_i, \eta_i Q_i\}$ can be evaluated based on the conditional density function of $(\mathbf{L}_i, \mathbf{U}_{b,i})$ given $(\bar{\mathbf{W}}_i, \eta_i Q_i)$ as follows.

$$E\{h(\mathbf{X}_i)|\bar{\mathbf{W}}_i, \eta_i Q_i\} = \frac{\int_l \int_{u_b} h(\mathbf{X}_i = l + u_b)\mathscr{L}(Q_i|\mathbf{X}_i = l + u_b)^{\eta_i}\mathscr{L}(\bar{\mathbf{W}}_i|\mathbf{L}_i = l)\mathscr{L}(\mathbf{L}_i = l)\mathscr{L}(\mathbf{U}_{b,i} = u_b)dldu_b}{\int_l \int_{u_b} \mathscr{L}(Q_i|\mathbf{X}_i = l + u_b)^{\eta_i}\mathscr{L}(\bar{\mathbf{W}}_i|\mathbf{L}_i = l)\mathscr{L}(\mathbf{L}_i = l)\mathscr{L}(\mathbf{U}_{b,i} = b)dldu_b},$$

where $\mathscr{L}(.)$ denotes density function. Integrals that are involved in the conditional expectations can be computed by means of the Gauss-Hermite quadrature.

4 SIMEX approach

The SIMEX method is a two-step procedure consisting of a simulation step and a subsequent extrapolation step. It was originally developed by [13] to deal with classical error in the covariates of a linear or nonlinear regression model and later studied by [3, 14, 31] and [32]. In addition, applications of this method to a mixture of errors of both types in the covariates of a generalized linear models are discussed in [3] and [24]. Here we provide an extension of the method to the Cox proportional hazards model when some covariates are subject to both classical and Berkson errors. We describe the implementation of the two steps of the procedure for our problem in what follows. The simulation step requires an estimating function that would yield a consistent estimator of $\boldsymbol{\beta}$ if there were no classical error ($\Sigma_c = 0$) and $\mathbf{X}_i$ were substituted with $\bar{\mathbf{W}}_i$ in that estimating function. The partial likelihood score in (4) would have been a suitable choice for this purpose if the error were simply of classical type. However, its use in our context is complicated by the presence of Berkson-type error in addition to classical error. We instead use estimation equations based on a maximum likelihood function.

It should be noted that if $\mathbf{X}_i$ were observed and $\lambda_0(.)$ known, the maximum likelihood estimator of $\boldsymbol{\beta}$ would solve the following equation:

$$\sum_{i=1}^{n} \int_0^{\tau} \left\{ \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} dN_i(u) - Y_i(u) \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} \exp(\boldsymbol{\beta}_x^{'}\mathbf{X}_i + \boldsymbol{\beta}_z^{'}\mathbf{Z}_i)\lambda_0(u)du \right\} = 0.$$

Also, it is useful to note that $\mathbf{L}_i = \bar{\mathbf{W}}_i$ if the error were simply of Berkson type ($\Sigma_c = 0$) under Model (1). Assuming that $\lambda_0(.)$ and the distribution of $(\mathbf{L}_i, \mathbf{Z}_i)$ do not involve $\boldsymbol{\beta}$, a consistent estimator for $\boldsymbol{\beta}$ in the presence of simply Berkson error ($\Sigma_c = 0$) would be obtained by solving

$$\sum_{i=1}^{n} \int_0^{\tau} \mathbb{E}\left[\left\{ \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} dN_i(u) - Y_i(u) \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} \exp(\boldsymbol{\beta}_x^{'}\mathbf{X}_i + \boldsymbol{\beta}_z^{'}\mathbf{Z}_i)\lambda_0(u)du \right\} |N_i, Y_i, \mathbf{L}_i, \mathbf{Z}_i \right] = 0. \qquad (5)$$

Also, a Breslow type estimator for the baseline hazards function $\lambda_0(.)$ could be expressed as $\hat{\lambda}_0(u) = \sum_{i=1}^{n} dN_i(u) / \sum_{i=1}^{n} Y_i(u)\mathbb{E}\left\{\exp(\boldsymbol{\beta}_x^{'}\mathbf{X}_i + \boldsymbol{\beta}_z^{'}\mathbf{Z}_i)|N_i, Y_i, \mathbf{L}_i, \mathbf{Z}_i\right\}$. However, $\mathbf{L}_i$ cannot be observed due to the presence of classical error (in addition to the Berkson error) and simply replacing $\mathbf{L}_i$ with $\bar{\mathbf{W}}_i$ in (5) will potentially lead to a biased estimator for $\boldsymbol{\beta}$ in our situation. In the simulation step of the proposed SIMEX-based method, a number of $\mathscr{R}$ (with $\mathscr{R} > 1$) naive estimates of $\boldsymbol{\beta}$ are obtained by solving (5), in which $\mathbf{L}_i$ is imputed with $\mathbf{W}_{\zeta,r,i} = \bar{\mathbf{W}}_i + \sqrt{\zeta}\mathbf{U}_{r,i}$, where ζ is a non-negative scalar and $\mathbf{U}_{r,i} \sim N(0, k_i^{-1}\Sigma_c)$, $i = 1, \ldots, n$, $r = 1, \ldots, \mathscr{R}$. Let $\hat{\boldsymbol{\beta}}_r(\zeta_j)$ denote the naive estimator for $\boldsymbol{\beta}$ using $\mathbf{W}_{\zeta_j,r,i}$ in place of L_i in (5) and denote $\hat{\boldsymbol{\beta}}(\zeta_j) = \mathscr{R}^{-1}\sum_{r=1}^{\mathscr{R}} \hat{\boldsymbol{\beta}}_r(\zeta_j)$ the pseudo-estimate of $\boldsymbol{\beta}$ using ζ_j, where $0 = \zeta_0 < \zeta_1 < \ldots < \zeta_J$, and $J > 1$ is the number of pseudo-estimates. The extrapolation step consists in fitting a regression model of each component of $\hat{\boldsymbol{\beta}}(\zeta_j)$ on the ζ_j's, $j = 1, \ldots, J$, using the ordinary least squares estimation method. The SIMEX estimate of each component of $\boldsymbol{\beta}$ is then obtained by extrapolating back to the case when $\zeta = -1$, which represents the situation of no classical error. Common extrapolation functions are polynomial in ζ of the form $\mathscr{P}(\zeta, \mathbf{a}) = a_0 + a_1\zeta + \ldots + a_q\zeta^q$, where $q \geq 1$ is the degree of the polynomial. For a particular component of $\boldsymbol{\beta}$, an estimator $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_q)^{'}$ for the parameter $\mathbf{a} = (a_0, \ldots, a_q)^{'}$ is obtained via the regression of the corresponding components of the $\hat{\boldsymbol{\beta}}(\zeta_j)'s$ on the ζ_j's, $j = 1, \ldots, J$. For example, letting $\hat{\boldsymbol{\beta}}_x(\zeta_j)$ represent the component of $\hat{\boldsymbol{\beta}}(\zeta_j)$ that corresponds to β_x, and $\hat{\boldsymbol{\beta}}_x(\boldsymbol{\zeta}) = \{\hat{\beta}_x(\zeta_1), \ldots, \hat{\beta}_x(\zeta_J)\}^{'}$, an estimator for $\mathbf{a}$ would be $\hat{\mathbf{a}} = \{\mathscr{D}(\boldsymbol{\zeta})^{'}\mathscr{D}(\boldsymbol{\zeta})\}^{-1}\mathscr{D}(\boldsymbol{\zeta})^{'}\hat{\boldsymbol{\beta}}_x(\boldsymbol{\zeta})$, where $\mathscr{D}(\boldsymbol{\zeta})$ is the $J \times q$-matrix whose jth row is $(1, \zeta_j, \ldots, \zeta_j^q)^{'}$. The SIMEX estimate of that particular component of $\boldsymbol{\beta}$ is then obtained as $\mathscr{P}(-1, \hat{\mathbf{a}}) = \hat{a}_0 - \hat{a}_1 + \ldots + \hat{a}_q(-1)^q$. Here $\mathscr{R}$, J and q are assumed known. The standard errors can be obtained based on the bootstraps method. The SIMEX method leads to a consistent estimator for $\boldsymbol{\beta}$ if the true expression of the extrapolation function is used. In practice, approximations to the form of the extrapolation function are necessary as its closed form expression is unknown. It is important to note that the performance of the method could be affected by the choice of the extrapolation function.

A conditional expectation of the form $\mathbb{E}\{h(\mathbf{X}_i)|N_i, Y_i, \mathbf{L}_i, \mathbf{Z}_i\}$ involved in (5) is given as follows.

$$\mathbb{E}\{h(\mathbf{X}_i)|N_i, Y_i, \mathbf{L}_i, \mathbf{Z}_i\} = \frac{\int_{u_b} h(\mathbf{X}_i = \mathbf{L}_i + u_b)\mathscr{L}(N_i, Y_i|\mathbf{X}_i = \mathbf{L}_i + u_b, \mathbf{Z}_i)\mathscr{L}(\mathbf{L}_i, \mathbf{Z}_i)\mathscr{L}(\mathbf{U}_{b,i} = u_b)du_b}{\int_{u_b} \mathscr{L}(N_i, Y_i|\mathbf{X}_i = \mathbf{L}_i + u_b, \mathbf{Z}_i)\mathscr{L}(\mathbf{L}_i, \mathbf{Z}_i)\mathscr{L}(\mathbf{U}_{b,i} = u_b)du_b},$$

where

$$\mathscr{L}(N_i, Y_i|\mathbf{X}_i, \mathbf{Z}_i) = \prod_{u \in [0,\tau]} \{Y_i(u)\lambda_0(u)\}^{dN_i(u)} \exp\left\{dN_i(u)(\boldsymbol{\beta}_x^{'}\mathbf{X}_i + \boldsymbol{\beta}_z^{'}\mathbf{Z}_i) - e^{\boldsymbol{\beta}_x^{'}\mathbf{X}_i + \boldsymbol{\beta}_z^{'}\mathbf{Z}_i} \int_0^{\tau} Y_i(u)\lambda_0(u)du\right\}.$$

The nuisance parameters $\boldsymbol{\mu}, \boldsymbol{\alpha}, \Sigma_l, \Sigma_b, \Sigma_c$ and σ_ϵ^2 are needed for the evaluation of the conditional expectations involved in the implementation of the RC and SIMEX methods. They may be estimated from the data if unknown. For the single covariate case for example, the vector of nuisance parameters is $\boldsymbol{\nu} = (\mu, \alpha_0, \alpha_1, \sigma_l^2, \sigma_b^2, \sigma_c^2, \sigma_\epsilon^2)$, where $\sigma_l^2 = \Sigma_l$, $\sigma_b^2 = \Sigma_b$ and $\sigma_c^2 = \Sigma_c$. All the components of $\boldsymbol{\nu}$ can be identified using the observations $(W_{i1}, \ldots, W_{ik_i}, \eta_i Q_i, T_i)$, $i = 1, \ldots, n$. Moreover, a consistent estimator of $\boldsymbol{\nu}$ can be obtained based on the method of moments, which leads to the following estimating equations:

$$\begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{k_i}(W_{ij} - \mu_l) = 0; \\ \sum_{i=1}^{n}\sum_{j=1}^{k_i-1}\sum_{j^{'}=j+1}^{k_i}\{(W_{ij} - \mu_l)(W_{ij^{'}} - \mu_l) - \sigma_l^2\} = 0; \\ \sum_{i=1}^{n}\sum_{j=1}^{k_i-1}\sum_{j^{'}=j+1}^{k_i}\{(W_{ij} - W_{ij^{'}})^2 - 2\sigma_c^2\} = 0; \\ \sum_{i=1}^{n}\eta_i(Q_i - \alpha_0 - \alpha_1\mu_l) = 0; \\ \sum_{i=1}^{n}\sum_{j=1}^{k_i}\eta_i\{(Q_i - \alpha_0 - \alpha_1\mu_l)(W_{ij} - \mu_l) - \alpha_1\sigma_l^2\} = 0; \\ \sum_{i=1}^{n}\eta_i\{(Q_i - \alpha_0 - \alpha_1\mu_l)^2 - \alpha_1^2(\sigma_l^2 + \sigma_b^2) - \sigma_\epsilon^2\} = 0; \\ \sum_{i=1}^{n}\sum_{j=1}^{k_i}\eta_i(T_i - \bar{T})\{(Q_i - \alpha_0 - \alpha_1\mu_l)\sigma_l^2 - \alpha_1(W_{ij} - \mu_l)(\sigma_l^2 + \sigma_b^2)\} = 0, \end{cases}$$

where $\bar{T} = n^{-1}\sum_{i=1}^{n} T_i$.

5 Simulation study

We performed a simulation study to examine the finite-sample performance of the proposed SIMEX method and compare it to the naive and RC methods for the case of a single covariate with a mixture of Berkson and classical measurement errors. The latent variable L_i was generated from $N(\mu_l, \sigma_l^2)$ with $\mu_l = 0$ and $\sigma_l^2 = 1$. The unobserved covariate X_i was simulated as $X_i = L_i + U_{b,i}$, where the Berkson error, $U_{b,i}$ followed a zero-mean normal distribution with variance σ_b^2. Two replicates, $W_{ij}, j = 1, 2 = k_i$, were generated from the classical error model $W_{ij} = L_i + U_{c,ij}$ with $U_{c,ij}$ generated from $N(0, \sigma_c^2)$. We set $\sigma_b^2 = 0.3$ or 0.5, and $\sigma_c^2 = 0.3$ or 0.5 to study the separate and combined effects of the Berkson error and classical error on the performances of the various estimators. The survival times were simulated following the Cox proportional hazards model with unit baseline hazard $\lambda(u|X_i, Z_i) = \exp(\beta_x X_i + \beta_z Z_i)$, where Z_i followed a Bernoulli trial with a success probability of 0.5, $\beta_x = \log(2)$ and $\beta_z = 0$. A common censoring was applied to subjects inducing a censoring rate of 50%. The instrumental variable was simulated following the model $Q_i = \alpha_0 + \alpha_1 X_i + \epsilon_i$, where ϵ_i was normal with mean 0 and variance σ_ϵ^2. We took $\alpha_0 = 1$, $\alpha_1 = 0.8$ and $\epsilon_i = 0.5$. The variable η_i, indicating whether Q_i is available or not was generated from the Bernoulli distribution with probability of success $P(\eta_i = 1) = c$. We set $c = 0.5$ or 0.7 to explore how the proportion of the calibration data influences the results.

A total of 500 Monte Carlo samples of size $n = 500$ or 1000 were generated in the simulations. We estimated the parameter of interest $\boldsymbol{\beta} = (\beta_x, \beta_z)^{'}$ based on the naive method (Naive), which replaces X_i by $\bar{W}_i = (W_{i1} + W_{i2})/2$, RC method (RC1) that uses $E(X_i|\bar{W}_i)$ in place of X_i, RC method (RC2) that substitutes X_i by $E(X_i|\bar{W}_i, Q_i)$ if $\eta_i = 1$ and $E(X_i|\bar{W}_i)$ otherwise and the SIMEX method with quadratic polynomial (SIM1) or polynomial of degree 4 (SIM2) as extrapolation function in extrapolation step. For the SIMEX method, we created $\mathscr{R} = 50$ additional data sets of measurement error at each point $\zeta \in \{0, 0.25, 0.5, 0.75, 1\}$ in the simulation step. In addition, the vector of nuisance parameters $\boldsymbol{\nu} = (\mu_l, \alpha_0, \alpha_1, \sigma_l^2, \sigma_b^2, \sigma_c^2, \sigma_\epsilon^2)^{'}$ was estimated by the method of moments. Integrals involved in the evaluation of condition expectations for the implementation of RC and SIMEX methods were computed using Gauss-Hermite integration techniques with 20 quadrature points. The estimators were evaluated with regard to their biases (Bias), sample standard deviation of the estimates (SD), average of the estimated standard erros (ASE) of the estimators, mean squared errors (MSE = $\text{Bias}^2 + \text{SD}^2$) and coverage probabilities (CP) of their 95% Wald confidence intervals. The standard errors for RC1, RC2 estimators were computed using the sandwich method, whereas the estimation of the standard error for the SIMEX estimators was based on the bootstrap method, re-sampling 30 times.

Table 1 shows the results of the simulations pertaining to the estimation of the parameter $\boldsymbol{\beta}$ for 500. For β_x, the naive estimator noticeably shows the worst performance among all the methods with respect to biases and coverage probabilities. Its biases and MSE's are substantially large and increase as the variance of the Berkson error or that of the classical error becomes large. The attenuation effect seems to be more serious with the classical error than the Berkson error. The bias problem is severely accentuated by the presence of errors of both types. Also, it can be seen that its coverage probabilities are well below the nominal level of 95%. Hence, it is important to correct for both types of errors when they are present in the covariates of a Cox regression model. RC methods appear to have reduced the bias from the naive estimator although not completely. RC2, which uses the data in the calibration subsample generally exhibits smaller biases (except when $\sigma_b^2 = 0.3$ and $P(\eta = 1) = 0.5$) and larger coverage probabilities than RC1. Both SIM1 and SIM2 display smaller biases than the RC-based estimators. SIM1 works well in correcting the bias from the Berkson error and performs acceptably when the magnitude of the classical error is small. However, its bias becomes noticeable when the variance of the classical error is large. This highlights the importance of the choice of the extrapolation function for SIMEX method, which may perform well in terms of bias reduction when the choice is appropriate. Meanwhile, SIM2 appears to show very small biases and good coverage probabilities, although its MSE is larger than the RC and SIM1 counterparts. It can also be noted that the biases and ASE's of RC2, SIM1 and SIM2 decrease to some extent as the proportion of the calibration data increases. Moreover, all methods, including the naive one perform equally well with respect to biases for the estimation of the parameter β_z. The results of the simulations for the sample size $n = 1000$ are shown in Table 2. The performances of the methods are similar to those observed in Table 1. Also, the biases and ASE's are smaller than those presented in Table 1 for all the methods.

Table 1: Simulation results for the estimation of $\boldsymbol{\beta}$: $n = 500$; $\lambda(u) = \exp(\beta_x X + \beta_z Z)$, $X = L + U_b$; $L \sim N(0,1)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $W_j = L + U_{c,j}$; $U_{c,j} \sim N(0, \sigma_c^2)$ is the classical error, $j = 1,2$; $\boldsymbol{\beta} = (\log(2), 0)'$; $Q = 1 + 0.8X + \epsilon$; $\varepsilon \sim N(0,0.5)$; $p(\eta = 1)$ is the proportion of the calibration data; $Z \sim Bernoulli(0.5)$; censoring rate is 50%; Naive, "naive" estimator; RC1, RC replacing X by $\mathrm{E}(X|\bar{W})$; RC2, RC replacing X by $\{\mathrm{E}(X|\bar{W}, Q)\}^{\eta}\{\mathrm{E}(X|\bar{W})\}^{1-\eta}$; SIM1, SIMEX using quadratic extrapolation function; SIM2, SIMEX using polynomial of degree 4 as extrapolation function.

σ_b^2	$\boldsymbol{\beta}$		$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$				
			Naive	**RC1**	**RC2**	**SIM1**	**SIM2**	**Naive**	**RC1**	**RC2**	**SIM1**	**SIM2**
						$\mathrm{P}(\eta = 1) = 0.5$						
0.3	β_x	Bias	−13.075	−4.551	−4.580	−0.325	0.442	−18.097	−5.135	−5.218	−1.689	0.392
		SD	6.238	7.228	7.812	8.776	12.506	5.902	7.518	8.318	9.003	13.858
		ASE	6.327	7.101	8.081	8.859	12.312	6.027	7.325	8.650	9.194	13.747
		MSE	2.099	0.730	0.820	0.771	1.566	3.623	0.829	0.964	0.839	1.922
		CP	44.490	89.780	93.587	95.190	93.788	16.633	87.174	92.986	94.790	94.188
	β_z	Bias	0.302	0.302	0.338	0.327	0.279	0.325	0.325	0.364	0.355	0.284
		SD	6.797	6.797	6.760	7.386	7.873	6.797	6.797	6.761	7.466	8.217
		ASE	6.346	6.338	6.353	6.984	7.688	6.346	6.336	6.354	7.075	8.069
		MSE	0.463	0.463	0.458	0.547	0.621	0.463	0.463	0.458	0.559	0.676
		CP	93.186	92.587	93.587	92.585	94.188	93.788	92.986	93.788	93.186	94.389
0.5	β_x	Bias	−14.668	−6.386	−6.018	−0.502	0.307	−19.522	−6.923	−6.567	−1.996	0.163
		SD	6.155	7.121	7.963	9.150	13.028	5.817	7.388	8.521	9.294	14.309
		ASE	6.301	7.070	8.462	9.270	12.823	6.006	7.301	9.091	9.571	14.244
		MSE	2.530	0.915	0.996	0.840	1.698	4.149	1.025	1.157	0.904	2.048
		CP	35.200	83.800	91.000	94.600	94.000	10.800	82.600	91.000	94.000	95.400
	β_z	Bias	0.352	0.352	0.385	0.388	0.349	0.354	0.354	0.388	0.393	0.340
		SD	6.724	6.724	6.678	7.542	8.040	6.714	6.714	6.672	7.592	8.344
		ASE	6.346	6.336	6.356	7.185	7.857	6.346	6.336	6.358	7.255	8.212
		MSE	0.453	0.453	0.447	0.570	0.648	0.452	0.452	0.447	0.578	0.697
		CP	93.600	93.600	93.600	93.200	95.200	94.000	94.200	93.200	93.400	95.000
								$\mathrm{P}(\eta = 1) = 0.7$				
0.3	β_x	Bias	−13.050	−4.525	−4.066	−0.575	0.162	−18.092	−5.130	−4.582	−2.010	−0.009
		SD	6.225	7.219	8.051	8.584	12.582	5.897	7.511	8.604	8.805	13.914
		ASE	6.326	7.100	7.993	8.693	12.128	6.027	7.326	8.567	8.999	13.515
		MSE	2.090	0.726	0.813	0.740	1.583	3.621	0.827	0.950	0.816	1.936
		CP	45.090	89.780	92.986	94.389	93.387	16.600	87.200	93.800	94.000	94.188
	β_z	Bias	0.319	0.319	0.291	0.331	0.269	0.317	0.317	0.288	0.336	0.257
		SD	6.804	6.804	6.742	7.326	7.813	6.793	6.793	6.732	7.394	8.139
		ASE	6.346	6.336	6.352	6.917	7.617	6.346	6.336	6.353	6.995	7.982
		MSE	0.464	0.464	0.455	0.538	0.611	0.462	0.462	0.454	0.548	0.663
		CP	93.186	92.585	93.788	92.986	93.788	93.800	93.000	93.600	93.200	93.600

0.5	β_x	Bias	−14.668	−6.386	−5.127	−0.822	0.038	−19.522	−6.923	−5.543	−2.353	−0.249
		SD	6.155	7.121	8.259	8.939	13.070	5.817	7.388	8.841	9.098	14.372
		ASE	6.301	7.070	8.328	9.063	12.619	6.006	7.301	8.956	9.330	13.985
		MSE	2.530	0.915	0.945	0.806	1.708	4.149	1.025	1.089	0.883	2.066
		CP	35.200	83.800	91.800	95.400	93.600	10.800	82.600	93.000	93.800	93.600
	β_z	Bias	0.352	0.352	0.300	0.361	0.298	0.354	0.354	0.302	0.367	0.281
		SD	6.724	6.724	6.653	7.460	7.959	6.714	6.714	6.648	7.513	8.268
		ASE	6.346	6.336	6.352	7.107	7.776	6.346	6.336	6.354	7.168	8.117
		MSE	0.453	0.453	0.444	0.558	0.634	0.452	0.452	0.443	0.566	0.684
		CP	93.600	93.600	93.000	93.200	95.200	94.000	94.200	93.600	93.000	95.000

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; MSE is the mean squared errors; CP represents the coverage probability of the 95% confidence intervals. The results are presented in the 10^{-2} scale.

Table 2: Simulation results for the estimation of $\boldsymbol{\beta}$: $n = 1000$; $\lambda(u) = \exp(\beta_x X + \beta_z Z)$, $X = L + U_b$; $L \sim N(0{,}1)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $W_j = L + U_{c,j}$; $U_{c,j} \sim N(0, \sigma_c^2)$ is the classical error, $j = 1{,}2$; $\boldsymbol{\beta} = (\log(2), 0)^{'}$; $Q = 1 + 0.8X + \epsilon$; $\varepsilon \sim N(0{,}0.5)$; $p(\eta = 1)$ is the proportion of the calibration data; $Z \sim$ *Bernoulli*(0.5); censoring rate is 50%; Naive, “naive” estimator; RC1, RC replacing X by $\mathrm{E}(X|\bar{W})$; RC2, RC replacing X by $\{\mathrm{E}(X|\bar{W}, Q)\}^{\eta}\{\mathrm{E}(X|\bar{W})\}^{1-\eta}$; SIM1, SIMEX using quadratic extrapolation function; SIM2, SIMEX using polynomial of degree 4 as extrapolation function.

σ_b^2	β						$\sigma_c^2 = 0.3$					$\sigma_c^2 = 0.5$
			Naive	**RC1**	**RC2**	**SIM1**	**SIM2**	**Naive**	**RC1**	**RC2**	**SIM1**	**SIM2**
						$\mathrm{P}(\eta = 1) = 0.5$						
0.3	β_x	Bias	−12.851	−4.364	−3.988	−0.286	0.412	−17.859	−4.956	−4.484	−1.703	0.272
		SD	4.335	5.023	5.342	6.152	8.675	4.171	5.319	5.840	6.372	9.693
		ASE	4.441	4.951	5.453	6.205	8.641	4.233	5.106	5.823	6.410	9.627
		MSE	1.839	0.443	0.444	0.379	0.754	3.364	0.529	0.542	0.435	0.940
		CP	17.234	85.772	85.772	94.389	94.990	1.600	82.000	87.000	93.200	94.400
	β_z	Bias	0.119	0.119	0.128	0.157	0.234	0.113	0.113	0.120	0.164	0.251
		SD	4.455	4.455	4.490	4.789	5.134	4.470	4.470	4.504	4.842	5.351
		ASE	4.482	4.484	4.492	4.882	5.362	4.481	4.483	4.492	4.930	5.610
		MSE	0.199	0.199	0.202	0.230	0.264	0.200	0.200	0.505	0.235	0.287
		CP	94.790	94.188	94.188	94.389	95.792	94.400	94.000	94.200	94.800	96.000
0.5	β_x	Bias	−14.510	−6.273	−5.316	−0.418	0.355	−19.339	−6.806	−5.727	−1.942	0.193
		SD	4.355	5.047	5.597	6.509	9.185	4.197	5.351	6.144	6.722	10.230
		ASE	4.423	4.924	5.664	6.525	9.043	4.218	5.083	6.077	6.711	10.028
		MSE	2.295	0.648	0.596	0.425	0.845	3.916	0.750	0.705	0.490	1.047
		CP	9.000	72.800	86.200	95.000	95.400	0.600	69.800	86.400	92.800	94.800
	β_z	Bias	0.103	0.103	0.120	0.141	0.211	0.094	0.094	0.108	0.147	0.229
		SD	4.473	4.473	4.508	4.952	5.268	4.486	4.486	4.521	4.992	5.465
		ASE	4.482	4.483	4.491	5.019	5.464	4.481	4.482	4.491	5.055	4.695
		MSE	0.200	0.200	0.203	0.245	0.278	0.201	0.201	0.205	0.249	0.299
		CP	94.600	95.000	94.600	94.800	95.600	95.000	94.600	94.400	95.600	95.800
							$\mathrm{P}(\eta = 1) = 0.7$					
0.3	β_x	Bias	−12.837	−4.351	−3.540	−0.400	0.281	−17.859	−4.956	−3.958	−1.839	0.110
		SD	4.333	5.022	5.313	5.965	8.553	4.171	5.319	5.801	6.196	9.561
		ASE	4.442	4.952	5.471	6.057	8.492	4.233	5.106	5.840	6.250	9.453
		MSE	1.836	0.441	0.408	0.357	0.732	3.364	0.529	0.493	0.418	0.914
		CP	17.234	85.772	90.581	94.589	94.990	1.600	82.000	89.800	92.400	94.400
	β_z	Bias	0.122	0.122	0.089	0.161	0.233	0.113	0.113	0.082	0.165	0.249
		SD	4.457	4.457	4.442	4.763	5.109	4.470	4.470	4.452	4.816	5.326
		ASE	4.482	4.484	4.492	4.863	5.346	4.481	4.483	4.491	4.910	5.590
		MSE	0.199	0.199	0.197	0.227	0.262	0.200	0.200	0.198	0.232	0.284
		CP	94.790	94.188	95.190	94.790	95.792	94.400	94.000	94.800	95.200	96.400

0.5	β_x	Bias	−14.510	−6.273	−4.626	−0.528	0.235	−19.339	−6.807	−4.945	−2.060	0.060
		SD	4.355	5.046	5.515	6.277	8.994	4.197	5.351	6.051	6.500	10.024
		ASE	4.423	4.924	5.702	6.344	8.866	4.218	5.083	6.113	6.510	9.819
		MSE	2.295	0.648	0.518	0.397	0.810	3.916	0.750	0.611	0.455	1.005
		CP	9.000	72.800	89.600	95.600	95.600	0.600	69.600	88.800	93.400	94.800
	β_z	Bias	0.103	0.103	0.052	0.142	0.209	0.094	0.094	0.045	0.146	0.225
		SD	4.473	4.473	4.476	4.919	5.238	4.486	4.486	4.488	4.958	5.435
		ASE	4.482	4.483	4.492	4.999	5.447	4.481	4.482	4.491	5.034	5.674
		MSE	0.200	0.200	0.200	0.242	0.275	0.201	0.201	0.201	0.246	0.296
		CP	94.600	95.000	95.000	94.800	95.600	95.000	94.600	94.800	94.800	96.000

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; MSE is the mean squared errors; CP represents the coverage probability of the 95% confidence intervals. The results are presented in the 10^{-2} scale.

The simulation results for the estimation of the nuisance parameters by the method of moments are presented in Table 3 for the setting with $\sigma_b^2 = 0.5$ and $\sigma_c^2 = 0.5$. The estimators generally show small bias and acceptable coverage probabilities. Also, it can be noted that the standard errors for the nuisance parameters involved in the distribution of the instrumental variable get smaller as the size of the calibration subsample increases. The biases and ASE's of the estimators decrease as the sample size increases.

Table 3: Simulation results for the estimation of the nuisance parameters: $\lambda(u) = \exp(\beta_x X + \beta_z Z)$, $X = L + U_b$; $L \sim N(0{,}1)$; $U_b \sim N(0, \sigma_b^2)$ is the Berkson error; $W_j = L + U_{c,j}$; $U_{c,j} \sim N(0, \sigma_c^2)$ is the classical error, $j = 1{,}2$; $\sigma_b^2 = 0.5$; $\sigma_c^2 = 0.5$; $\boldsymbol{\beta} = (\log(2), 0)^{'}$; $Q = 1 + 0.8X + \epsilon$; $\varepsilon \sim N(0{,}0.5)$; $p(\eta = 1)$ is the proportion of the calibration data; $Z \sim \mathit{Bernoulli}(0.5)$; censoring rate is 50%.

	$n = 500$							$n = 1000$						
	μ_l	α_0	α_1	σ_l^2	σ_b^2	σ_c^2	σ_ϵ^2	μ_l	α_0	α_1	σ_l^2	σ_b^2	σ_c^2	σ_ϵ^2
							P($\eta = 1$) = 0.5							
Bias	0.275	−0.345	−0.231	−0.422	8.841	0.010	−3.130	0.119	−0.296	0.359	−0.088	2.419	−0.083	−0.854
SD	4.862	6.897	9.783	8.289	36.235	3.259	16.555	3.409	4.984	6.781	5.682	24.974	2.220	12.437
ASE	4.989	7.003	9.541	8.017	36.459	3.157	17.686	3.533	4.955	6.762	5.695	23.800	2.226	11.876
MSE	0.237	0.477	0.958	0.689	13.911	0.106	2.839	0.116	0.249	0.461	0.323	6.295	0.049	1.554
CP	95.000	96.200	93.000	94.800	94.000	95.000	97.000	96.200	94.400	95.800	94.200	94.000	95.600	94.400
							P($\eta = 1$) = 0.7							
Bias	0.275	−0.052	0.172	−0.418	5.314	0.009	−2.745	0.119	0.014	0.128	−0.085	1.562	−0.084	−0.844
SD	4.862	5.724	7.850	8.290	30.977	3.259	14.627	3.409	4.059	5.324	5.682	20.041	2.221	10.285
ASE	4.989	5.660	7.663	8.017	28.925	3.157	14.843	3.533	4.010	5.403	5.695	19.269	2.226	10.012
MSE	0.237	0.328	0.616	0.689	9.878	0.106	2.215	0.116	0.165	0.284	0.323	4.041	0.049	1.065
CP	95.000	93.400	94.800	94.800	94.200	95.000	96.600	96.200	94.200	94.200	94.200	94.800	95.600	95.000

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; MSE is the mean squared errors; CP represents the coverage probability of the 95% confidence intervals. The results are presented in the 10^{-2} scale.

In addition, we conducted some simulations to investigate the performances of the methods when X_i is not normally distributed. The simulation settings were similar to those in Table 1 with the exception that L_i was simulated from gamma distribution $G(1{,}1)$, $U_{b,i}$ followed the uniform $U(-\sigma_b\sqrt{3}, \sigma_b\sqrt{3})$, $\sigma_b^2 = 0.5$, $\sigma_c^2 = 0.5$ and $P(\eta_i = 1) = 0.5$. We estimated the parameter $\boldsymbol{\beta}$ by all the methods under the correctly specified or misspecified models for L_i and $U_{b,i}$ to gain insight into potential effects of models misspecification on the performances of the methods. The misspecified model for $U_{b,i}$ assumed normality for $U_{b,i}$ even though this assumption did not hold. The vector of nuisance parameters was estimated using the method of moments. The results of these simulations are reported in Table 4. The performances of the methods under the correctly specified models are similar to those observed in Table 1 and Table 2 with the SIMEX methods showing smaller biases and better coverage probabilities than the RC methods. Also, the SIMEX estimators show similar performances under both correct and misspecified models situations, suggesting that the method may not be very affected by a misspecification of the distribution for the latent variable or the measurement error.

Table 4: Results of simulation assessing the effects of a misspecified distribution for the Berkson error: $\lambda(u) = \exp(\beta_x X + \beta_z Z)$, $X = L + U_b$; $L \sim G(1,1)$; $U_b \sim U(-\sigma_b\sqrt{3}, \sigma_b\sqrt{3})$ is the Berkson error; $W_j = L + U_{c,j}$; $U_{c,j} \sim N(0, \sigma_c^2)$ is the classical error, $j = 1,2$; $\boldsymbol{\beta} = (\log(2), 0)'$; $Q = 1 + 0.8X + \epsilon$; $\varepsilon \sim N(0,0.5)$; $Z \sim$ *Bernoulli*(0.5); censoring rate is 50%; $\sigma_b^2 = 0.5$; $\sigma_c^2 = 0.5$; $p(\eta = 1) = 0.5$; Naive, "naive" estimator; RC1, RC replacing X by $\mathrm{E}(X|\bar{W})$; RC2, RC replacing X by $\{\mathrm{E}(X|\bar{W}, Q)\}^{\eta}\{\mathrm{E}(X|\bar{W})\}^{1-\eta}$; SIM1, SIMEX using quadratic extrapolation function; SIM2, SIMEX using polynomial of degree 4 as extrapolation function.

n	$\boldsymbol{\beta}$		Correct model					Misspecified model				
			Naive	**RC1**	**RC2**	**SIM1**	**SIM2**	**Naive**	**RC1**	**RC2**	**SIM1**	**SIM2**
500	β_x	Bias	−16.233	−6.988	−5.931	−0.319	0.543	−16.233	−2.569	−2.648	−0.202	0.823
		SD	5.923	6.945	7.916	9.202	13.333	5.923	7.555	8.962	9.258	13.589
		ASE	5.523	5.989	6.522	8.389	12.409	5.523	6.506	8.496	8.279	12.198
		MSE	2.986	0.971	0.978	0.848	1.781	2.986	0.637	0.873	0.857	1.853
		CP	15.102	74.694	79.388	91.224	93.878	15.102	88.163	94.898	90.612	93.673
	β_z	Bias	−0.138	−0.142	−0.100	−0.126	−0.024	−1.38	−0.138	−0.104	−0.129	−0.033
		SD	6.334	6.369	6.299	7.080	8.094	6.334	6.334	6.290	7.063	8.084
		ASE	6.359	6.377	6.465	7.259	8.205	6.359	6.367	6.410	7.233	8.181
		MSE	0.401	0.406	0.397	0.501	0.655	0.401	0.401	0.396	0.499	0.653
		CP	95.306	95.102	94.490	95.918	94.490	95.306	95.510	94.490	96.122	94.490
1000	β_x	Bias	−16.484	−7.462	−5.376	−0.915	−0.237	−16.484	−2.993	−2.041	−0.899	−0.214
		SD	4.016	4.756	5.240	6.067	9.175	4.024	5.175	5.894	6.066	9.179
		ASE	3.876	4.155	4.340	5.972	8.720	3.876	4.530	5.688	5.957	8.705
		MSE	2.878	0.783	0.564	0.376	0.842	2.879	0.357	0.389	0.376	0.843
		CP	1.000	53.600	70.800	93.200	93.000	1.004	85.141	93.373	92.972	93.173
	β_z	Bias	0.007	−0.011	0.067	0.018	−0.153	−0.007	−0.007	0.093	−0.003	−0.177
		SD	4.485	4.524	4.505	4.932	5.661	4.485	4.485	4.462	4.934	5.672
		ASE	4.482	4.488	4.490	5.004	5.634	4.482	4.485	4.493	4.994	5.625
		MSE	0.201	0.205	0.203	0.243	0.321	0.201	0.201	0.199	0.243	0.322
		CP	94.800	94.600	94.400	94.200	94.400	94.779	94.779	94.980	94.378	93.976

Note: SD denotes the sample standard deviation of the estimates; ASE is the average of the estimated standard errors; MSE is the mean squared errors; CP represents the coverage probability of the 95% confidence intervals. The results are presented in the 10^{-2} scale.

6 Application

In this section, we apply the estimation methods to the analysis of the data from the ACTG 175 study, which was designed to compare Zidovudine (*ZDV*) alone, Zidovudine combined with Didanosine (*ZDV* + *ddI*), Zidovudine combined with Zalcitabine (*ZDV* + *ddC*) and Didanosine (*ddI*) alone in HIV-1 infected patients. A detailed description of the data can be found in [25]. These data have been previously analyzed by a number of authors under different settings [16, 17]. In this application, we considered the subset of participants with two CD4 cell counts taken prior to the start of treatment and within three weeks of randomization. There were in total 2441 subjects in the analyzed dataset and only 1459 of them had available CD8 cell counts taken at randomization. Also, the median follow-up time for the participants was 144.3 weeks with an interquartile range of 42.2 weeks. There was a total of 306 events, consisting of death or AIDS disease development during follow-up. The two CD4 cell counts, measured before the beginning of treatment and within three weeks of randomization were treated as replicates for the baseline CD4 cell counts. Furthermore, CD4 and CD8 measurements were log-transformed for variance stabilization and approximate normality achievement. The scatter plot of the first replicate versus the second replicate of the baseline $\log_{10}(1 + CD4)$ and the scatter plot of the $\log_{10}(1 + CD8)$ against each of these replicates in the calibration subsample are shown in the upper panel of Figure 1, which also displays the qq-plots of these variables in its lower panel.

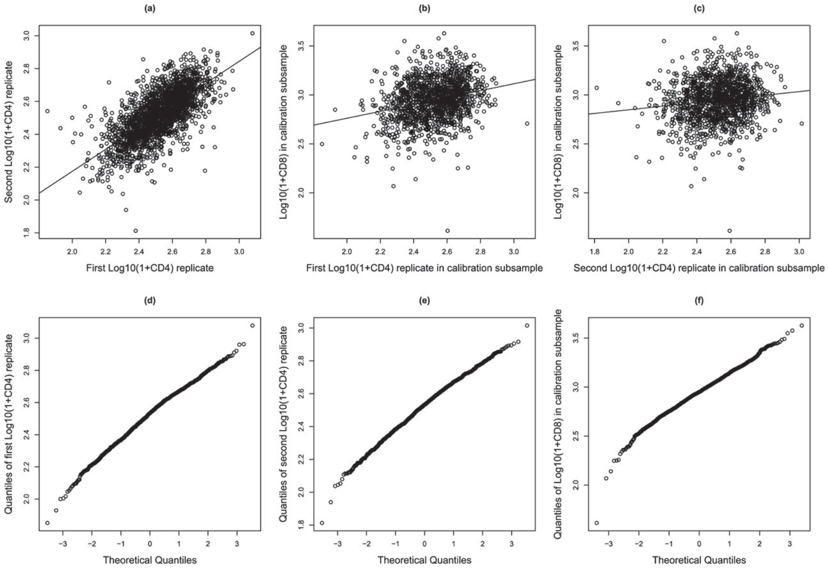

Figure 1: Scatter plots and qq-plots of duplicates of baseline $\log_{10}(1 + CD4)$ and $\log_{10}(1 + CD8)$ in calibration sample. **(a)** scatter plot of first replicate versus second replicate for baseline $\log_{10}(1 + CD4)$. **(b)** Scatter plot of first $\log_{10}(1 + CD4)$ replicate versus $\log_{10}(1 + CD8)$ in calibration. **(c)** Scatter plot of second $\log_{10}(1 + CD4)$ replicate versus $\log_{10}(1 + CD8)$ in calibration. **(d)** qq-plot of first $\log_{10}(1 + CD4)$ replicate. **(e)** qq-plot of second $\log_{10}(1 + CD4)$ replicate. **(f)** qq-plot of $\log_{10}(1 + CD8)$ in calibration subsample.

Our interest in this application was in assessing the association between the true baseline $\log_{10}(1 + CD4)$ and the time to progression to AIDS disease or death adjusting for the treatment effect in the following proportional hazards model:

$$\lambda(u) = \lambda_0(u)\exp(\beta_X X_i + \boldsymbol{\beta}_Z^{'}\mathbf{Z}_i),$$

where X_i is the true baseline $\log_{10}(1 + CD4)$, $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, Z_{3i})^{'}$, Z_{1i}, Z_{2i} and Z_{3i} represent the indicator variables for *ZDV* + *ddI*, *ZDV* + *ddC* and *ddI*, respectively, for subject i, $\lambda_0(.)$ is an unspecified baseline hazards function, β_X and $\boldsymbol{\beta}_Z = (\beta_{Z1}, \beta_{Z2}, \beta_{Z3})^{'}$ are the regression coefficients to be estimated. Moreover, we coded the first and second replicates for baseline $\log_{10}(1 + CD4)$ as W_{i1} and W_{i2}, respectively, and denoted the $\log_{10}(1 + CD8)$ measured at randomization by Q_i for subject i. The measurement error and Q_i were modeled as in (1) and (2), respectively. We estimated the main parameter $\boldsymbol{\beta} = (\beta_X, \boldsymbol{\beta}_Z^{'})^{'}$ by the RC and SIMEX methods. The implementations of the methods were done similarly as in the simulation study and assumed normality for the measurement errors. The vector of nuisance parameters $\boldsymbol{\nu}$ was estimated using the method of moments.

The results of the analysis are reported in Table 5. The estimates of the nuisance parameters by the method of moments are shown in the lower portion of this table. They suggest that the error involved in the measurements of the true baseline $\log_{10}(1 + CD4)$ is essentially of the classical type. Furthermore, the results of the estimation of β_X by all the methods in the upper part of the table indicate that the $\log_{10}(1+CD4)$ is significantly associated with the time to AIDS disease or death. Larger values of CD4 cell counts lead to lower risk of death or AIDS development. The estimates obtained based on the RC and SIMEX methods show stronger $\log_{10}(1+CD4)$ effect than the naive method. In addition, it appears from the estimation of $\boldsymbol{\beta}_Z$ that Didanosine and the combination of Zidovudine and Didanosine treatments are significantly different from Zidovudine in terms of their effects on time to progression to AIDS or death.

Table 5: Results of the ACTG 175 data analysis: Naive, "naive" estimator; RC1, RC not using data in calibration subsample; RC2, RC making use of calibration subsample data; SIM1, SIMEX using quadratic extrapolation function; SIM2, SIMEX using polynomial of degree 4 as extrapolation function.

$\boldsymbol{\beta}$			**RC**		**SIMEX**			
		Naive	**RC1**	**RC2**	**SIM1**	**SIM2**		
β_X	Estimate	−4.5310	−5.5057	−5.4179	−5.4011	−5.1889		
	SE	0.4086	0.4618	0.6690	0.6163	0.8546		
	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001		
β_{Z1}	Estimate	−0.5114	−0.5114	−0.5095	−0.5330	−0.5282		
	SE	0.1613	0.1612	0.1616	0.1783	0.1907		
	p-value	0.0015	0.0015	0.0016	0.0028	0.0056		
β_{Z2}	Estimate	−0.2950	−0.2950	−0.2934	−0.3044	−0.2978		
	SE	0.1540	0.1539	0.1537	0.1788	0.1933		
	p-value	0.0554	0.0552	0.0562	0.0886	0.1234		
β_{Z3}	Estimate	−0.4321	−0.4321	−0.4343	−0.4424	−0.4427		
	SE	0.1576	0.1575	0.1578	0.1888	0.1981		
	p-value	0.0061	0.0061	0.0059	0.0191	0.0254		
	$\boldsymbol{\nu}$	μ_l	α_0	α_1	σ_l^2	σ_b^2	σ_c^2	σ_ϵ^2
	Estimate	2.5243	1.8035	0.4526	0.0150	10^{-5}	0.0065	0.0386
	SE	0.0027	0.2141	0.0847	0.0005	0.0136	0.0003	0.0030
	p-value	<0.0001	<0.0001	<0.0001	<0.0001	0.9996	<0.0001	<0.0001

β_X is the coefficient of $\log_{10}(1+CD4)$; β_{Z1}, β_{Z2} and β_{Z3} are the coefficients of the indicator variables for $ZDV + ddI$, $ZDV + ddC$ and ddI, respectively; SE means standard error and p-value is the Wald test-based p-value.

7 Conclusion

We have developed a new estimation method for Cox proportional hazards models with some covariates subject to a mixture of Berkson and classical measurement errors. The proposed method is based on the simulation extrapolation algorithm and leads to a consistent estimator of the vector of parameters of the proportional hazards model provided that the true extrapolation function is known. In practice, polynomial functions can be used to approximate this function. In addition, the variances of the classical and Berkson errors are estimated using the method of moments with no assumption about the mixture percentage of the error variances. The method makes use of two replicates of the error-prone covariates and data on an instrumental variable that are available for some subjects in a calibration subsample only.

Extensive simulations have shown that the proposed SIMEX method outperforms the regression calibration method with regard to biases and coverage probabilities in many finite sample size situations. A sensitivity analysis has revealed that the method may perform satisfactorily when the distribution of the Berkson error is misspecified.

Acknowledgements

This research was supported in parts by the National Institute of Health grants R01ES017030 (Wang and Tapsoba), HL121347 (Wang and Tapsoba) and GM100573 (Chao, Wang and Tapsoba).

References

[1] Cox DR. Regression models and life tables (with discussion). J R Stat Soc Ser B. 1972;34:187–220.

[2] Tsiatis AA, DeGruttola V, Wulfsohm MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. J Am Stat Assoc. 1995;90:27–37.

[3] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Nonlinear measurement error models, a modern perspective, 2nd ed. London: Chapman and Hall, 2006.

[4] Prentice RL. Covariate measurement errors and parameter estimates in failure time regression. Biometrika. 1982;69:331–42.

[5] Wang CY. Robust best linear estimation for regression analysis using surrogate and instrumental variables. Biostatistics. 2012;13:326–40. DOI: 10.1093/biostatistics/kxr051.

[6] Kim HY, Yasui Y, Burtyn I. Attenuation in rsik-estimation in logistic and Cox proportional hazards models due to grouped-based exposure assessment strategy. Ann Occup Hyg. 2006;50:623–35.

[7] Schafer DW, Gilbert ES. Some statistical implications of dose uncertainty in radiation dose response analyses. Radiat Res. 2006;166:303–12.

[8] Huang Y, Wang CY. Cox regression with accurate covariates unascertainable: a nonparametric correction approach. J Am Stat Assoc. 2000;95:1209–19

[9] Huges MD. Regression dilution in the proportional hazards model. Biometrics. 1993;49:1056–66.

[10] Kuchenhoff H, Bender R, Langner I. Effect of Berkson measurement error on parameter estimates in Cox regression models. Lifetime Data Anal. 2007;13:261–272.

[11] Wang CY. Robust sandwich covariance estimation for regression estimator in Cox regression with measurement error. Stat Probab Lett. 1999;45:371–8.

[12] Wang CY, Wang N, Wang S. Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. Biometrics. 2000;56:487–95.

[13] Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc. 1994;89:1314–28.

[14] Wang CY, Huang Y. Error in timing regression with observed longitudinal measurements. Stat Med. 2003;22:2577–90. DOI: 10.1002/sim.1435.

[15] Nakamura T. Proportional hazards model with covariates subject to measurement error. Biometrics. 1992;48:829–38.

[16] Song X, Wang CY. Proportional hazards model with covariates measurement error and instrumental variables. J Am Stat Assoc. 2014. DOI: 10.1080/01621459.2014.896805.

[17] Tapsoba JD, Wang CY, Lee SM. Joint modeling of survival time and longitudinal data with subject-specific change points in the covariates. Stat Med. 2011;30:232–49. DOI: 10.1002/sim.4107.

[18] Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika. 2001;88:447–58.

[19] Hu P, Tsiatis AA, Davidian M. Estimating the parameters in the Cox model when covariate variables are measured with error. Biometrics. 1998;54:1407–19.

[20] Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. Biometrics. 2002;58:742–53.

[21] Wang CY. Nonparametric maximum likelihood estimator for Cox regression with subject-specific measurement error. Scand J Stat. 2008;35:613–28.

[22] Yan Y, Yi GY. A corrected profile likelihood method for survival data with covariate measurement error under the Cox model. Canad J Stat. 2015;43:454–80.

[23] Li Y, Guolo A, Hoffman FO, Carroll RJ. Shared uncertainty in measurement error problems, with application to Nevada Test Site fallout data. Biometrics. 2007;63:1226–36. DOI: 10.1111/j.1541-0420.2007.00810.x.

[24] Tapsoba JD, Wang CY, Lee SM. Expected estimating equation using calibration data for generalized linear models with a mixture of Berkson and classical errors in covariates. Stat Med. 2014;33:675–92. DOI: 10.1002/sim.5966.

[25] Hammer SM, Katzenstein DA, Huges MD, Gundacker H, Schooley RT, Haubrich MR, Henry WK, Lederman MM, Phair JP, Niu M, Hirch 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. N Engl J Med. 1996;335:1081–90.

[26] 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.

[27] Carroll RJ, Delaigle A, Hall P. Non-parametric regression estimation from data contaminated by a mixture of Berkson and classical errors. J R Stat Soc., Ser B. 2007;69:859–78. DOI: 10.1111/j.1467-9868.2007.00614.x.

[28] Wang CY. Corrected score estimator for joint modeling of longitudinal and failure time data. Stat Sin. 2006;16:235–53.

[29] Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. Biometrics. 1997;53:131–45.

[30] Whitemore AS, Keller JB. Approximations for regression with covariate measurement error. J Am Stat Assoc. 1988;83:1057–66.

[31] Apanasovich TV, Carroll RJ, Maity A. SIMEX and standard error estimation in semiparametric measurement error models. Electron J Stat. 2009;3:318–48. DOI: 10.1214/08-EJS341.

[32] Carroll RJ, Kuchenhoff H, Lombard F, Stefanski LA. Asymptotics for SIMEX estimator in nonlinear measurement error models. J Am Stat Assoc. 1996;91:242–50. DOI: 10.1111/j.1467-9868.2007.00614.x.