Received: 17 March 2020 | Revised: 5 August 2020 | Accepted: 11 October 2020

DOI: 10.1002/sim.8793

RESEARCH ARTICLE

Statistics in Medicine WILEY

Raking and regression calibration: Methods to address bias from correlated covariate and time-to-event error

Eric J. Oh[1] | Bryan E. Shepherd[2] | Thomas Lumley[3] | Pamela A. Shaw[1]

[1]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Department of Biostatistics, Vanderbilt University, Nashville, Tennessee, USA

[3]Department of Statistics, University of Auckland, Auckland, New Zealand

Correspondence
Eric J. Oh, 423 Guardian Drive,
Philadelphia, PA 19104.
Email: ericoh@pennmedicine.upenn.edu

Funding information
National Institutes of Health, Grant/Award Numbers: P30-AI110527, R01-AI093234, R01-AI131771, U01-AI069918, U01-AI069923; Patient-Centered Outcomes Research Institute, Grant/Award Number: R-1609-36207

Medical studies that depend on electronic health records (EHR) data are often subject to measurement error, as the data are not collected to support research questions under study. These data errors, if not accounted for in study analyses, can obscure or cause spurious associations between patient exposures and disease risk. Methodology to address covariate measurement error has been well developed; however, time-to-event error has also been shown to cause significant bias, but methods to address it are relatively underdeveloped. More generally, it is possible to observe errors in both the covariate and the time-to-event outcome that are correlated. We propose regression calibration (RC) estimators to simultaneously address correlated error in the covariates and the censored event time. Although RC can perform well in many settings with covariate measurement error, it is biased for nonlinear regression models, such as the Cox model. Thus, we additionally propose raking estimators which are consistent estimators of the parameter defined by the population estimating equation. Raking can improve upon RC in certain settings with failure-time data, require no explicit modeling of the error structure, and can be utilized under outcome-dependent sampling designs. We discuss features of the underlying estimation problem that affect the degree of improvement the raking estimator has over the RC approach. Detailed simulation studies are presented to examine the performance of the proposed estimators under varying levels of signal, error, and censoring. The methodology is illustrated on observational EHR data on HIV outcomes from the Vanderbilt Comprehensive Care Clinic.

KEYWORDS

calibration, electronic health records, measurement error, misclassification, raking, survival analysis

1 | INTRODUCTION

Biomedical research relies increasingly on electronic health records (EHR) data, either as the sole or supplemental source of data, due to the vast amount of data these resources contain and their relatively low cost compared with prospectively collected data. However, EHR data and other large cohort databases have been observed to be error-prone. These errors, if not accounted for in the data analysis, can bias associations of patient exposures and disease risk. There exists a large body of literature describing the impact of and methods to correct for covariate measurement error;[1] however, much less attention has been given to errors in the outcome. For linear models, independent random (classical) errors in the outcome variable do not bias regression estimates; however, errors correlated with either predictors in the model or errors in those predictors could bias associations. For nonlinear models, even classical outcome errors can bias estimated associations of interest.[1] There are many examples in clinical research where the outcome of interest relies on an imprecisely measured event time. Researchers studying the epidemiology of chronic conditions may enroll subjects sometime after an initial diagnosis, and so research questions focused on the timing of events post diagnosis may need to rely on patient recall or chart review of electronic medical records for the date of diagnosis, both of which are subject to error. Errors in the time origin can be systematic, as subject characteristics can influence the amount of error in recall. Methods to handle a misclassified outcome have been developed for binary outcomes[2-4] and discrete failure time data,[5-7] where estimates of sensitivity and specificity can be incorporated into the bias correction. However, methods to handle errors in a continuous failure time have largely been ignored.

Additionally, as more and more observational studies utilize data primarily collected for nonresearch purposes (eg, administrative databases or electronic health records), it is increasingly common to have errors in both the outcome and exposures that are correlated. For example, in some observational studies of HIV/AIDS, the date of antiretroviral therapy (ART) initiation has been observed to have substantial errors.[8,9] These errors can lead to errors in event times, defined as time since ART initiation, and errors in exposures of interest, such as CD4 count at ART initiation. Furthermore, certain types of records are often more likely to have errors (eg, records from a particular study site), records with errors often tend to have errors across multiple variables, and the magnitude of these errors cannot be assumed uncorrelated. Ignoring correlated outcome and exposure errors could lead to positive or negative bias in estimates of regression parameters.

In some settings, data errors can be corrected by retrospectively reviewing and validating medical records; however, this is expensive and time-consuming to do for a large number of records. Instead, we can perform data validation on a subset of selected records and use this information to correct estimates based on the larger, unvalidated data set. In this article, we propose regression calibration and raking estimators as two methods to correct the bias induced from such correlated errors by incorporating information learned in a validation subset to the large unvalidated data set.

Regression calibration (RC), introduced by Prentice,[10] is a method to address covariate measurement error that is widely used due to ease of implementation and good numerical performance in a wide range of settings. Although most RC methods assume measurement error in covariates only, Shaw et al[11] examined a way to apply RC to correlated errors in a covariate and a continuous outcome; to date these methods have not addressed correlated errors between failure time outcomes and exposures.

Raking is a method in survey sampling that makes use of auxiliary information available on the population to improve upon the Horvitz-Thompson (HT) estimator for regression parameters in two-phase designs. The HT estimator is known to be inefficient[12] but raking improves statistical efficiency, without changing the target of inference, by adjusting the standard HT weights by tuning them to auxiliary variables. Raking also takes advantage of the known sampling probabilities with validation studies such as those considered in this article. These survey sampling ideas, while not new, have not been carefully studied in the measurement error setting. Breslow et al[13] considered raking estimators for modeling case-cohort data with missing covariates. Lumley et al[14] considered a raking estimator using simulated data in a covariate measurement error context with a validation subset. In this article, we consider raking estimators for more general settings allowing for errors in the covariate and a time-to-event outcome, including misclassification, and discuss various possibilities for the auxiliary variables, how different choices affect the degree of improvement over the HT estimator, and ways to implement these methods using standard statistical software.

Our contributions in this article are 2-fold. First, we develop regression calibration estimators to address both censored event time error alone and correlated covariate and censored event time errors together. To our knowledge, no RC estimators have been developed for these settings. Second, we develop raking estimators that are consistent and, in some settings, improve upon the RC estimators. These methods are important given the increased use of error-prone data in biomedical research and the paucity of methods that simultaneously handle errors in covariates and times-to-event. The rest of the article proceeds as follows. We present our survival time model and the considered measurement error frameworks in Section 2. Sections 3 and 4 present the proposed regression calibration and raking methods, respectively. Section 5 compares the relative performance of the proposed estimators with simulation studies for various parameter settings and error distributions. In Section 6, we apply our methods to an HIV cohort and ascertain their robustness to misclassification. We conclude with a discussion in Section 7.

2 | TIME-TO-EVENT MODEL AND ERROR FRAMEWORK

We consider the Cox proportional hazards model. Let T_i and C_i, be the failure time and right censoring time, respectively, for subjects $i = 1, \ldots, n$ on a finite follow-up time interval, $[0, \tau]$. Define $U_i = \min(T_i, C_i)$ and the corresponding failure indicator $\Delta_i = I(T_i \leq C_i)$. Let $Y_i(t) = I(U_i \geq t)$ and $N_i(t) = I(U_i \leq t, \Delta_i = 1)$ denote the at-risk indicator and counting process for observed events, respectively. Let X_i be a p-dimensional vector of continuous covariates that are measured with error and Z_i a q-dimensional vector of precisely measured discrete and/or continuous covariates that may be correlated with X_i. We assume C_i is independent of T_i given (X_i, Z_i) and that (T_i, C_i, X_i, Z_i) are i.i.d. Let the hazard rate for subject i at time t be given by $\lambda_i(t) = \lambda_0(t)\exp(\beta_X' X_i + \beta_Z' Z_i)$, where $\lambda_0(t)$ is an unspecified baseline hazard function. We consider β_X to be the parameter(s) of interest, which is estimated by solving the partial likelihood score for $\beta = (\beta_X, \beta_Z)$.

$$\sum_{i=1}^{n} \int_0^{\tau} \left\{ \{X_i, Z_i\}' - \frac{n^{-1} \sum_{j=1}^{n} Y_j(t) \{X_j, Z_j\}' \exp(\beta_X' X_j + \beta_Z' Z_j)}{n^{-1} \sum_{j=1}^{n} Y_j(t) \exp(\beta_X' X_j + \beta_Z' Z_j)} \right\} dN_i(t) = 0. \tag{1}$$

2.1 | Additive measurement error structure

Oftentimes, errors seen in electronic health records data or other data sets used for observational studies will not be simple random error and will depend on other variables in the data set. For example, when the time-to-event error is due to a mismeasured time origin, this timing error can cause correlated errors in the baseline observations for exposures that are associated with the true survival outcome. In addition, errors induced in the exposures and censored time-to-event outcome can vary systematically with subject characteristics that could make a subject's record more error-prone. Thus, we consider the error setting involving additive systematic and random error in both the covariates and time-to-event.

Instead of observing (X, Z, U, Δ), we observe $(X^\star, Z, U^\star, \Delta)$, where

$$X^\star = \alpha_0 + \alpha_1' X + \alpha_2' Z + \epsilon, \tag{2}$$

$$U^\star = U + \gamma_0 + \gamma_1' X + \gamma_2' Z + \nu = U + \omega. \tag{3}$$

Note that X and Z in the above formulation do not necessarily represent the full vector of covariates (eg, some elements in the vectors α_1, α_2, γ_1, and γ_2 may be 0). We assume that ϵ and ν are mean 0 random variables with variance $\Sigma_{\epsilon\epsilon}$ and $\Sigma_{\nu\nu}$, respectively, and are independent of all other variables with the exception that we allow their covariance, $\Sigma_{\epsilon\nu}$, to be nonzero. We refer to this setting as the *additive error structure*. In this setting the error in the observed censored failure time U^* is a mistiming error but there are no errors in the event indicator Δ.

2.2 | More general error structure

We will see in the sections to follow that raking estimators, contrary to regression calibration estimators, do not require modeling the measurement error structure explicitly. Thus, we will also consider a more general error model that also involves a misspecified event. Whereas the additive error structure in Section 2.1 might be expected in scenarios involving only an error-prone baseline time (eg, self-reported baseline time), the general error model relaxes this assumption to allow the timing of the failure, and thus the failure indicator, to be error-prone as well. Instead of observing (X, Z, U, Δ), one observes $(X^\star, Z, U^\star, \Delta^\star)$, where errors in the event may be coming from both a mistiming error and also from misclassification of the event indicator. Note that with this error structure we also make no assumptions regarding the additivity of errors or their correlation with other variables.

2.3 | Two-phase design

We consider the two-phase design in which the true, error-free variables are measured retrospectively for a subsample of subjects at the second phase. Let R_i be an indicator for whether subject $i = 1, \ldots, n$ is selected to be in the second phase and let $0 < \pi_i \leq 1$ be their known sampling probability. In general, the sampling probabilities are known in validation studies based on observational data utilizing EHR, which are becoming increasingly common. This sampling scheme also accommodates scenarios where the subsample size is fixed (eg, simple random sampling) and where the subsample size is random (eg, Bernoulli sampling), as well as stratified designs (eg, case-cohort). We assume that at phase one, the random variables $(X_i^\star, Z_i, U_i^\star, \Delta_i)$ (or $(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$ in a setting with misclassification) are observed for n subjects as a random sample from the population. At phase two, $m < n$ subjects are selected from the phase one population according to the aforementioned sampling probability and the random variables (X_i, U_i) (or (X_i, U_i, Δ_i)) are additionally observed for those subjects. From this point on, we refer to the phase two subjects as the validation subset.

3 | PROPOSED REGRESSION CALIBRATION METHODOLOGY

In this section, we give a brief introduction to the original RC and risk set regression calibration (RSRC) methods for classical covariate measurement error and then develop their extensions for our considered error settings that include error in the censored outcome alone and correlated errors in the censored outcome and covariates. Under regularity conditions similar to those in Andersen and Gill,[15] the RC and RSRC estimators developed in this section for error in the censored outcome and potentially correlated errors in the censored outcome and covariates are asymptotically normal, although not necessarily consistent for β. The proof is similar to that in the covariate error only setting, which was shown in Wang et al.[16] For more detail see Appendix A of the Supplementary Materials.

3.1 | Regression calibration for covariate error

Prentice[10] introduced the regression calibration method for the setting of Cox regression and classical measurement error in the covariate. Shaw and Prentice[17] applied regression calibration for the covariate error structure assumed in Section 2.1. The idea of regression calibration is to estimate the unobserved true variable with its expectation given the data. Prentice[10] showed that under the independent censoring assumption, the induced hazard function based on the error-prone data is given by $\lambda(t; X^\star, Z) = \lambda_0(t)\exp(\beta_Z' Z)\mathrm{E}(\exp\{\beta_X' X\}|X^\star, Z, U \geq t)$. He then showed that for rare events and moderate β_X, $\mathrm{E}(\exp\{\beta_X' X\}|X^\star, Z, U \geq t) \approx \exp(\beta_X' \mathrm{E}(X|X^\star, Z))$. $\mathrm{E}(X|X^\star, Z)$ can be estimated using the following first-order approximation

$$\mathrm{E}(X|X^\star, Z) = \mu_X + \begin{bmatrix} \Sigma_{XX^\star} & \Sigma_{XZ} \end{bmatrix} \begin{bmatrix} \Sigma_{X^\star X^\star} & \Sigma_{X^\star Z} \\ \Sigma_{ZX^\star} & \Sigma_{ZZ} \end{bmatrix}^{-1} \begin{bmatrix} X^\star - \mu_{X^\star} \\ Z - \mu_Z \end{bmatrix}, \tag{4}$$

where the validation subset is used to calculate the moments involving X (see Shaw and Prentice[17]). Define $\hat{X} = \mathrm{E}(X|X^\star, Z; \hat{\zeta}_x)$, where $\hat{\zeta}_x$ is the vector of nuisance parameters in (4) estimated from the data. $\hat{X}$ is then imputed for X in the partial likelihood score (1) instead of the observed $X^\star$ to solve for β, which yields the corrected estimates.[17] Note, for simplicity we generally suppress the notation of the dependence of terms such as $\mathrm{E}(X|X^\star, Z)$ on the nuisance parameter ζ_x, unless it is important for clarity, such as to refer to its estimator $\mathrm{E}(X|X^\star, Z; \hat{\zeta}_x)$.

3.2 | Proposed regression calibration extension for time-to-event error

Assume only the time-to-event error structure given in (3) in Section 2.1, that is, we observe $(X, Z, U^\star, \Delta)$. Given the additivity of the outcome errors in (3), we can take the expectation of the censored event time, $U^\star$, given the observed covariates and rearrange to obtain $\mathrm{E}(U|X, Z) = \mathrm{E}(U^\star|X, Z) - \mathrm{E}(\omega|X, Z)$. We use $\mathrm{E}(\omega|X, Z)$ to correct $U^\star$ and then impute as our estimate of the true censored event time. Since the true $\mathrm{E}(\omega|X, Z)$ is unknown, we can estimate it using the following first-order approximation

$$\mathrm{E}(\omega|X, Z; \zeta_\omega) = \mu_\omega + \begin{bmatrix} \Sigma_{\omega X} & \Sigma_{\omega Z} \end{bmatrix} \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}^{-1} \begin{bmatrix} X - \mu_X \\ Z - \mu_Z \end{bmatrix}, \tag{5}$$

where the validation subset is used to calculate the moments involving ω and ζ_ω is the vector of nuisance parameters in (5). Adjusting $U^\star$ to have the correct expectation gives us $\hat{U} = U^\star - \mathrm{E}(\omega|X, Z; \hat{\zeta}_\omega)$, which we use instead of $U^\star$ to solve the partial likelihood score (1) for the corrected β estimates.

3.3 | Proposed regression calibration extension for covariate and time-to-event error

Assume the additive error structure for both $X^\star$ and $U^\star$ in Section 2.1, that is, we observe $(X^\star, Z, U^\star, \Delta)$. Given the additivity of the outcome errors in (2.3), we can take the expectation of the censored event time, $U^\star$, given the observed covariates and rearrange to obtain $\mathrm{E}(U|X^\star, Z) = \mathrm{E}(U^\star|X^\star, Z) - \mathrm{E}(\omega|X^\star, Z)$. We use $\mathrm{E}(\omega|X^\star, Z)$ to correct $U^\star$ and then impute as our estimate of the true censored event time. Due to the error-prone $X^\star$, we impute $\mathrm{E}(X|X^\star, Z)$ for X as well, similar to Prentice.[10] Given that the true $\mathrm{E}(X|X^\star, Z; \zeta_x)$ is unknown, we estimate it using the same first-order approximation described in Section 3.1. In addition, we estimate $\mathrm{E}(\omega|X^\star, Z; \zeta_\omega)$ using a similar first-order approximation to that described in Section 3.2, constructed using the observed data $(X^\star, Z)$, giving us $\hat{U} = U^\star - \mathrm{E}(\omega|X^\star, Z; \hat{\zeta}_\omega)$ as the estimate of the true censored time-to-event. Thus, we impute $\hat{U}$ and $\hat{X} = \mathrm{E}(X|X^\star, Z; \hat{\zeta}_x)$ in the partial likelihood score (1) instead of the observed $U^\star$ and $X^\star$ and solve for β to obtain our corrected estimates.

3.4 | Proposed risk set regression calibration (RSRC) extension

We also considered improving our regression calibration estimators by applying the idea of recalibrating the mismeasured covariate within each risk set developed by Xie et al[18] for classical measurement error and extended to the covariate error model in Section 2.1 by Shaw and Prentice.[17] Since the risk set membership likely depends on subject specific covariates whose distribution is changing over time, we may be able to obtain better RC estimates by performing the calibration at every risk set as events occur. In particular, this method was shown to decrease the bias significantly for the setting of covariate measurement error when the hazard ratio is quite large, a case in which ordinary RC has been observed to perform poorly. Specifically for covariate measurement error, the risk set regression calibration estimator solves the partial likelihood score (1) using $\hat{X}(t)$ instead of X, where $\hat{X}(t)$ is recalculated using RC at each event time using data from only those individuals still in the risk set at that event time.

In the presence of time-to-event error, however, the necessary moments needed to estimate the conditional expectations in Sections 3.2 and 3.3 at the ith individuals' censored event time will be incorrect due to the fact that the risk sets defined by $U^\star$ will not be the same as those defined by U, leading to biased estimates. Thus, to extend the RSRC idea to the settings of error in the censored outcome and correlated error in the covariate and censored outcome, we propose a two-stage RSRC estimator where the first stage involves obtaining the estimate $\hat{U}$ using ordinary RC. The second stage then assumes $\hat{U}$ is the observed event time instead of $U^\star$ and recalibrates $\hat{U}$ and $X^\star$ at risk sets defined by $\hat{U}$ using the methods described in Sections 3.2 and 3.3.

4 | PROPOSED GENERALIZED RAKING METHODOLOGY

In this section, we develop design-based estimators by applying generalized raking (raking for short),[19,20] which leverages the error-prone data available on the entire sample to improve the efficiency of consistent estimators calculated using the error-free validation subset. We give a brief overview of the general raking method and then propose our estimators for the correlated measurement error settings under consideration. Under suitable regularity conditions, the proposed raking estimators have been shown to be $\sqrt{n}$ consistent, asymptotically normal estimators of β for all two-phase designs described in Section 2.3. For the proof, see Saegusa and Wellner.[21]

4.1 | Generalized raking overview

Let $P_i(\beta)$ denote the population score equations for the true underlying Cox model with corresponding target parameter β, the log hazard ratio we would estimate if we had error-free data on the full cohort. Then the HT estimator of β is given by the solution to $\sum_{i=1}^{n} \frac{R_i}{\pi_i} P_i(\beta) = 0$, which is known to be a consistent estimator of β. Consider A_i, a vector of auxiliary variables that are available for everyone at phase one and are correlated with the phase two data. Raking estimators modify the design weights $w_{i,des} = \frac{1}{\pi_i}$ to new weights $w_{i,cal} = \frac{g_i}{\pi_i}$ such that they are as close as possible to $w_{i,\,des}$ while $\sum_{i=1}^{n} A_i$ is exactly estimated by the validation subset. Thus, given a distance measure $d(.,.)$, the objective is

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{n} R_i d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right) \\ &\text{subject to} \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} R_i \frac{g_i}{\pi_i} A_i. \end{aligned} \tag{6}$$

Note that the constraints above are known as the calibration equations. Deville et al[20] give several options for choosing the distance function, and the resulting constrained minimization problem can be solved to yield a solution for g_i. The generalized raking estimator is then defined as the solution to

$$\sum_{i=1}^{n} R_i \frac{g_i}{\pi_i} P_i(\beta) = 0. \tag{7}$$

4.2 | Proposed raking estimators

For our setting of the Cox model, we use the distance function $d(a, b) = a \log\left(\frac{a}{b}\right) + (b - a)$ in the objective function of (6) to ensure positive weights. Solving the constrained minimization problem for g_i using λ, a $p + q$-dimensional vector of Lagrange multipliers, then yields $g_i = \exp(-\hat{\lambda}' A_i)$. After plugging in g_i to the calibration equations, Deville and Särndal[19] show that the solution for λ satisfies

$$\hat{\lambda} = \hat{B}^{-1} \left(\sum_{i=1}^{N} \frac{R_i}{\pi_i} A_i - \sum_{i=1}^{N} A_i \right) + O_p(n^{-1}),$$

where $\hat{B} = \sum_{i=1}^{N} \frac{R_i}{\pi_i} A_i' A_i$. Finally, we construct auxiliary variables, A_i, that yield efficient estimators.

Breslow et al[13] derived the asymptotic expansion for the solution to (7) and showed that the optimal auxiliary variable is given by $A_i^{\text{opt}} = \text{E}(\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)|V)$, where $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ denotes the efficient influence function contributions from the population model had the true outcome and covariates been observed for everyone in phase one and $V = (X^\star, Z, U^\star, \Delta)$ (or $(X^\star, Z, U^\star, \Delta^\star)$ in a setting with misclassification). However, calculating A_i^{opt} involves a conditional distribution of unobserved variables and thus is generally not practically obtainable. Kulich and Lin[22] proposed a "plug-in" method that approximates this conditional expectation by using the influence functions from a model fit using phase one data. Specifically, they proposed to use the phase two data to fit models that impute the missing information from the phase one data only and then to obtain the influence functions from the desired model that uses imputed values in place of the missing data. They further proposed using a jackknife approximation of the influence function, a delta-beta-type residual typically available in most software programs. We will propose two different imputations for the missing data, which will lead to two different choices of A_i that approximate A_i^{opt}.

The first proposed approximation of A_i^{opt} is given by $A_{N,i} = \tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i)$, the influence function for the naive estimator that used the error prone data instead of the unobserved true values. One can estimate $A_{N,\,i}$ empirically using

$$\begin{aligned} \tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i) \approx{}& \Delta_i \left\{ \{X_i^\star, Z_i\}' - \frac{S^{(1)\star}(\beta, t)}{S^{(0)\star}(\beta, t))} \right\} \\ &- \sum_{i=1}^{n} \int_0^\tau \frac{\exp(\beta_X' X_i^\star + \beta_Z' Z_i)}{S^{(0)\star}(\beta, t)} \left\{ \{X_i^\star, Z_i\}' - \frac{S^{(1)\star}(\beta, t)}{S^{(0)\star}(\beta, t))} \right\} dN_i^\star(t), \end{aligned}$$

where $S^{(r)\star}(\beta, t) = n^{-1} \sum_{j=1}^{n} Y_j^\star(t) \{X_j^\star, Z_j\}'^{\otimes r} \exp(\beta_X' X_j^\star + \beta_Z' Z_j)$ ($a^{\otimes 1}$ is the vector a and $a^{\otimes 0}$ is the scalar 1). For measurement error settings including an error-prone failure indicator, we approximate A_i^{opt} with $A_{N,i} = \tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$.

The second proposed approximation of A_i^{opt} is given by $A_{\text{RC},i} = \tilde{\ell}_0(\hat{X}_i(\hat{\zeta}_x), Z_i, \hat{U}_i(\hat{\zeta}_\omega), \Delta_i)$, that is, the influence function for the target estimator that uses the calibrated estimates $(\hat{X}_i(\hat{\zeta}_x), \hat{U}_i(\hat{\zeta}_\omega))$ in place of the unobserved true data (X_i, U_i). One can again use the empirical approximation

$$\tilde{\ell}_0(\hat{X}_i(\hat{\zeta}_x), Z_i, \hat{U}_i(\hat{\zeta}_\omega), \Delta_i) \approx \Delta_i \left\{ \{\hat{X}_i(\hat{\zeta}_x), Z_i\}' - \frac{\hat{S}^{(1)}(\beta, \hat{\zeta}, t)}{\hat{S}^{(0)}(\beta, \hat{\zeta}, t))} \right\}$$
$$- \sum_{i=1}^{n} \int_0^\tau \frac{\exp(\beta_X' \hat{X}_i(\hat{\zeta}_x) + \beta_Z' Z_i)}{\hat{S}^{(0)}(\beta, \hat{\zeta}, t)} \left\{ \{\hat{X}_i(\hat{\zeta}_x), Z_i\}' - \frac{\hat{S}^{(1)}(\beta, \hat{\zeta}, t)}{\hat{S}^{(0)}(\beta, \hat{\zeta}, t))} \right\} d\hat{N}_i(t; \hat{\zeta}_\omega),$$

where $\hat{S}^{(r)}(\beta, \hat{\zeta}, t) = n^{-1} \sum_{j=1}^{n} \hat{Y}_j(t; \hat{\zeta}_\omega)\{\hat{X}_j(\hat{\zeta}_x), Z_j\}'^{\otimes r} \exp(\beta_X' \hat{X}_j(\hat{\zeta}_x) + \beta_Z' Z_j)$ ($a^{\otimes 1}$ is the vector a and $a^{\otimes 0}$ is the scalar 1). For measurement error settings including an error-prone failure indicator, we approximate A_i^{opt} with $A_{\text{RC},i} = \tilde{\ell}_0(\hat{X}_i(\hat{\zeta}_x), Z_i, \hat{U}_i(\hat{\zeta}_\omega), \Delta_i^\star)$. Thus, the two proposed raking estimators are:

1. Generalized raking naive (GRN): solution to (7) using $A_{N,i}$
2. Generalized raking regression calibration (GRRC): solution to (7) using $A_{\text{RC},i}$,

where both estimators utilize $g_i = \exp(-\hat{\lambda}' A_i)$.

The efficiency gain from the raking estimator over the HT estimator depends on the correlation between the auxiliary variables and the target variables. Breslow and Wellner[23] showed that the variance of HT parameter estimates is the sum of the model-based variance due to sampling from an infinite population with no missing data and the design-based variance resulting from estimation of the unknown full cohort total of efficient influence function contributions. Thus, we consider $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ to be our target variables. We expect the regression calibration estimators to be less biased than the naive estimators and therefore conjecture that A_{RC} would be more highly correlated with A^{opt} than A_N. Note that in general, when the parameter of interest is a regression parameter, choosing the auxiliary variables to be the observed, error-prone variables will not improve efficiency. For more details, see chapter 8 of Lumley.[24]

In addition, the raking estimators share a close relationship with the augmented inverse probability weighted (AIPW) estimators proposed by Robins et al.[12] The class of AIPW estimators studied by Robins et al[12] contains all regular asymptotically linear estimators consistent for the design-based parameter of interest, including the raking estimators. The class of raking estimators, however, includes all of the most efficient AIPW estimators.[14] Thus, raking estimators are asymptotically efficient among design-based estimators and can provide simple, easy to compute AIPW estimators.

4.3 | Calculating raking estimators

Instead of explicitly calculating $A_{N,i}$ and $A_{\text{RC},i}$ with the influence function formulas given above, we propose to utilize standard software to calculate the A_i so that practitioners may easily implement these methods. In R, the influence functions can be approximated with negligible error using the *dfbeta* (delta-beta)-type residual in the *resid* function. Thus, the raking estimates can be computed as follows:

1. Fit a candidate Cox model using all phase one subjects.
2. Construct the auxiliary variables A_i as imputed *dfbetas* from the model fit in step 1.
3. Estimate regression parameters β using weights raked to A_i by solving (7).

For step 1, we consider the naive Cox model using the error-prone data (GRN) and the regression calibration approach described in Section 3 (GRRC). For step 3, we utilize the *survey* package by Lumley[25] in R, which provides standard software for obtaining raking estimates.

5 | SIMULATION STUDIES

We examined the finite sample performance of our proposed RC, RSRC, GRRC, and GRN estimators through simulation for the error framework described in Section 2. These four estimators were compared with those from the true model, a Cox proportional hazards regression model fit with the true covariates and event times, a naive Cox model fit with the error-prone covariates and/or error-prone censored event times, and the complete-case estimator using only the true covariates and event times in the validation subset. We note that unless noted otherwise, all validation subsets were selected as simple random samples with known sampling probability, meaning the complete-case estimator is equivalent to the HT estimator. We additionally consider validation subsets sampled using a case-control design to compare the performance of the methods under outcome-dependent sampling designs. Following Section 2.1, we considered the additive error structure with correlated covariate and time-to-event error. In addition to this case, we also considered the censored outcome error only setting. We further considered correlated covariate and censored outcome error under the special case where the covariates are only subject to random error, namely classical measurement error ($(\alpha_0, \alpha_2) = \vec{0}; \alpha_1 = \vec{1}$). In addition, we considered the general error structure described in Section 2.2, where there exists errors in the time-to-event that result from mistiming as well as misclassification in addition to additive covariate error. We present %biases, average bootstrap standard errors (ASE) for the four proposed estimators or average model standard errors (ASE) for the naive and complete case estimators, empirical standard errors (ESE), mean square errors (MSE), and 95% coverage probabilities (CP) for varying values of the log hazard ratio β_X, %censoring, and error variances and covariances. We additionally present type 1 error results for $\beta_X = 0$ and $\alpha = .05$.

5.1 | Simulation setup

All simulations were run 2000 times using R version 3.4.2. The error-prone covariate X was generated as a standard normal distribution and the error-free covariate as $Z \sim \text{N}(2, 1)$, with $\rho_{X,Z} = 0.5$. We set the true log hazard ratios to be $\beta_X \in \{\log(1.5), \log(3)\}$, which we refer to as moderate and large, respectively, and $\beta_Z = \log(2)$. The true survival time T was generated from an exponential distribution with rate equal to $\lambda_0 \exp(\beta_X X + \beta_Z Z)$, where $\lambda_0 = 0.1$. We then simulated 25% and 75% censoring, which we refer to as common and rare event settings, respectively, by generating separate random right censoring times for each β_X to yield the desired %censored event times. Censoring times were generated as Uniform distributions with length 2 and 0.4 for each %censored time, respectively, to mimic studies of different lengths. For the error terms ϵ and ν, we considered normal distributions with means 0, variances $(\Sigma_{\epsilon\epsilon} = \sigma^2_\epsilon, \Sigma_{\nu\nu} = \sigma^2_\nu) \in \{0.5, 1\}$, and $(\Sigma_{\epsilon\nu} = \sigma_{\epsilon\nu}) \in \{0.15, 0.3\}$, resulting in correlations ranging from 0.15 to 0.60. The error-prone covariate and censored event time were generated with parameters $(\alpha_0, \alpha_1, \alpha_2) = (0, 0.9, -0.2)$ and $(\gamma_0, \gamma_1, \gamma_2) = (\sigma_\nu \times 3, 0.2, -0.3)$. The choice of γ_0 is such that the error-prone time is a valid event time (ie, greater than zero) with high probability. The few censored event times that were less than 0 were reflected across 0 to generate valid outcomes.

For the error terms ϵ and ν, we also considered a mixture of a point mass at zero and a shifted gamma distribution with the same means and covariances as the normal distributions to determine the robustness of our methods to non-normality of errors. Note that while the RC and RSRC estimators are expected to be challenged by such departures from normality, the raking estimators are not affected by the structure of the measurement error other than by the strength of the correlation between the auxiliary variables and the target variables. The mixture probability was set to be 0.5 for both covariate and outcome error.

For the misclassification example, we set $\beta_X = \log(1.5)$, $\sigma^2_\epsilon = \sigma^2_\nu = 0.5$, $\sigma_{\epsilon\nu} = 0.15$, with normally distributed error terms and 75% censoring. In addition, the sensitivity and specificity for Δ were set to 90% by adding Bernoulli error ($p = 0.10$). For all simulations, we set the number of subjects to be 2000. Unless noted otherwise, we selected the validation subsets as simple random samples of size 200, or $\pi_i = \pi = 0.1$. We additionally considered validation subsets selected using case-control sampling where 100 cases and 100 controls were randomly sampled. The data example in Section 6 additionally considers selecting the validation subsets using unequal sampling probabilities via outcome-dependent sampling.

Standard errors for the RC, GRRC, and GRN estimates were obtained using the bootstrap method with bootstrap sampling stratified on the validation subset membership and using 300 bootstrap samples. Note that while the raking estimators have known sandwich variance estimators for the asymptotic variance, we used the bootstrap to calculate standard errors and coverage probabilities (see Appendix B of the Supplementary Materials for an empirical comparison). The RSRC standard errors were also calculated similarly using the bootstrap; however, only 100 bootstrap samples were utilized due to its computational burden. In addition, the RSRC estimators were recalibrated at deciles of the observed event times.

5.2 | Simulation results

For all discussed tables, we observed that the naive estimates had very large bias with 95% coverage hovering around 0%. By contrast, the complete case estimates were nearly unbiased for all settings discussed, but suffered from large standard errors, particularly for rare event settings when there were only a few subjects who had events in the validation subset. The coverage of the complete case estimates was near 95% for all settings. In the discussion of simulation results to follow, we focus on the four proposed estimators and how their relative performance differed across settings.

Table 1 presents the relative performance for estimating β_X in the presence of the time-to-event error described in Section 2.1 and no covariate error, with $\nu \sim \text{N}(0, \sigma_\nu^2)$. The RC estimates had moderate to large bias (−13% to −33%) and coverage ranging from 0.87 to 0, depending on if β_X was moderate or large. We observed around a 50% decrease in bias for the RSRC estimates compared with RC for moderate β_X and common events and a range of 5% to 30% bias reduction for other settings, with coverage around 87% to 93% and 0% for moderate and large β_X, respectively. The reduction in bias for the RSRC estimates resulted in a lower MSE for all settings except under moderate β_X and rare events, a setting in which RC is known to perform well. Both raking estimates were nearly unbiased across all parameter settings, had uniformly lower standard errors than the complete case estimates, and had coverage near 95%. Interestingly, the performances of the GRRC and GRN estimators were virtually indistinguishable, with similar bias, standard errors, MSE, and coverage. Overall, RSRC had the lowest MSE for all moderate β_X settings whereas the raking estimates had the lowest MSE for all large β_X settings.

Tables 2 and 3 consider the relative performance for estimating a moderate log hazard ratio in the setting of correlated additive errors in the outcome and covariate as described in Section 2.1 for normally distributed error terms and common and rare events, respectively. The RC estimates had relatively moderate bias (−13% to −19%) and coverage ranging from 0.74 to 0.92. For common events, the RSRC estimates had around 50% less bias than the RC estimates, whereas for rare events, they yielded only a small decrease in bias. Even in these more complex error settings, both raking estimates remained nearly unbiased, had lower standard errors than the complete case estimates, and maintained coverage around 95% across varying error variances and covariances. We noticed that for all parameter settings, the GRRC and GRN estimators were again nearly indistinguishable. Overall for the common event settings, the RSRC estimates had the lowest MSE when the error variances were both 0.5; otherwise, the raking estimates had the lowest MSE for all other settings. For the rare event settings, the RC estimates had the lowest MSE across all variance and covariance settings.

We present the relative performance for estimating a larger log hazard ratio, keeping other parameters the same as in Tables 2 and 3, in Table 4 and Supplementary Materials Table 1 in Appendix C. Both the RC and RSRC estimates had large bias, ranging from −31% to −37% and −23% to −32%, respectively, as well as coverage 50% or below. Again, both raking estimates remained nearly unbiased, had lower standard errors than the complete case estimates, and maintained coverage around 95% across varying error variances and covariances, with the GRRC and GRN estimates indistinguishable. Across all error settings, the raking estimates had the lowest MSE.

Table 5 presents the type 1 error, ASE, ESE, and MSE when $\beta_X = 0$ in the presence of correlated, additive measurement error in the outcome and covariate X with normally distributed errors. For both levels of censoring, the type 1 error of the RC and RSRC estimates ranged from 0.044 to 0.059 and the raking estimates were around 0.042 and 0.046 for common and rare events, respectively. It is of note that the type 1 error for the naive estimator is 1 for both levels of censoring, meaning the null hypothesis was falsely rejected in every simulation run.

Results for β_Z, for the settings presented in Tables 1 to 4, are presented in Tables 2 to 5 of Appendix C in the Supplementary Materials. The conclusions for this parameter were similar to those of β_X; however, the raking estimates had the lowest MSE across more settings. Tables 6 to 8 in Appendix D of the Supplementary Materials present simulation results for β_X in a setting where the covariates are only subject to classical measurement error, keeping all other settings the same as Tables 2 to 4. Results are similar to those presented above.

We consider the relative performance for when the error distributions were generated as a mixture of a point mass at 0 and shifted gamma distribution, with settings otherwise the same as those in Tables 1-4, in Tables 9 to 12 of Appendix E in the Supplementary Materials. The RC and RSRC estimators were challenged by such departures from normality, with generally more bias and higher MSE, while the raking estimators remained unbiased with lower MSE.

Table 13 in Appendix F of the Supplementary Materials considers the relative performance of the estimators in the presence of misclassification errors in addition to the correlated additive errors in the time-to-event and covariate X, as described in Section 2.2. The RC and RSRC estimates had very large bias and coverage between 61% and 68% as these methods were not developed to directly handle misclassification. As expected, the GRRC and GRN estimates were nearly

TABLE 1 Simulation results for β_X under additive measurement error only in the outcome with normally distributed error and 25% and 75% censoring for the true event times

%Censoring	β_X	σ_v^2	Method	%Bias	ASE	ESE	MSE	CP
25	log(1.5)		True	−0.025	0.030	0.031	0.001	0.947
		0.5	RC	−12.677	0.042	0.043	0.004	0.752
			RSRC	−5.056	0.048	0.050	0.003	0.928
			GRRC	0.074	0.059	0.058	0.003	0.957
			GRN	0.271	0.060	0.059	0.003	0.958
			Naive	−37.562	0.030	0.031	0.024	0.002
			Complete	0.321	0.098	0.098	0.010	0.952
		1	RC	−18.522	0.046	0.047	0.008	0.624
			RSRC	−7.991	0.055	0.056	0.004	0.910
			GRRC	−0.025	0.066	0.065	0.004	0.956
			GRN	0.074	0.066	0.065	0.004	0.958
			Naive	−40.891	0.030	0.030	0.028	0.000
			Complete	0.321	0.098	0.098	0.010	0.954
	log(3)		True	0.046	0.037	0.036	0.001	0.951
		0.5	RC	−26.879	0.054	0.056	0.090	0.001
			RSRC	−19.188	0.060	0.063	0.048	0.070
			GRRC	−0.983	0.103	0.102	0.010	0.938
			GRN	−1.010	0.104	0.104	0.011	0.939
			Naive	−37.347	0.031	0.040	0.170	0.000
			Complete	0.819	0.118	0.118	0.014	0.954
		1	RC	−33.042	0.056	0.058	0.135	0.000
			RSRC	−23.466	0.065	0.067	0.071	0.027
			GRRC	−0.883	0.108	0.105	0.011	0.940
			GRN	−0.847	0.108	0.106	0.011	0.942
			Naive	−41.88	0.030	0.039	0.213	0.000
			Complete	0.819	0.118	0.118	0.014	0.955
75	log(1.5)		True	0.074	0.054	0.054	0.003	0.948
		0.5	RC	−15.340	0.079	0.080	0.010	0.872
			RSRC	−12.874	0.087	0.089	0.011	0.898
			GRRC	−0.099	0.113	0.112	0.012	0.957
			GRN	0.543	0.116	0.117	0.014	0.955
			Naive	−69.204	0.054	0.055	0.082	0.000
			Complete	0.444	0.176	0.182	0.033	0.950
		1	RC	−17.338	0.081	0.084	0.012	0.845
			RSRC	−15.488	0.089	0.092	0.012	0.873
			GRRC	−0.444	0.118	0.118	0.014	0.952
			GRN	0.247	0.120	0.121	0.015	0.953
			Naive	−57.638	0.054	0.056	0.058	0.016
			Complete	−0.099	0.177	0.182	0.033	0.946

(Continues)

TABLE 1 (Continued)

%Censoring	β_X	σ_v^2	Method	%Bias	ASE	ESE	MSE	CP
	log(3)		True	0.118	0.058	0.059	0.003	0.950
		0.5	RC	−31.030	0.085	0.088	0.124	0.024
			RSRC	−28.827	0.094	0.097	0.110	0.087
			GRRC	−0.901	0.166	0.163	0.027	0.951
			GRN	−0.446	0.168	0.175	0.031	0.950
			Naive	−52.357	0.053	0.062	0.335	0.000
			Complete	1.912	0.191	0.197	0.039	0.946
		1	RC	−33.060	0.087	0.091	0.140	0.024
			RSRC	−31.567	0.095	0.099	0.130	0.055
			GRRC	−0.774	0.171	0.170	0.029	0.940
			GRN	−0.501	0.171	0.172	0.030	0.942
			Naive	−48.680	0.053	0.061	0.290	0.000
			Complete	1.930	0.193	0.202	0.041	0.946

Note: For 2000 simulated data sets, the bias, average bootstrap standard error (ASE) for the four proposed estimators, average model standard error (ASE) for naive and complete case, empirical standard error (ESE), mean squared error (MSE), and 95% coverage probabilities (CP) are presented.

unbiased because the raking estimators do not depend on the structure of the measurement error. Overall, the raking estimators had the lowest MSE in this more complex error setting.

Tables 14 and 15 in Appendix G of the Supplementary Materials considers the relative performance of the estimators when the validation subset is selected using case-control sampling, with settings otherwise the same as those in Table 3 and Supplementary Materials Table 1 in Appendix C. The RC and RSRC estimates had moderate to large bias for both values of β_X, as in the simple random sampling settings. The GRRC and GRN estimates, however, were nearly unbiased and had the lowest MSE for almost all settings due to the lower standard errors compared with the estimates from the simple random sampling scenarios.

6 | DATA EXAMPLE

We applied the four proposed methods to electronic health records data from a large HIV clinic, the Vanderbilt Comprehensive Care Clinic (VCCC). The VCCC is an outpatient clinic that provides care to HIV patients and collects clinical data over time that is electronically recorded by nurses and physicians.[26] The VCCC fully validated all key variables for all records, resulting in an unvalidated, error-prone data set and a fully validated data set that we consider to be correct. Thus, this observational cohort is ideal for directly assessing the relative performance of the proposed regression calibration and raking estimators compared with the naive and HT estimators. Note that the naive estimator was calculated using only the unvalidated data set as if the validated data set did not exist. In addition, the HT estimator was calculated using a subsample of the fully validated data set. Throughout this example, we considered the estimates from the fully validated data set to be the "truth" and defined these as the parameters of interest. In addition, all considerations of bias were relative to these target parameters. We considered two different failure time outcomes of interest: time from the start of antiretroviral therapy (ART) to the time of virologic failure and to the time of first AIDS defining event (ADE). For the former analysis, virologic failure was defined as an HIV-RNA count greater than or equal to 400 copies/mL and patients were censored at the last available test date after ART initiation. The HIV-RNA assay, and hence time at virologic failure was largely free of errors, whereas the time at ART start was error-prone, corresponding to errors in U. The ADE outcome was defined as the first opportunistic infection (OI) and patients were censored at age of death if it occurred or last available test date after ART initiation. For this failure time, both time of ART initiation and time at first ADE were error-prone, corresponding to errors in U and Δ. We studied the association between the outcomes of interest and the CD4 count and age at ART initiation. Since date of ART initiation was error prone, CD4 and age at ART initiation may

TABLE 2 Simulation results for $\beta_X = \log 1.5$ under correlated, additive measurement error in the outcome and covariate X with normally distributed error and 25% censoring for the true event time

β_X	σ_ν^2	σ_ϵ^2	$\sigma_{\nu,\epsilon}$	**Method**	**%Bias**	**ASE**	**ESE**	**MSE**	**CP**
log(1.5)				True	−0.025	0.030	0.031	0.001	0.947
	0.5	0.5	0.15	RC	−13.762	0.059	0.059	0.007	0.804
				RSRC	−6.338	0.070	0.068	0.005	0.922
				GRRC	0.173	0.083	0.084	0.007	0.947
				GRN	0.345	0.083	0.084	0.007	0.946
				Naive	−79.760	0.024	0.025	0.105	0.000
				Complete	0.321	0.098	0.098	0.010	0.952
			0.30	RC	−13.491	0.060	0.060	0.007	0.813
				RSRC	−6.116	0.071	0.069	0.005	0.928
				GRRC	0.296	0.083	0.084	0.007	0.947
				GRN	0.567	0.083	0.084	0.007	0.945
				Naive	−97.024	0.024	0.025	0.155	0.000
				Complete	0.173	0.098	0.099	0.010	0.954
		1	0.15	RC	−13.836	0.072	0.071	0.008	0.843
				RSRC	−7.054	0.084	0.083	0.008	0.922
				GRRC	0.049	0.089	0.090	0.008	0.948
				GRN	0.148	0.089	0.090	0.008	0.952
				Naive	−86.099	0.020	0.020	0.122	0.000
				Complete	0.271	0.098	0.098	0.010	0.952
			0.30	RC	−13.639	0.073	0.072	0.008	0.845
				RSRC	−6.955	0.086	0.084	0.008	0.914
				GRRC	0.074	0.089	0.090	0.008	0.947
				GRN	0.271	0.089	0.089	0.008	0.945
				Naive	−97.912	0.020	0.020	0.158	0.000
				Complete	0.222	0.098	0.098	0.010	0.957
	1	0.5	0.15	RC	−19.237	0.065	0.065	0.010	0.746
				RSRC	−9.520	0.078	0.076	0.007	0.902
				GRRC	0.123	0.085	0.086	0.007	0.944
				GRN	0.247	0.085	0.086	0.007	0.944
				Naive	−79.686	0.024	0.025	0.105	0.000
				Complete	0.321	0.098	0.098	0.010	0.954
			0.30	RC	−19.311	0.066	0.066	0.010	0.743
				RSRC	−9.693	0.079	0.077	0.008	0.903
				GRRC	0.148	0.085	0.086	0.007	0.945
				GRN	0.345	0.085	0.085	0.007	0.946
				Naive	−95.027	0.024	0.025	0.149	0.000
				Complete	0.173	0.098	0.098	0.010	0.955
		1	0.15	RC	−19.213	0.079	0.079	0.012	0.801

(Continues)

TABLE 2 (Continued)

β_X	σ_ν^2	σ_ϵ^2	$\sigma_{\nu,\epsilon}$	Method	%Bias	ASE	ESE	MSE	CP
				RSRC	−10.235	0.095	0.092	0.010	0.908
				GRRC	−0.025	0.090	0.092	0.008	0.945
				GRN	0.074	0.090	0.091	0.008	0.946
				Naive	−86.049	0.020	0.020	0.122	0.000
				Complete	0.148	0.098	0.099	0.010	0.952
			0.30	RC	−19.213	0.080	0.080	0.012	0.798
				RSRC	−10.580	0.096	0.093	0.010	0.902
				GRRC	0.123	0.090	0.091	0.008	0.947
				GRN	0.247	0.090	0.091	0.008	0.948
				Naive	−96.556	0.020	0.020	0.154	0.000
				Complete	0.321	0.098	0.098	0.010	0.953

Note: For 2000 simulated data sets, the bias, average bootstrap standard error (ASE) for the four proposed estimators, average model standard error (ASE) for naive and complete case, empirical standard error (ESE), mean squared error (MSE), and 95% coverage probabilities (CP) are presented.

also have errors. Appendix H of the Supplementary Materials provides detail on the eligibility criteria and statistics for the covariate and time-to-event error for both analyses.

The analysis of the virologic failure outcome included 1863 patients with moderate censoring rates of 46.1% and 47.2% in the unvalidated and validated data set, respectively. We observed highly (slightly) skewed error in CD4 count at ART start (observed event times) and very small amounts of misclassification. The validation subset was selected as a simple random sample of 20%, resulting in 373 patients. For this sampling design, the HT estimator is equivalent to the complete case estimator. The hazard ratios and their corresponding confidence intervals comparing the estimators are displayed graphically in the first row of Figure 1 and shown in Table 16 in Appendix I of the Supplementary Materials. We note that the standard errors for all estimators (including the true, naive, and HT) were calculated using the bootstrap with 300 replicates, which were somewhat larger than the model SEs likely due to a lack of fit of the Cox model. The RSRC estimators were recalibrated at vigintiles of the observed event times. For this analysis, there was little bias in the naive estimators of a 100 cell/mm^3 increase in CD4 count at ART initiation and 10 year increase of age at ART initiation (1.87% and 2.17%, respectively). For both covariates, RC and RSRC provided very minimal improvements in bias, albeit with slightly wider confidence intervals. Small bias notwithstanding, we noticed that both the GRRC and GRN estimators had smaller bias compared with the naive estimator and had narrower confidence intervals than the HT estimator. The GRRC and GRN estimators had very little differentiating them, similar to what was observed in the simulations.

The analysis of the ADE outcome included 1595 patients with very high censoring rates of 84.5% and 93.8% in the unvalidated and validated data set, respectively. We observed highly (slightly) skewed error in CD4 count at ART start (observed event times) and a misclassification rate of 11% that was largely due to false positives (positive predictive value = 35%). While the RC and RSRC methods developed in this article do not explicitly handle misclassification, we were nevertheless interested in seeing how they would perform in this real data scenario in comparison with the raking methods that can handle misclassification. Due to ADE being a rare event, we utilized a case-cohort sampling scheme to select the validation subset. Specifically, we selected a simple random sample of 7%, or 112 patients, from the full error-prone data and then added the remaining 227 subjects classified as cases by the error-prone ADE indicator to the validation subset. Note that due to the outcome-dependent sampling scheme of the case-cohort design, the estimates of the conditional expectations involved in the RC and RSRC estimators cannot be calculated in the same manner as under simple random sampling. Thus, we used IPW least squares to estimate the conditional expectations for RC, RSRC, and GRRC (step 1 of calculating raking estimates as detailed in Section 4.3) . The hazard ratios and their corresponding confidence intervals comparing the estimators are displayed graphically in the second row of Figure 1 and shown in Table 16 in Appendix I of the Supplementary Materials. The standard errors for all estimators were again calculated using the bootstrap with 300 replicates. We noticed significantly more bias in the naive estimators of a 100 cell/mm^3 increase in CD4 count at ART initiation and 10 year increase of age at CD4 count measurement (31.44% and 31.2%, respectively). In

TABLE 3 Simulation results for $\beta_X = \log 1.5$ under correlated, additive measurement error in the outcome and covariate X with normally distributed error and 75% censoring for the true event time

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	%Bias	ASE	ESE	MSE	CP
log(1.5)				True	0.074	0.054	0.054	0.003	0.948
	0.5	0.5	0.15	RC	−15.143	0.109	0.108	0.015	0.906
				RSRC	−12.677	0.120	0.120	0.017	0.925
				GRRC	0.222	0.154	0.153	0.023	0.955
				GRN	0.987	0.156	0.156	0.024	0.956
				Naive	−120.208	0.046	0.046	0.240	0.000
				Complete	0.444	0.176	0.182	0.033	0.950
			0.30	RC	−14.477	0.109	0.108	0.015	0.900
				RSRC	−11.715	0.121	0.119	0.016	0.922
				GRRC	0.099	0.154	0.152	0.023	0.954
				GRN	1.406	0.154	0.154	0.024	0.954
				Naive	−167.043	0.048	0.049	0.461	0.000
				Complete	0.444	0.177	0.183	0.034	0.948
		1	0.15	RC	−14.896	0.134	0.131	0.021	0.920
				RSRC	−13.047	0.146	0.146	0.024	0.931
				GRRC	−0.099	0.166	0.164	0.027	0.962
				GRN	0.271	0.168	0.166	0.028	0.958
				Naive	−113.623	0.038	0.038	0.214	0.000
				Complete	0.271	0.177	0.183	0.034	0.952
			0.30	RC	−14.650	0.133	0.131	0.021	0.922
				RSRC	−12.381	0.146	0.145	0.024	0.936
				GRRC	0.839	0.166	0.164	0.027	0.958
				GRN	1.430	0.168	0.167	0.028	0.956
				Naive	−143.465	0.039	0.039	0.340	0.000
				Complete	1.208	0.177	0.182	0.033	0.948
	1	0.5	0.15	RC	−16.993	0.113	0.114	0.018	0.890
				RSRC	−15.316	0.123	0.123	0.019	0.907
				GRRC	−0.370	0.156	0.155	0.024	0.954
				GRN	0.444	0.158	0.157	0.024	0.952
				Naive	−102.228	0.045	0.046	0.174	0.000
				Complete	−0.099	0.177	0.182	0.033	0.946
			0.30	RC	−17.264	0.113	0.112	0.017	0.892
				RSRC	−15.464	0.124	0.124	0.019	0.904
				GRRC	−0.222	0.155	0.154	0.024	0.956
				GRN	0.814	0.156	0.155	0.024	0.958
				Naive	−132.613	0.046	0.046	0.291	0.000
				Complete	0.296	0.176	0.182	0.033	0.950
		1	0.15	RC	−17.091	0.138	0.136	0.023	0.918
				RSRC	−15.562	0.150	0.152	0.027	0.916

(Continues)

TABLE 3 (Continued)

β_X	σ_ν^2	σ_ϵ^2	$\sigma_{\nu,\epsilon}$	Method	%Bias	ASE	ESE	MSE	CP
				GRRC	−0.222	0.166	0.165	0.027	0.957
				GRN	0.123	0.168	0.167	0.028	0.955
				Naive	−101.587	0.037	0.038	0.171	0.000
				Complete	−0.074	0.176	0.182	0.033	0.948
			0.30	RC	−17.042	0.138	0.135	0.023	0.916
				RSRC	−15.291	0.151	0.151	0.027	0.916
				GRRC	0.123	0.167	0.165	0.027	0.954
				GRN	0.814	0.169	0.167	0.028	0.952
				Naive	−121.86	0.038	0.038	0.246	0.000
				Complete	0.617	0.177	0.180	0.032	0.954

Note: For 2000 simulated data sets, the bias, average bootstrap standard error (ASE) for the 4 proposed estimators, average model standard error (ASE) for naive and complete case, empirical standard error (ESE), mean squared error (MSE), and 95% coverage probabilities (CP) are presented.

fact, the naive point estimate for age was in the wrong direction compared with the true estimate, yielding anticonservative bias. The RC and RSRC estimators provided little to no bias improvement for both covariates. However, the GRRC and GRN estimates were both nearly unbiased with narrower confidence intervals than those of the HT estimator. In this analysis, the HT estimator appeared to have some bias due to random sampling variability; we evaluated its performance across 10 different validation subsets using case-cohort sampling. The mean of the 10 estimates is given in Table 17 in Appendix I of the Supplementary Materials and shows minimal bias for the HT estimator. Again, we noticed that the GRRC and GRN estimators gave similar estimates, with GRRC (GRN) having narrower confidence intervals for the CD4 (age) hazard ratios. In this analysis, we noticed huge improvements in bias from the GRRC and GRN estimators compared with the naive estimators and decreased standard errors compared with the HT estimator even in the presence of appreciable misclassification, which the RC and RSRC estimators could not handle.

The R package RRCME at https://github.com/ericoh17/RRCME implements our methods on a simulated data set that mimics the structure of the VCCC data. Additionally, Appendix J of the Supplementary Materials contains code that implements the RC and GRN estimators for this simulated data to demonstrate ease of application of these estimators.

7 | DISCUSSION

Data collected primarily for nonresearch purposes, such as those from administrative databases or EHR, can have errors in both the outcome and exposures of interest, which can be correlated. Using EHR data from the VCCC HIV cohort, we observed that Cox regression models using the unvalidated data set compared with the fully validated data set resulted in a 3-fold underestimation of the CD4 hazard ratio for ADE and overestimation of the age hazard ratio in the wrong direction such that the null hypothesis of a unit hazard ratio was nearly rejected. Spurious associations driven by such unvalidated outcomes and exposures can misdirect clinical researchers and can be harmful to patients down the line. Even when variables are reviewed and validated for a subset of the records, the additional information gained from these validation procedures are not often utilized in estimation.

The existing literature does not adequately address such complex error across multiple variables; in particular, the timing error in the censored failure time outcome. In this article, we developed four different estimators that incorporate an internal validation subset in the analysis to try to obtain unbiased and efficient estimates. The RC and RSRC estimators approximate the true model by estimating the true outcome and/or exposure given the unvalidated data and information on the error structure from the validation subset. This approximation lacks consistency in most cases for nonlinear models and the RC and RSRC estimators can have appreciable bias for some error settings. However, in settings with a modest hazard ratio and rare events, RC outperformed the other estimators with respect to having the lowest MSE. RSRC had the lowest MSE for settings with a modest hazard ratio and common events under only censored outcome error and for

TABLE 4 Simulation results for $\beta_X = \log 3$ under correlated, additive measurement error in the outcome and covariate X with normally distributed error and 25% censoring for the true event time

β_X	σ_ν^2	σ_ϵ^2	$\sigma_{\nu,\epsilon}$	Method	%Bias	ASE	ESE	MSE	CP
log(3)				True	0.055	0.037	0.036	0.001	0.952
	0.5	0.5	0.15	RC	−31.239	0.077	0.077	0.124	0.026
				RSRC	−23.038	0.092	0.092	0.072	0.239
				GRRC	0.337	0.113	0.112	0.012	0.950
				GRN	0.346	0.112	0.111	0.012	0.950
				Naive	−70.243	0.025	0.027	0.596	0.000
				Complete	0.819	0.118	0.118	0.014	0.954
			0.30	RC	−31.904	0.079	0.080	0.129	0.030
				RSRC	−23.102	0.097	0.096	0.074	0.274
				GRRC	0.410	0.113	0.111	0.012	0.952
				GRN	0.473	0.112	0.111	0.012	0.954
				Naive	−76.842	0.024	0.026	0.713	0.000
				Complete	0.810	0.118	0.118	0.014	0.955
		1	0.15	RC	−31.895	0.094	0.093	0.132	0.086
				RSRC	−24.394	0.111	0.110	0.084	0.329
				GRRC	0.373	0.116	0.115	0.013	0.954
				GRN	0.410	0.116	0.114	0.013	0.952
				Naive	−79.473	0.020	0.022	0.763	0.000
				Complete	0.719	0.118	0.118	0.014	0.956
			0.30	RC	−32.359	0.096	0.095	0.135	0.092
				RSRC	−24.540	0.115	0.113	0.086	0.351
				GRRC	0.391	0.116	0.114	0.013	0.957
				GRN	0.455	0.115	0.114	0.013	0.954
				Naive	−83.888	0.020	0.021	0.850	0.000
				Complete	0.737	0.118	0.118	0.014	0.956
	1	0.5	0.15	RC	−35.900	0.079	0.079	0.162	0.014
				RSRC	−26.916	0.095	0.094	0.096	0.163
				GRRC	0.328	0.114	0.112	0.013	0.950
				GRN	0.337	0.114	0.112	0.013	0.951
				Naive	−71.372	0.025	0.027	0.616	0.000
				Complete	0.819	0.118	0.118	0.014	0.955
			0.30	RC	−36.528	0.080	0.081	0.168	0.014
				RSRC	−27.334	0.098	0.097	0.100	0.181
				GRRC	0.337	0.114	0.112	0.013	0.949
				GRN	0.364	0.114	0.112	0.012	0.954
				Naive	−76.997	0.024	0.026	0.716	0.000
				Complete	0.728	0.118	0.118	0.014	0.956
		1	0.15	RC	−36.246	0.096	0.096	0.168	0.052
				RSRC	−28.409	0.114	0.113	0.110	0.253

(Continues)

TABLE 4 (Continued)

β_X	σ_ν^2	σ_ϵ^2	$\sigma_{\nu,\epsilon}$	Method	%Bias	ASE	ESE	MSE	CP
				GRRC	0.391	0.117	0.115	0.013	0.950
				GRN	0.401	0.116	0.115	0.013	0.950
				Naive	−80.256	0.020	0.022	0.778	0.000
				Complete	0.755	0.118	0.118	0.014	0.952
			0.30	RC	−36.674	0.098	0.097	0.172	0.056
				RSRC	−28.754	0.116	0.115	0.113	0.264
				GRRC	0.428	0.117	0.114	0.013	0.952
				GRN	0.446	0.116	0.114	0.013	0.954
				Naive	−84.015	0.020	0.021	0.852	0.000
				Complete	0.746	0.118	0.118	0.014	0.954

Note: For 2000 simulated data sets, the bias, average bootstrap standard error (ASE) for the four proposed estimators, average model standard error (ASE) for naive and complete case, empirical standard error (ESE), mean squared error (MSE), and 95% coverage probabilities (CP) are presented.

TABLE 5 Type 1 error results for $\beta_X = 0$ under correlated, additive measurement error in the outcome and covariates with normally distributed error and 25% and 75% censoring for the true event time

%Censoring	σ_ν^2	σ_ϵ^2	$\sigma_{\nu,\epsilon}$	Method	Type 1 error	ASE	ESE	MSE
25	0.5	0.5	0.15	RC	0.044	0.054	0.054	0.003
				RSRC	0.050	0.063	0.062	0.004
				GRRC	0.043	0.077	0.075	0.006
				GRN	0.042	0.078	0.075	0.006
				Naive	1.000	0.025	0.026	0.019
				Complete	0.049	0.097	0.097	0.010
75	0.5	0.5	0.15	RC	0.050	0.102	0.102	0.010
				RSRC	0.059	0.112	0.116	0.014
				GRRC	0.046	0.141	0.141	0.020
				GRN	0.046	0.143	0.143	0.021
				Naive	1.000	0.045	0.047	0.080
				Complete	0.056	0.170	0.178	0.032

Note: For 2000 simulated data sets, the type 1 error, average bootstrap standard error (ASE) for the four proposed estimators, average model standard error (ASE) for naive and complete case, empirical standard error (ESE), and mean squared error (MSE) are presented.

settings with a modest hazard ratio, common events, and small error variance under correlated outcome and covariate error. The proposed regression calibration methods were considered for the proportional hazards model; however, we expect they would work quite well more generally in accelerated failure time models where an additive error structure is assumed. In fact, some forms of error in the outcome will bias the proportional hazards parameter but not the acceleration parameter.[27]

The generalized raking estimators are consistent whenever the design-weighted complete case estimating equations (eg, HT estimator) yields consistent estimators; they use influence functions based on the unvalidated data as auxiliary variables to improve efficiency over the complete case estimator and can be used under outcome-dependent sampling. The raking estimators are not sensitive to the measurement error structure, which is in contrast to the RC and RSRC estimators that can perform poorly when the error structure is not correctly specified. In particular, we noticed in our data example and simulations that in the presence of misclassification as well as timing errors, GRRC and GRN yield

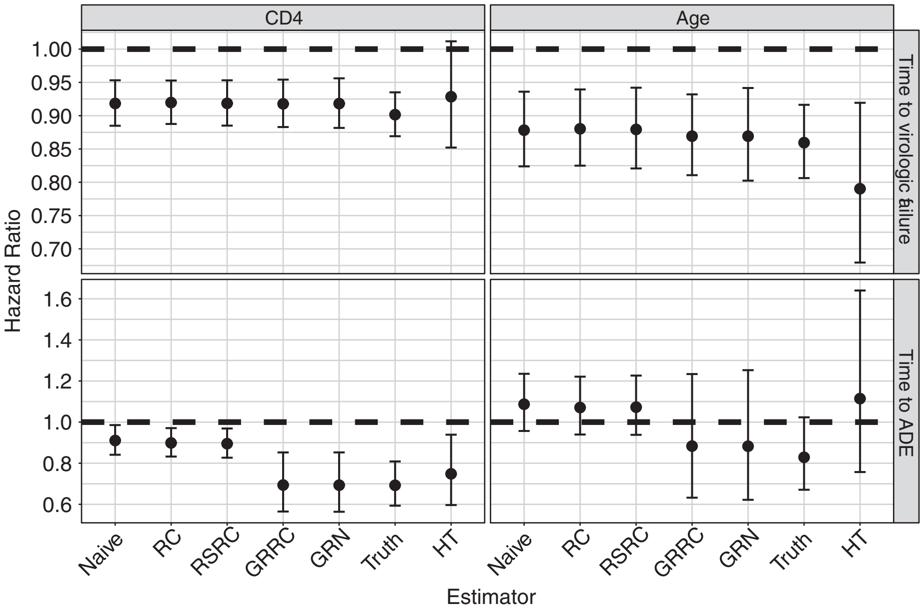

FIGURE 1 The hazard ratios and their corresponding 95% confidence intervals (CI) for a 100 cell/mm^3 increase in CD4 count at ART initiation and 10 year increase in age at CD4 count measurement. Estimates and their CIs are calculated using the bootstrap for the regression calibration (RC), risk set regression calibration (RSRC), generalized raking regression calibration (GRRC), and generalized raking naive (GRN) estimators

nearly unbiased estimates, while RC and RSRC are substantially biased. Generally, the raking estimators performed well, with little small sample bias and, in most cases, the smallest MSE. The raking estimators had large efficiency gains in settings with a large hazard ratio as well those with a modest hazard ratio, common events, and large error variances. For all settings considered, GRRC and GRN performed similarly. GRN has the added advantage that it can be applied with standard statistical software, for example, the survey package in R.[25]

As noted above, the performance of the GRRC and GRN estimators was virtually identical, contrary to our hypothesis that the GRRC estimates would be more efficient than those of GRN. This result was unknown for previous applications of raking[13,14] and in fact goes against their recommendation to build imputation models for the partially missing variables. For the setting of only classical covariate measurement error and no time-to-event error, we derived (not shown) that the influence functions for Cox regression using $X^{\star}$ vs $\hat{X}$ are scalar multiples of each other. Thus, the solutions to (7) under both auxiliary variables are equivalent. For the more complex error settings considered in this article (Sections 2.1 and 2.2), an explicit characterization of the relationship between the two auxiliary variables is more difficult, but we hypothesize that an approximation of a similar type holds for the settings studied.

The motivating example for this article was to develop methods where there were only errors in the failure time outcome but not in the failure indicator. We additionally considered methods, namely, GRRC and GRN, that are able to address more general error structures. We believe future research investigating RC methods to directly correct for misclassification resulting from time-to-event error would be worthwhile. In addition, while theory demonstrates that generalized raking estimators are consistent, we noticed that the small sample bias (and efficiency) can depend on the specific validation subsample. Developing optimal subsampling schemes to maximize efficiency would not only improve the complete case analysis, but also increase the efficiency gains of the raking estimators and is an area of future work.

ACKNOWLEDGEMENTS

The authors would like to thank Timothy Sterling, MD and the coinvestigators of the Vanderbilt Comprehensive Care Clinic (VCCC) for use of their data. This work was supported by a Patient Centered Outcomes Research Institute (PCORI) Award (R-1609-36207) and the US National Institutes of Health (NIH) (R01-AI131771, P30-AI110527, R01-AI093234, U01-AI069923, and U01-AI069918). The statements in this article are solely the responsibility of the authors and do not necessarily represent the views of PCORI or NIH.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the Vanderbilt Comprehensive Care Clinic (VCCC) with restrictions. A simulated data set that mimics the structure of the VCCC data and R code that supports the findings of this manuscript are available on GitHub at https://github.com/ericoh17/RRCME.

ORCID

Eric J. Oh https://orcid.org/0000-0002-7568-2371
Pamela A. Shaw https://orcid.org/0000-0003-1883-8410

REFERENCES

1. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton, FL: CRC Press; 2006.
2. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol*. 1997;146(2):195-203.
3. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am J Epidemiol*. 2013;177(9):904-912.
4. Wang L, Shaw PA, Mathelier HM, Kimmel SE, French B. Evaluating risk-prediction models using data from electronic health records. *Ann Appl Stat*. 2016;10(1):286.
5. Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*. 2003;59(4):947-954.
6. Magaret AS. Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Stat Med*. 2008;27(26):5456-5470.
7. Hunsberger S, Albert PS, Dodd L. Analysis of progression-free survival data using a discrete time survival model that incorporates measurements with and without diagnostic error. *Clin Trials*. 2010;7(6):634-642.
8. Shepherd BE, Yu C. Accounting for data errors discovered from an audit in multiple linear regression. *Biometrics*. 2011;67(3):1083-1091.
9. 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.
10. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*. 1982;69(2):331-342.
11. Shaw PA, He J, Shepherd BE. Regression calibration to correct correlated errors in outcome and exposure; 2018. arXiv preprint arXiv:1811.10147.
12. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.
13. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Stat Biosci*. 2009;1(1):32-49.
14. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. *Int Stat Rev*. 2011;79(2):200-220.
15. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982;10(4):1100-1120.
16. Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. *Biometrics*. 1997;53(1):131-145.
17. Shaw PA, Prentice RL. Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics*. 2012;68(2):397-407.
18. Xie SX, Wang CY, Prentice RL. A risk set calibration method for failure time regression by using a covariate reliability sample. *J Royal Stat Soc SerB (Stat Methodol)*. 2001;63(4):855-870.
19. Deville JC, Särndal CE. Calibration estimators in survey sampling. *J Am Stat Assoc*. 1992;87(418):376-382.
20. Deville JC, Särndal CE, Sautory O. Generalized raking procedures in survey sampling. *J Am Stat Assoc*. 1993;88(423):1013-1020.
21. Saegusa T, Wellner JA. Weighted likelihood estimation under two-phase sampling. *Ann Stat*. 2013;41(1):269-295.
22. Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. *J Am Stat Assoc*. 2004;99(467):832-844.
23. Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand J Stat*. 2007;34(1):86-102.
24. Lumley T. *Complex Surveys: A Guide to Analysis Using R*. Vol 565. Hoboken, NJ: John Wiley & Sons; 2011.
25. Lumley T. *Survey: Analysis of Complex Survey Samples. R Package Version 3.32*; 2016. https://cran.r-project.org/web/packages/survey/survey.pdf.
26. Lemly DC, Shepherd BE, Hulgan T, et al. Race and sex differences in antiretroviral therapy use and mortality among HIV-infected persons in care. *J Infect Dis*. 2009;199(7):991-998.
27. Oh EJ, Shepherd BE, Lumley T, Shaw PA. Considerations for analysis of time-to-event outcomes measured with error: bias and correction with SIMEX. *Stat Med*. 2018;37(8):1276-1289.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Oh EJ, Shepherd BE, Lumley T, Shaw PA. Raking and regression calibration: Methods to address bias from correlated covariate and time-to-event error. *Statistics in Medicine*. 2021;40:631–649. https://doi.org/10.1002/sim.8793