2020

Statistical Approaches To Address Correlated Measurement Error In A Failure-Time Outcome And Covariates

Eric Oh
University of Pennsylvania

Statistical Approaches To Address Correlated Measurement Error In A Failure-Time Outcome And Covariates

Abstract

Biomedical studies are increasingly relying on electronic health records (EHR) as either the sole or supplementary source of data. While these data sources have enormous potential to support the discovery of associations between exposures and disease risk, they are subject to measurement error, leading to bias in estimates of effects of interest. Covariate measurement error has been well studied in the literature, with published work spanning descriptions of its impact as well as methods to address it; however, errors in the outcome has not received as much attention. Furthermore, the error found in EHR data often involves errors in both covariates and a failure-time outcome that can be correlated. In this dissertation, we address these gaps by developing methodology in the paradigm of the Cox model for: (1) correlated errors in the time-to-event and covariate, (2) event-indicator misclassification as well as correlated time-to-event and covariate error, and (3) multiplicative error in the time-to-event. In Chapter 2, we develop two classes of estimators, regression calibration (RC) and generalized raking, to address the bias in Cox regression coefficients resulting from correlated errors in the time-to-event and covariate of interest. The RC estimators have lower relative MSE in moderate signal and high censoring settings; however, they are biased for the Cox model. The raking estimators are consistent, require no explicit modeling of the error structure, and have lower relative MSE for many error settings. In Chapter 3, we develop raking estimators for error settings involving misclassification by constructing auxiliary variables utilizing multiple imputation. We provide rationale for why the previously proposed raking estimators can be expected to be inefficient in the presence of event-indicator misclassification and demonstrate that the proposed raking estimators are more efficient in this setting. In Chapter 4, we compare the performance of the Cox and Weibull AFT models in error settings with random multiplicative time-to-event error. In addition, we develop an extension of the SIMEX method to correct the bias in hazard ratio estimates from the Cox model under multiplicative time-to-event error. We illustrate the proposed methods in the three chapters by applying them to observational EHR data on HIV outcomes from the Vanderbilt Comprehensive Care Clinic.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Pamela A. Shaw

Subject Categories

Biostatistics

STATISTICAL APPROACHES TO ADDRESS CORRELATED MEASUREMENT ERROR IN A FAILURE-TIME OUTCOME AND COVARIATES

Eric J. Oh

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

_____

Pamela A. Shaw, Associate Professor of Biostatistics

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Sharon X. Xie, Professor of Biostatistics

Jinbo Chen, Professor of Biostatistics

Eric J. Tchetgen Tchetgen, Professor of Statistics

Robert Gross, Associate Professor of Medicine & Epidemiology

STATISTICAL APPROACHES TO ADDRESS CORRELATED MEASUREMENT ERROR IN A FAILURE-TIME OUTCOME AND COVARIATES

ACKNOWLEDGEMENT

This dissertation is the culmination of five years of work. Five years of struggle and triumph and hardship and perseverance. Although it is my name on this dissertation, neither the highs nor the lows were borne alone. Although I do not have the space to appropriately thank everyone who has helped me along the way, I will do my best.

First and foremost, I would like to extend my deepest gratitude to my advisor Pam. I consider myself incredibly lucky to have gotten to work with her over the last five years. I came to Penn barely understanding rudimentary statistics and she patiently shaped me to be able to write this dissertation. Beyond the technical training, which is a baseline requirement for any advisor worth their salt in my eyes, Pam truly instilled in me a deep appreciation and pride in my work. Even on days when I was disillusioned with science, her enthusiasm for and enjoyment of her work rubbed off to motivate me through another week. This healthy passion for research is something I will take with me my entire lifetime.

To my committee members, Sharon, Jinbo, Eric, and Robert, thank you for all of your insightful comments and moral support. I want to thank Sharon, Jinbo, and Eric, in particular, for being such incredible teachers over the years. Between the three of you, I feel like I was taught everything I need to know about survival analysis, inference, semiparametric theory, and causal inference. Your wisdom, patience, and approach to research are things I will take with me forever.

To Peter and Raj, it has been an incredible honor and pleasure to have worked together the last four years. Our ECG work has been some of the most exciting I have worked on during my time at Penn and always provided me with a refreshing change of pace from my usual research. More than the work, I learned the power of collaborative research through our work and the opportunities that can arise from it. I thank you both for your mentorship over the years.

To all the past and present Penn biostat students who overlapped with me, I thank you all. I feel lucky to have been so inspired on a daily basis by the amazing work you all do. Special thanks to Julie, Jiaqi, Bret, Lior, Leah, Ali, Arman, and Angela for the many outings filled with good food, beer, and laughter.

To Mo, who was my only friend outside of the Penn biostat department who truly understood what

I was going through. I could not have done this without your support and friendship.

To my friends Jay, Awjin, Winnie, Caleb, Amy, Jonah, Raymond, Joey, and Linhan, I thank you for reminding me that there is much joy to be had outside of research. Even though we all live in different parts of the country, the reunions with everyone powered me through months of research. I would not have been able to make it this far without your friendship.

To my family, thank you for all of your sacrifice and love through the years. I would not have even made it to graduate school without your never-ending support and encouragement.

To Bryton, thank you for everything. This dissertation is as much yours as it is mine. You have been my emotional pillar through the good and the bad, never failing to lift my spirits. Every single day I wonder how I was lucky enough to find someone as supportive and as loving as you are. To say that I would not have made it through graduate school without you would be an understatement.

ABSTRACT

STATISTICAL APPROACHES TO ADDRESS CORRELATED MEASUREMENT ERROR IN A FAILURE-TIME OUTCOME AND COVARIATES

Eric J. Oh

Pamela A. Shaw

Biomedical studies are increasingly relying on electronic health records (EHR) as either the sole or supplementary source of data. While these data sources have enormous potential to support the discovery of associations between exposures and disease risk, they are subject to measurement error, leading to bias in estimates of effects of interest. Covariate measurement error has been well studied in the literature, with published work spanning descriptions of its impact as well as methods to address it; however, errors in the outcome has not received as much attention. Furthermore, the error found in EHR data often involves errors in both covariates and a failure-time outcome that can be correlated. In this dissertation, we address these gaps by developing methodology in the paradigm of the Cox model for: (1) correlated errors in the time-to-event and covariate, (2) event-indicator misclassification as well as correlated time-to-event and covariate error, and (3) multiplicative error in the time-to-event. In Chapter 2, we develop two classes of estimators, regression calibration (RC) and generalized raking, to address the bias in Cox regression coefficients resulting from correlated errors in the time-to-event and covariate of interest. The RC estimators have lower relative MSE in moderate signal and high censoring settings; however, they are biased for the Cox model. The raking estimators are consistent, require no explicit modeling of the error structure, and have lower relative MSE for many error settings. In Chapter 3, we develop raking estimators for error settings involving misclassification by constructing auxiliary variables utilizing multiple imputation. We provide rationale for why the previously proposed raking estimators can be expected to be inefficient in the presence of event-indicator misclassification and demonstrate that the proposed raking estimators are more efficient in this setting. In Chapter 4, we compare the performance of the Cox and Weibull AFT models in error settings with random multiplicative time-to-event error. In addition, we develop an extension of the SIMEX method to correct the bias in hazard ratio estimates from the Cox model under multiplicative time-to-event error. We illustrate the proposed methods in the three chapters by applying them to observational EHR data on HIV

outcomes from the Vanderbilt Comprehensive Care Clinic.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Modern biomedical studies are increasingly utilizing electronic health records (EHR) as either the sole or supplemental source of data. This surge in EHR utilization followed from the Health Information Technology for Economic and Clinical Health Act of 2009 (Adler-Milstein and Jha, 2017), which provided financial incentives for hospitals to adopt and demonstrate meaningful use of EHR systems and thus increased their availability for researchers. Accordingly, these data sources have enormous potential for medical discovery due to their low cost and large sample sizes compared to prospectively collected data. Specifically, EHRs collect patient-level information on healthcare resource utilization, clinical events, and various risk factors on a diverse population over time, allowing researchers to study associations between exposures and disease risk more quickly and cost-effectively.

The potential of EHR, however, carries with it the challenge that much of this data has been observed to be error-prone. For example, EHRs might fail to capture clinical events, data entry of risk factors could be based on patient recall or error-prone chart review, and the manual entry of data in EHR could be error-prone. These types of errors are magnified in statistical analyses due to the fact that variables of interest are often not directly observed in EHR data; instead they need to be derived from existing error-prone variables in the data, potentially causing correlations in errors. If such errors are not accounted for in the data analysis, estimated effects of interest can be biased, which in turn can mislead researchers and potentially harm patients down the line. There is a rich body of knowledge on the impact of and methods to handle measurement error in covariates, particularly for time-to-event outcomes. These include approximate methods such as regression calibration (Prentice, 1982) and SIMEX (Cook and Stefanski, 1994) as well as methods such as the corrected score (Huang and Wang, 2000, 2006; Nakamura, 1992) and conditional score (Tsiatis and Davidian, 2001) that are unbiased under various assumptions.

In contrast, errors in the failure-time outcome have not been as well studied. For linear models, it is known that random outcome error will not bias estimates of regression parameters; however, for nonlinear models such as the Cox model, even random error can bias regression parameter

estimates (Carroll et al., 2006). 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 some time 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 EHR for the date of diagnosis, both of which are subject to error. In addition, errors in the time origin can be systematic, as subject characteristics can influence the amount of error in recall. There has been some recent work to address errors in binary outcomes (Edwards et al., 2013; Magder and Hughes, 1997; Wang et al., 2016), discrete failure-time outcomes (Hunsberger, Albert, and Dodd, 2010; Magaret, 2008; Meier, Richardson, and Hughes, 2003); however, less attention has been given to errors in continuous failure-time outcomes.

Additionally, the measurement error found in EHR data often consists of errors in both covariates and a failure-time outcome that are correlated. This complexity stems from the fact that to utilize EHR data for statistical analyses, variables of interest are often derived from the existing variables in the data. For example, HIV/AIDS studies might be interested in evaluating the association between a lab measurement at the date of antiretroviral therapy (ART) initiation and the time from ART initiation to some event of interest. Both the exposure and outcome in the above example depend on the ART initiation date; thus, if the initiation date is incorrect, the outcome and covariate in the analysis will both contain measurement error that is correlated. This is an even more challenging statistical problem that has not received much attention in the literature. In this dissertation, we develop methodology to address errors in a continuous time-to-event outcome as well as correlated errors in both covariates and a continuous time-to-event outcome.

In Chapter 2, we develop two classes of estimators, regression calibration (RC) and generalized raking, that utilize an internal validation subset to address the bias resulting from correlated errors in the time-to-event and covariate of interest. The RC estimators estimate the true error-free variables with its expectation given the observed error-prone data. These estimates, however, are approximate for the Cox model and result in bias that can be appreciable for certain settings. The raking estimators are design-based estimators that incorporate the error-prone phase-one data as auxiliary variables to improve efficiency over the Horvitz-Thompson estimator. These estimators are consistent whenever the HT estimator yields consistent estimates and require no explicit modeling of the error structure. In Chapter 3, we develop generalized raking estimators that improve

the efficiency of the raking estimators considered in Chapter 2 for settings involving event-indicator misclassification. We demonstrate why the previously proposed raking estimators can be expected to be inefficient for these settings and propose raking estimators utilizing auxiliary variables constructed using multiple imputation to improve efficiency. Furthermore, we investigate the use of an outcome-dependent sampling design to select the validation subset to improve efficiency in rare-event settings. In Chapter 4, we consider the effects of random multiplicative error in the time-to-event outcome on the Cox and Weibull AFT models. In addition, we develop an extension of the SIMEX method to correct the bias in hazard ratio estimates from the Cox model under multiplicative time-to-event error. We illustrate the proposed methods in the three chapters by applying them to observational EHR data on HIV outcomes from the Vanderbilt Comprehensive Care Clinic.

CHAPTER 2

RAKING AND REGRESSION CALIBRATION: METHODS TO ADDRESS BIAS FROM CORRELATED COVARIATE AND TIME-TO-EVENT ERROR

2.1. Abstract

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.

2.2. 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 to 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

(Carroll et al., 2006); 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 non-linear models, even classical outcome errors can bias estimated associations of interest (Carroll et al., 2006). 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 some time 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 (Edwards et al., 2013; Magder and Hughes, 1997; Wang et al., 2016) and discrete failure time data (Hunsberger, Albert, and Dodd, 2010; Magaret, 2008; Meier, Richardson, and Hughes, 2003), 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 non-research purposes (e.g. 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 (Duda et al., 2012; Shepherd and Yu, 2011). 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 (e.g. 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 dataset. In this manuscript, 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 dataset.

Regression calibration (RC), introduced by Prentice (1982), is a method to address covariate measurement error that is widely used due to ease of implementation and good numerical performance in a broad range of settings. Although most RC methods assume measurement error in covariates only, Shaw, He, and Shepherd (2018) 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 (Robins, Rotnitzky, and Zhao, 1994) 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 manuscript. These survey sampling ideas, while not new, have not been carefully studied in the measurement error setting. Breslow et al. (2009) considered raking estimators for modeling case-cohort data with missing covariates. Lumley, Shaw, and Dai (2011) considered a raking estimator using simulated data in a covariate measurement error context with a validation subset. In this manuscript, 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 manuscript are twofold. 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 paper proceeds as follows. We present our survival time model and the considered measurement error frameworks in Section 2.3. Sections 2.4 and 2.5 present the proposed regression calibration and raking methods, respectively. Section 2.6 compares the relative performance of the proposed estimators with simulation studies for various parameter settings and error distributions. In Section 2.7, we apply our methods to an HIV cohort and ascertain their robustness to misclassification. We conclude with a discussion in Section 2.8.

2.3. 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 the data 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 $\beta = (\beta_X, \beta_Z)$.

$$\sum_{i=1}^{n} \int_0^{\tau} \left\{ \{X_i, Z_i\}' - \frac{n^{-1} \sum_{j=1}^{n} Y_j(t) \{X_j, Z_j\}' \exp(\beta_X' X_j + \beta_Z' Z_j)}{n^{-1} \sum_{j=1}^{n} Y_j(t) \exp(\beta_X' X_j + \beta_Z' Z_j)} \right\} dN_i(t) = 0 \qquad (2.1)$$

2.3.1. Additive Measurement Error Structure

Oftentimes, errors seen in electronic health records data or other datasets used for observational studies will not be simple random error and will depend on other variables in the dataset. 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

$$X^\star = \alpha_0 + \alpha_1' X + \alpha_2' Z + \epsilon \tag{2.2}$$

$$U^\star = U + \gamma_0 + \gamma_1' X + \gamma_2' Z + \nu = U + \omega. \tag{2.3}$$

Note that X and Z in the above formulation do not necessarily represent the full vector of covariates (e.g. 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 non-zero. 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.

2.3.2. 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 2.3.1 might be expected in scenarios involving only an error-prone baseline time (e.g. 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.

2.3.3. 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 (e.g. simple random sampling) and where the subsample size is random (e.g. Bernoulli sampling), as well as stratified designs (e.g. 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.

2.4. 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 (1982), 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. (1997). For more detail see Appendix A.1.

2.4.1. Regression Calibration for Covariate Error

Prentice (1982) introduced the regression calibration method for the setting of Cox regression and classical measurement error in the covariate. Shaw and Prentice (2012) applied regression calibration for the covariate error structure assumed in Section 2.3.1. The idea of regression calibration is to estimate the unobserved true variable with its expectation given the data. Prentice (1982) 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

$$\mathrm{E}(X | X^\star, Z) = \mu_X + \begin{bmatrix} \Sigma_{XX^\star} & \Sigma_{XZ} \end{bmatrix} \begin{bmatrix} \Sigma_{X^\star X^\star} & \Sigma_{X^\star Z} \\ \Sigma_{ZX^\star} & \Sigma_{ZZ} \end{bmatrix}^{-1} \begin{bmatrix} X^\star - \mu_{X^\star} \\ Z - \mu_Z \end{bmatrix}, \tag{2.4}$$

where the validation subset is used to calculate the moments involving X (see Shaw and Prentice (2012)). Define $\hat{X} = \mathrm{E}(X | X^\star, Z; \hat{\zeta}_x)$, where $\hat{\zeta}_x$ is the vector of nuisance parameters in equation (2.4) estimated from the data. $\hat{X}$ is then imputed for X in the partial likelihood score (2.1) instead of the observed $X^\star$ to solve for β, which yields the corrected estimates (Shaw and Pren-

tice, 2012). 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 $\mathrm{E}(X|X^\star, Z; \hat{\zeta}_x)$.

2.4.2. Proposed Regression Calibration Extension for Time-to-Event Error

Assume the time-to-event error structure in Section 2.3.1, i.e., we observe $(X, Z, U^\star, \Delta)$. Given the additivity of the outcome errors, 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

$$\mathrm{E}(\omega|X, Z; \zeta_\omega) = \mu_\omega + \begin{bmatrix} \Sigma_{\omega X} & \Sigma_{\omega Z} \end{bmatrix} \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}^{-1} \begin{bmatrix} X - \mu_X \\ Z - \mu_Z \end{bmatrix}, \tag{2.5}$$

where the validation subset is used to calculate the moments involving ω and ζ_ω is the vector of nuisance parameters in (2.5). 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 (2.1) for the corrected β estimates.

2.4.3. 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 2.3.1, i.e., 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 $X^\star$ as well, similar to Prentice (1982). 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 2.4.1. In addition, we propose to estimate $\mathrm{E}(\omega|X^\star, Z; \zeta_\omega)$ using the same first order approximation described in Section 2.4.2 except using $X^\star$ instead of X, 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 (2.1) instead of the observed $U^\star$ and $X^\star$ and solve for β to obtain our corrected estimates.

2.4.4. 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, Wang, and Prentice

(2001) for classical measurement error and extended to the covariate error model in Section 2.3.1 by Shaw and Prentice (2012). 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 (2.1) 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 2.4.2 and 2.4.3 at the i^{th} 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 Section 2.4.2 and Section 2.4.3.

2.5. Proposed Generalized Raking Methodology

In this section, we develop design-based estimators by applying generalized raking (raking for short) (Deville and Särndal, 1992; Deville, Särndal, and Sautory, 1993), 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 2.3.3. For the proof, see Saegusa and Wellner (2013).

2.5.1. 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 set of auxiliary variables that are available for everyone at phase one and are correlated with the phase two subsample variables. 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

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{n} R_i d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right) \\ &\text{subject to} \sum_{i=1}^{n} A_i = \sum_{i=1}^{n} R_i \frac{g_i}{\pi_i} A_i. \end{aligned} \tag{2.6}$$

Note that the constraints above are known as the calibration equations. Deville, Särndal, and Sautory (1993) 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

$$\sum_{i=1}^{n} R_i \frac{g_i}{\pi_i} P_i(\beta) = 0. \tag{2.7}$$

2.5.2. 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 (2.6) to ensure positive weights. Solving the constrained minimization problem for g_i then yields $g_i = \exp\left(-\hat{\lambda}' A_i\right)$. After plugging in g_i to the calibration equations, Deville and Särndal (1992) 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. (2009) derived the asymptotic expansion for the solution to (2.7) and showed that the optimal auxiliary variable is given by $A_i^{\text{opt}} = \mathrm{E}(\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)|V)$, 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 (2004)

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 *dfbeta* type residual, which is readily available in statistical software, to estimate the influence function from the approximate model. 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 $S^{(r)\star}(\beta,t) = n^{-1}\sum_{j=1}^n Y_j^\star(t) \left\{X_j^\star, Z_j\right\}^{'\otimes r} \exp(\beta_X' X_j^\star + \beta_Z' Z_j)$ ($a^{\otimes 1}$ 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 $A_{\text{RC},i} = \tilde{\ell}_0(\hat{X}_i(\hat{\zeta}_x), Z_i, \hat{U}_i(\hat{\zeta}_\omega), \Delta_i)$, i.e., the influence function for the target estimator that uses the calibrated estimates $(\hat{X}_i(\hat{\zeta}_x), \hat{U}_i(\hat{\zeta}))$ in place of the unobserved true data (X_i, U_i). One can again use the empirical approximation

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

where $\hat{S}^{(r)}(\beta,\hat{\zeta},t) = n^{-1}\sum_{j=1}^n \hat{Y}_j(t;\hat{\zeta}_\omega) \left\{\hat{X}_j(\hat{\zeta}_x), Z_j\right\}^{'\otimes r} \exp(\beta_X' \hat{X}_j(\hat{\zeta}_x) + \beta_Z' Z_j)$ ($a^{\otimes 1}$ is the vector a and $a^{\otimes 0}$ is the scalar 1). For measurement error settings including an error-prone failure indi-

cator, 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 (2.7) using $A_{N,i}$
2. Generalized raking regression calibration (GRRC): solution to (2.7) using $A_{\text{RC},i}$

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

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 (2007) 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 (2011).

2.5.3. 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 as a *dfbeta* type residual. 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 (2.7).

For step one, we consider the naive Cox model using the error-prone data (GRN) and the regression calibration approach described in Section 2.4 (GRRC). For step three, we utilize the *survey* package by Lumley (2016) in R, which provides standard software for obtaining raking estimates.

2.6. 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 2.3. These four estimators were compared to 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 all validation subsets were selected as simple random samples with known sampling probability, meaning the complete-case estimator is equivalent to the HT estimator. Following Section 2.3.1, 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 $\left((\alpha_0, \alpha_2) = \vec{0}; \alpha_1 = \vec{1}\right)$. In addition, we considered the general error structure described in Section 2.3.2, 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 4 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 $\alpha = 0.05$.

2.6.1. Simulation Set-up

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 $Z \sim \mathsf{N}(2, 1)$, 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 $\left(\Sigma_{\epsilon\epsilon} = \sigma^2_\epsilon, \Sigma_{\nu\nu} = \sigma^2_\nu\right) \in \{0.5, 1\}$, 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 (i.e., 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 and selected the validation subsets as simple random samples of size 200, or $\pi_i = \pi = 0.1$. The data example in Section 2.7 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 A.2 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.

2.6.2. Simulation Results

For all discussed tables, we observed that the naive estimates had very large bias with 95% coverage hovering around 0%. In 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 4 proposed estimators and how their relative performance differed across settings.

Table 2.1 presents the relative performance for estimating β_X in the presence of the time-to-event error described in Section 2.3.1 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 to RC for moderate β_X and common events and a range of $5 - 30\%$ bias reduction for other settings, with coverage around $87 - 93\%$ 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 2.2 and 2.3 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 2.3.1 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 parame-

Table 2.1: 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 time. 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.

$\%$ Censoring	β_X	σ^2_ν	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
	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

Table 2.2: 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. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	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
				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

Table 2.3: 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. 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.

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
				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

ters the same as in Tables 2.2 and 2.3, in Table 2.4 and Table A.1 in Appendix A.3. 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 2.5 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 2.1-2.4, are presented in Tables A.2-A.5 of Appendix A.3. The conclusions for this parameter were similar to those of β_X; however, the raking estimates had the lowest MSE across more settings. Tables A.6-A.8 in Appendix A.4 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 2.2-2.4. 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 2.1-2.4, in Tables A.9-A.12 of Appendix A.5. 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 A.13 in Appendix A.6 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 2.3.2. 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 unbiased because the raking estimators do not depend on the structure of the measurement error. Overall, the raking estimators had the

Table 2.4: 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. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	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
				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

Table 2.5: 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. For 2000 simulated data sets, the type 1 error, average bootstrap standard error (ASE) for the 4 proposed estimators, average model standard error (ASE) for naive and complete case, empirical standard error (ESE), and mean squared error (MSE) are presented.

$\%$ Censoring	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	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

lowest MSE in this more complex error setting.

2.7. 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 (Lemly et al., 2009). The VCCC fully validated all key variables for all records, resulting in an unvalidated, error-prone dataset and a fully validated dataset 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 to the naive and HT estimators. Note that the naive estimator was calculated using only the unvalidated dataset as if the validated dataset did not exist. In addition, the HT estimator was calculated using a subsample of the fully validated dataset. Throughout this example, we considered the estimates from the fully validated dataset 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 also have errors. Appendix A.7 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 dataset, 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 2.1 and shown in Table A.14 in Appendix A.8. 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 to 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 dataset, 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 paper do not explicitly handle misclassification, we were nevertheless interested in seeing how they would perform in this real data scenario in comparison to the raking methods that can handle misclassification. Due to ADE being a rare event, we utilized a case-cohort

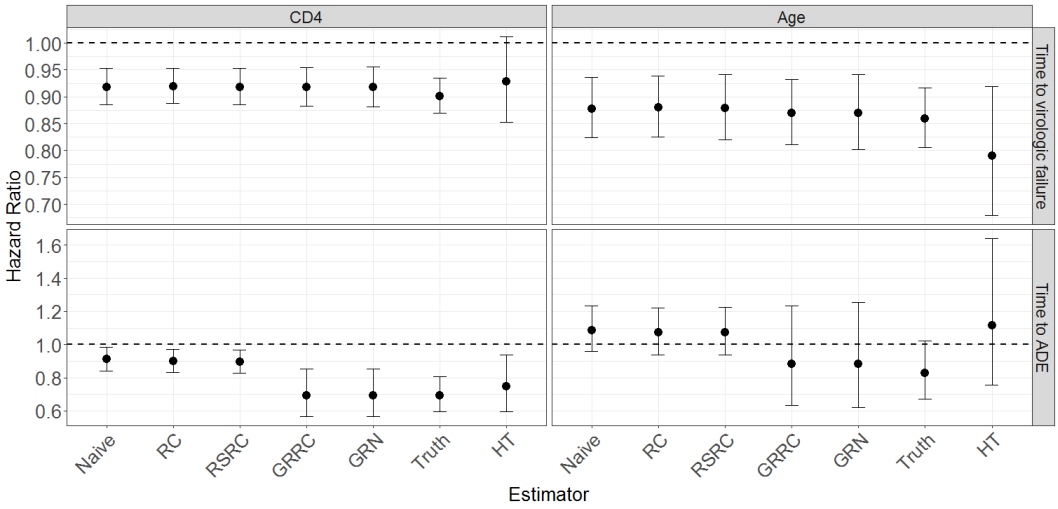


Figure 2.1: The hazard ratios and their corresponding 95% confidence intervals (CI) for a 100 cell/mm^3 increase in CD4 count at ART initiation and 10 year increase in age at CD4 count measurement. Estimates and their CIs are calculated using the bootstrap for the Regression Calibration (RC), Risk Set Regression Calibration (RSRC), Generalized Raking Regression Calibration (GRRC), and Generalized Raking Naive (GRN) estimators.

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 biased 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 one of calculating raking estimates as detailed in Section 2.5.3). The hazard ratios and their corresponding confidence intervals comparing the estimators are displayed graphically in the second row of Figure 2.1 and shown in Table A.14 in Appendix A.8. 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 fact, the naive point estimate for age was in the wrong direction compared to 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 ten different validation subsets using case-cohort sampling. The mean of the ten estimates is given in Table A.15 in Appendix A.8 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 to the naive estimators and decreased standard errors compared to 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 A.9 contains code that implements the RC and GRN estimators for this simulated data to demonstrate ease of application of these estimators.

2.8. Discussion

Data collected primarily for non-research 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 dataset as compared to the fully validated dataset 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 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 (Oh et al., 2018).

The generalized raking estimators are consistent whenever the design-weighted complete case estimating equations (e.g. 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 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, e.g. the survey package in R (Lumley, 2016).

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 (Breslow et al., 2009; Lumley, Shaw, and Dai, 2011) 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 $X^\star$ versus $\hat{X}$ are scalar multiples of each other. Thus, the solutions to (2.7) under both auxiliary variables are equivalent. For the more complex error settings considered in this paper (Sections 2.3.1, 2.3.2), 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 paper 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.

CHAPTER 3

IMPROVED GENERALIZED RAKING ESTIMATORS TO ADDRESS CORRELATED COVARIATE AND FAILURE-TIME OUTCOME ERROR

3.1. Abstract

Biomedical studies that use electronic health records (EHR) data for inference are often subject to bias due to measurement error. The measurement error present in EHR data is typically complex, consisting of errors of unknown functional form in covariates and the outcome, which can be dependent. To address the bias resulting from such errors, generalized raking has recently been proposed as a robust method that yields consistent estimates without the need to model the error structure. We provide rationale for why these previously proposed raking estimators can be expected to be inefficient in failure-time outcome settings involving misclassification of the event indicator. We propose raking estimators that utilize multiple imputation, to impute either the target variables or auxiliary variables, to improve the efficiency. We also consider outcome-dependent sampling designs and investigate their impact on the efficiency of the raking estimators, either with or without multiple imputation. We present an extensive numerical study to examine the performance of the proposed estimators across various measurement error settings. We then apply the proposed methods to our motivating setting, in which we seek to analyze HIV outcomes in an observational cohort with electronic health records data from the Vanderbilt Comprehensive Care Clinic.

3.2. Introduction

Modern biomedical studies are increasingly using non-traditional data sources such as electronic health records (EHR), which are not primarily collected for research purposes. These data sources have enormous potential to advance research of population-level health outcomes due to their large sample sizes and low cost compared to prospectively collected data (Beresniak et al., 2016; Hillestad et al., 2005; Jensen, Jensen, and Brunak, 2012; Staa et al., 2014). EHR data, however, have also been shown to be vulnerable to measurement error (Botsis et al., 2010; Duda et al., 2012; Floyd et al., 2012; Kiragga et al., 2011; Weiskopf and Weng, 2013). If such errors are not accounted for in the data analysis, estimated effects of interest can be biased, which in turn can mislead researchers and potentially harm patients.

The measurement error found in EHR data can be complex, consisting of errors in both an outcome and covariates of interest, which in turn can be dependent. This complexity stems from the fact that variables of interest are often not directly observed in EHR data; instead, they need to be derived from other existing variables in the data. For example, HIV/AIDS studies might be interested in evaluating the association between a lab value at the date of antiretroviral therapy (ART) initiation and the time from ART initiation to some event of interest. Both the exposure and outcome in the above example depend on the ART initiation date; thus, if the initiation date is incorrect, the outcome and covariate in the analysis will both contain measurement error that is dependent (in addition to potential misclassification of the event).

Covariate measurement error, particularly classical measurement error or extensions of it, has been well studied in the literature and methods to correct the bias resulting from such error have been well developed (Carroll et al., 2006). Although less attention has been given to errors in an outcome of interest, there has been some recent work looking at errors in binary outcomes (Edwards et al., 2013; Magder and Hughes, 1997; Wang et al., 2016), discrete time-to-event outcomes (Hunsberger, Albert, and Dodd, 2010; Magaret, 2008; Meier, Richardson, and Hughes, 2003), and to a lesser extent, continuous time-to-event outcomes (Gravel et al., 2018; Oh et al., 2018). There has been even less work to understand the impact of errors in both covariates and a time-to-event outcome, but it has recently been shown that ignoring such errors can cause severe bias in estimates of effects of interest (Boe, Tinker, and Shaw, 2020; Giganti et al., 2020; Oh et al., 2019).

In some cases, errors can be handled by retrospectively reviewing records and correcting all data points; however in most scenarios this will be too time-consuming and expensive to feasibly carry out. Instead, one can use a two-phase design, which involves reviewing and correcting only a subset of the records, to obtain consistent estimates of effects of interest. There have been some methods proposed recently that employ this framework to incorporate the large error-prone data with the smaller validated data to improve statistical inference, including regression calibration (Boe, Tinker, and Shaw, 2020; Oh et al., 2019), multiple imputation (Giganti et al., 2020), and generalized raking (Oh et al., 2019). Generalized raking in particular has been shown to be robust to the structure of the measurement error, which can be quite complex for EHR data (Han, Shaw, and Lumley, 2019; Oh et al., 2019). Specifically, generalized raking estimators use the error-prone data as auxiliary variables to improve the efficiency of the analysis of the validated data without having

to model the error structure, making them appealing for EHR settings where the true structure is likely unknown. Thus, we focus on the generalized raking methods in this manuscript.

In the measurement error setting, an error-prone version of the target variable is generally available on all subjects at phase one, which can be used to construct auxiliary variables for raking. While generalized raking estimators are robust, their statistical efficiency is dependent on the quality of the raking variables. Specifically, the efficiency of raking estimators depends on the (linear) correlation between the auxiliary variables and the target variable (Deville and Särndal, 1992). We show that for a time-to-event outcome, where the event indicator is subject to misclassification, this linear correlation is generally low and results in inefficient estimates. In this manuscript, we propose generalized raking estimators that utilize multiple imputation to construct improved auxiliary variables using imputed values of either the error-prone data or direct imputation of the auxiliary variables themselves to improve the linear correlation and ultimately, the efficiency of the raking estimator.

Our contributions in this manuscript are twofold. First, we develop generalized raking estimators that utilize multiple imputation to construct improved auxiliary variables in the presence of event indicator misclassification. Second, we evaluate the performance of various sampling designs with respect to their impact on the efficiency of the standard or proposed raking estimators. The rest of the paper proceeds as follows. We present our time-to-event outcome model and measurement error framework, and we introduce generalized raking estimators in Section 3.3. Section 3.4 discusses how the auxiliary variables relate to the efficiency of raking estimators and the need for their improvement in time-to-event settings with event indicator misclassification. Section 3.5 develops the proposed generalized raking estimators using multiple imputation. Section 3.6 compares the relative performance of the proposed estimators with simulation studies for various parameter settings and study designs. In Section 3.7, we apply our methods to evaluate HIV outcomes in an HIV cohort with error-prone EHR data. We conclude with a discussion in Section 3.8.

3.3. Model setup and design framework

This section introduces the design and estimation framework, including the time-to-event outcome model, measurement error framework, and generalized raking methods used to estimate parameters of interest.

3.3.1. Time-to-Event outcome 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.

In this paper, we consider estimating the parameters of a Cox proportional hazards model. 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. Then to estimate $\beta = (\beta_X, \beta_Z)$, we solve the partial likelihood score equation

$$\sum_{i=1}^{N} \int_0^{\tau} \left\{ \{X_i, Z_i\}' - \frac{\sum_{j=1}^{N} Y_j(t) \{X_j, Z_j\}' \exp(\beta_X' X_j + \beta_Z' Z_j)}{\sum_{j=1}^{N} Y_j(t) \exp(\beta_X' X_j + \beta_Z' Z_j)} \right\} dN_i(t) = 0 \tag{3.1}$$

Error framework

Instead of observing (X, Z, U, Δ), we observe $(X^\star, Z, U^\star, \Delta^\star)$, where $X^\star$, $U^\star$, and $\Delta^\star$ are the error-prone versions of X, U, and Δ, respectively. We do not impose any assumptions on the structure of the measurement error except that the error must have finite variance. In addition, we allow any of the errors to be correlated.

3.3.2. Two-phase design

We consider a retrospective two-phase design where at phase one, a set of possibly error-prone covariates and outcome information is collected on a large group of subjects. At phase two, the large cohort is augmented by selecting a subset of the subjects ($n < N$) to be validated, i.e., to have error-free covariates and outcome information measured. As a result, the phase two data is often referred to as the validation subset. Since the validation subset is selected retrospectively, the sampling probabilities are known. This type of sampling strategy accommodates both fixed subsample sizes (e.g. simple random sampling) as well as more complex designs with random subsample sizes (e.g. case-cohort). Specifically, let R_i be the indicator for whether subject $i = 1, \ldots, N$ is selected to be in the validation subset with known sampling probability $0 < \pi_i \leq 1$. Then

the observed data is given by $(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$ for $R_i = 0$ and $(X_i^\star, X_i, Z_i, U_i^\star, U_i, \Delta_i^\star, \Delta_i)$ for $R_i = 1$.

3.3.3. Generalized Raking

To estimate parameters in the two-phase design framework, we use generalized raking, a design-based estimator that combines the error-prone phase one data with the error-free phase two data to obtain efficient estimates that take advantage of all the measured data. Let β_0 denote the parameter defined by the population estimating equations $\sum_{i=1}^N \psi_i(\beta_0) = 0$. One classical estimator for two-phase designs is the Horvitz-Thompson (HT) estimator, $\hat{\beta}_{\mathrm{HT}}$, which is defined as the solution to $\sum_{i=1}^N \frac{R_i}{\pi_i}\psi_i(\beta) = 0$. Under suitable regularity conditions, $\hat{\beta}_{\mathrm{HT}}$ is a consistent estimator of β_0; however, it has been shown to be inefficient due to not using all of the available data at phase one (Robins, Rotnitzky, and Zhao, 1994). Let A_i denote a $p+q$-dimensional vector of auxiliary variables that are available for all N phase one subjects and correlated with the phase two data. Then generalized raking estimators modify the HT estimator design weights to new weights that incorporate the auxiliary variables such that $\sum_{i=1}^N A_i$, the known population total of auxiliary variables, is exactly estimated by the phase 2 subset. However, the new weights are constructed so that they are as close as possible to the HT weights while still satisfying the constraint. Specifically, for some distance measure $d(.,.)$, the objective can be written

$$\text{minimize} \quad \sum_{i=1}^N R_i d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right) \quad \text{subject to} \quad \sum_{i=1}^N A_i = \sum_{i=1}^N R_i \frac{g_i}{\pi_i} A_i,$$

where $\frac{g_i}{\pi_i}$ are the raking weights that can be solved for using Lagrange multipliers (Deville and Särndal, 1992). Note that the constraints above are known as the calibration equations. Therefore, the generalized raking estimator is defined by the solution to

$$\sum_{i=1}^N R_i \frac{g_i}{\pi_i}\psi_i(\beta) = 0. \tag{3.2}$$

Under suitable regularity conditions, the solution to (3.2) has been shown to be a $\sqrt{N}$ consistent and asymptotically normal estimator of β_0 (Saegusa and Wellner, 2013). When β_0 are the regression parameters from a correctly specified Cox proportional hazards model, $\psi_i(\beta) = \psi(X_i, Z_i, U_i, \Delta_i; \beta)$ is the Cox partial score equation (3.1) and the distance measure $d(a,b) = a\log(a/b) - a + b$ is used. Let λ denote a $p+q$-dimensional vector of Lagrange multipliers. Then solving the constrained minimization problem yields $g_i = \exp\left(\hat{\lambda}' A_i\right)$, where $\hat{\lambda} = \hat{B}^{-1}\left(\sum_{i=1}^N \frac{R_i}{\pi_i} A_i - \sum_{i=1}^N A_i\right) + O_p(n^{-1})$

and $\hat{B} = \sum_{i=1}^{N} \frac{R_i}{\pi_i} A_i' A_i$ (Deville and Särndal, 1992).

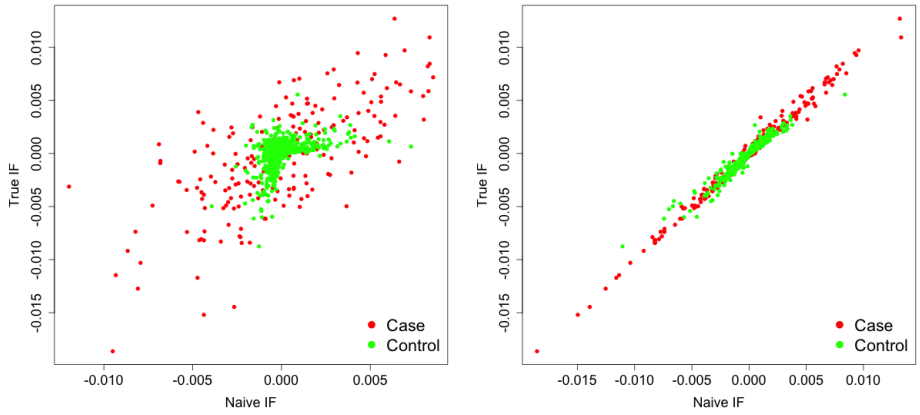
3.4. Construction of Better Auxiliary Variables

To quantify the gain in efficiency of raking estimators compared to the HT estimator, it is useful to consider the calibration equations, which constrain the raking weights to exactly estimate the known population total of the auxiliary variables. Deville and Särndal (1992) argued that "weights that perform well for the auxiliary variable also should perform well for the study variable" to provide support for such a construction. Note that study variable in this context represents the variable that is only observed on the phase two sample. Furthermore, there is an implicit assumption underlying this argument; namely that there exists a linear relationship between the variable of interest and the auxiliary variables of the form $S_i = \gamma_0 + \gamma_1 A_i + \epsilon_i$, where S_i and A_i are the variable of interest and auxiliary variables, respectively, and ϵ_i is random error. Thus, the efficiency gain of raking estimators depends directly on the (linear) correlation between the variable of interest and auxiliary variables. For more details, see Lumley, Shaw, and Dai (2011). The true relationship between S_i and A_i determines how to best use the auxiliary variables, which we hope to capture with the working model. If the true relationship between the study variable and auxiliary variables is nonlinear, standard generalized raking could be quite inefficient.

Assessing whether a linear working model is appropriate requires precise definitions for the variable of interest and auxiliary variables. In the setting of estimating regression parameters, many common estimators can be written as a population mean of influence function (or efficient influence function for semiparametric models) terms, $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$, using their asymptotically linear expansion. Thus, $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ is considered to be the variable of interest and the auxiliary variables should be constructed to be highly correlated with the influence function contributions. The optimal auxiliary variable was shown by Breslow et al. (2009) to be $\mathrm{E}(\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)|V)$, where $V = (X^\star, Z, U^\star, \Delta^\star)$, which is unavailable in practice. Oh et al. (2019), however, proposed an approximation, $\tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$, as the auxiliary variable, motivated by settings involving correlated measurement error in covariates and a censored event-time only.

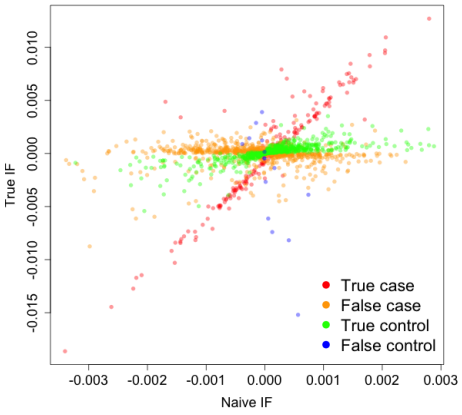
Thus, the linear working model underlying the estimator from Oh et al. (2019) is given by $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i) = \gamma_0 + \gamma_1 \tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i^\star) + \epsilon_i$. To assess whether the linear fit is appropriate, we plot $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ against $\tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$ from simulated data for various measurement error scenarios. Specifically, we plot empirical approximations of $\tilde{\ell}_0$ using delta-beta residuals (see

Figure 3.1: Plots of the true influence function $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ against the error-prone version $\tilde{\ell}_0^\star$ with the variables subject to measurement error noted in the graph subtitle. For example, (a) displays $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ against $\tilde{\ell}_0(X_i^\star, Z_i, U_i, \Delta_i)$. Univariate and normally distributed X and Z were generated. Survival times were generated from an exponential distribution with rate $\lambda_0 \exp(\beta_X X + \beta_Z Z)$, where $\lambda_0 = 0.1$, $\beta_X = \log(1.5)$, and $\beta_Z = \log(0.5)$, with 90% independent censoring. The error was generated as $X^\star = 0.2 + X - 0.1Z - 0.4\Delta + 0.25U + \epsilon$, $U^\star = U + \sigma_\nu \cdot 3 - 0.2X - 1.05Z + \nu$, and $\Delta^\star = \text{Bernoulli}\left(\text{expit}\left(-1.1 + 3\Delta - 0.3X - 0.2U + 0.1Z\right)\right)$, where (ϵ, ν) were normally distributed with $(\mu_\epsilon, \mu_\nu) = (0, 0)$, variances $(\sigma_\epsilon^2, \sigma_\nu^2) = (0.5, 0.5)$, and $\rho_{\epsilon,\nu} = 0.5$.

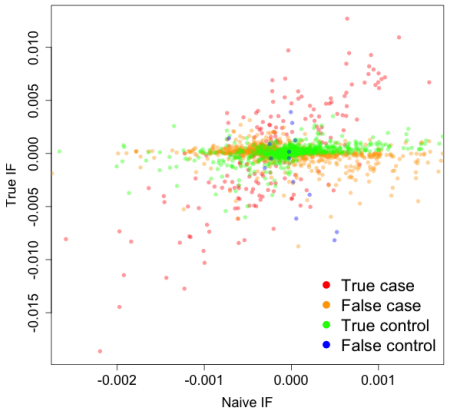


(a) X

(b) U



(c) Δ



(d) Δ, U, X

Oh et al. (2019) for more detail on their calculation) for settings with covariate error, time-to-event error, and misclassification only, as well as combinations of all three in Figure 3.1. The plots of $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ against $\tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$ for additive errors in the time-to-event or covariate show that the assumption of a linear relationship is mostly justified, albeit with some heteroscedasticity. However, when there is misclassification of the event indicator, a linear working model appears to be a very poor fit, and including additional errors in variables as in Figure 3.1d worsens the fit.

3.4.1. Model-calibration

Wu and Sitter (2001) proposed an alternative calibration method to handle settings where the true relationship between the variable of interest and the auxiliary variables may be nonlinear. Specifically, they assume the relationship between S_i and A_i can be characterized by the first and second moments, $\mathrm{E}(S_i|A_i) = \mu(A_i;\theta)$ and $\mathrm{Var}(S_i|A_i) = v_i^2\sigma^2$, where μ is a known function of A_i and θ, v is a known function of A_i or μ, and (θ, σ^2) are unknown parameters. Then using the validation subset, one obtains fitted values of $\mu(x_i;\theta)$, $\mu(x_i;\hat{\theta})$, and performs the raking procedure using them as auxiliary variables. Specifically, the generalized raking objective can be written as

$$\text{minimize} \quad \sum_{i=1}^{N} R_i d\left(\frac{g_i}{\pi_i}, \frac{1}{\pi_i}\right) \quad \text{subject to} \quad \sum_{i=1}^{N} \mu(x_i;\hat{\theta}) = \sum_{i=1}^{N} R_i \frac{g_i}{\pi_i} \mu(x_i;\hat{\theta}) \tag{3.3}$$

Wu and Sitter (2001) showed that this method yields more efficient estimates than the traditional raking estimator but still retains all of its statistical properties for a true nonlinear relationship between the variable of interest and auxiliary variables. Inspired by the model-calibration approach, we propose a data imputation approach that imputes the true Δ to obtain an auxiliary variable that has higher linear correlation with $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ than $\tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \Delta_i^\star)$ does. Additionally, we propose a novel application of the Wu and Sitter (2001) approach that directly imputes $\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i)$ based on a (potentially nonlinear) working model.

3.5. Proposed Multiple Imputation Methods for Generalized Raking

In this section, we propose methods to improve the efficiency of the generalized raking estimators under measurement error settings involving event indicator misclassification. Our methods use multiple imputation to impute the event indicator and then constructs new auxiliary variables using the imputed values to solve the raking estimating equation. For settings involving errors beyond just misclassification (e.g. additional time-to-event and/or covariate error), we propose a method using the fully conditional specification multiple imputation procedure that additionally imputes the

other error-prone variables iteratively. These methods are related to those of Han (2016), who proposed combining an empirical likelihood approach with multiple imputation to construct multiply robust estimators that are consistent if one of the sampling models or data generating models are correctly specified. Our approach differs in that we assume known phase two sampling probabilities possibly specified using a complex sampling design and study specific efficiency issues for time-to-event data. We additionally consider directly imputing the true population influence functions via a working model to use as auxiliary variables as a novel application of Wu and Sitter (2001). Lastly, we consider various study designs, including outcome-dependent sampling designs, for the selection of the validation subset in the two-phase design framework and discuss their varying impact on the efficiency of the proposed methods.

Note that due to raking being a design-based method, it will yield consistent estimates of the parameter that would be estimated with error-free data on the full cohort. The proposed methods all focus on adjusting the working model of the population influence functions to construct auxiliary variables closer to the optimal auxiliary variable. If the working model is misspecified, or does not capture the true relationship well, the proposed estimators still yield consistent and asympotically normal estimates (Breslow et al., 2009). If however, the working model is correct, the estimators will yield the most efficient design-consistent estimator (Han, 2016).

3.5.1. Multiple Imputation for the Event Indicator

Traditional multiple imputation in missing data settings (Rubin, 2004) involves developing statistical models for the distributions of the variables subject to missingness conditional on the fully observed variables. The missing variables are sampled M times from their distribution to generate M imputations of the missing data. The original data is augmented with the imputations, yielding M complete imputed datasets. Each of the M imputed datasets are then used to separately estimate the parameters of interest and the average of the M estimates is the multiple imputation estimator. The variance of the estimates can be calculated using Rubin's rules (Barnard and Rubin, 1999) or the estimators proposed by Robins and Wang (2000).

Multiple imputation for generalized raking follows similarly, with the exception that the M imputed datasets are first used to construct auxiliary variables for the influence functions for the target parameters.

First, we posit an imputation model for Δ, $f(\Delta|\Delta^\star, X^\star, U^\star, Z; \eta)$, with parameter vector η, and

specify a non-informative prior distribution, $f(\eta)$. We then fit the imputation model using the validation subset, generate the posterior distribution for η, and then sample M times from this posterior distribution to obtain $\eta_\star^{(1)}, \ldots, \eta_\star^{(M)}$. The parameter draws are used to sample $\hat{\Delta}_i^{(m)} \sim f(\Delta|\Delta_i^\star, X_i^\star, U_i^\star, Z_i; \eta_\star^{(m)})$ for all N phase one subjects and $m = 1, \ldots, M$. $\hat{\Delta}^{(1)}, \ldots, \hat{\Delta}^{(M)}$ are then augmented with the phase one data to yield M complete imputed datasets. Then for $m = 1, \ldots, M$, the estimating equation $\sum_{i=1}^N \psi(X_i^\star, Z_i, U_i^\star, \hat{\Delta}_i^{(m)}; \beta) = 0$ is solved to obtain $\hat{\beta}^{(m)}$. For each subject $i = 1, \ldots, N$, the auxiliary variable $\hat{A}_i$ is defined as

$$\hat{A}_i = \frac{1}{M} \sum_{m=1}^{M} \tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \hat{\Delta}_i^{(m)}; \hat{\beta}^{(m)}),$$

where $\tilde{\ell}_0(X_i^\star, Z_i, U_i^\star, \hat{\Delta}_i^{(m)})$ is the influence function for the estimating equation from the m-th imputation and can be empirically approximated as

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

where $S^{(r)\star}(\beta, t) = n^{-1} \sum_{j=1}^n Y_j^\star(t) \left\{X_j^\star, Z_j\right\}^{'\otimes r} \exp(\beta_X' X_j^\star + \beta_Z' Z_j)$ ($a^{\otimes 1}$ is the vector a and $a^{\otimes 0}$ is the scalar 1), $Y_j^\star(t) = I(U_j^\star \geq t)$, and $\hat{N}_i(t) = I(U_i^\star \leq t, \hat{\Delta}_i^{(m)} = 1)$.

Finally, to obtain estimates of the parameter of interest, we solve the raking estimating equation with adjusted weights calculated using $\hat{A}_i$ as auxiliary variables in (3.2).

3.5.2. Fully Conditional Specification Multiple Imputation

If there exists measurement error in variables beyond just the event indicator (e.g. additional time-to-event and/or covariate error), it is possible to gain efficiency by additionally imputing all error-prone variables iteratively using the fully conditional specification multiple imputation (FCSMI) method (Van Buuren, 2007). FCSMI involves specifying univariate models for the conditional distribution of each of the variables observed only at phase two given all phase one variables. Each missing variable is repeatedly imputed using the specified models and conditioning on the most recent imputations of the other variables. We explicate the FCSMI method for generalized raking in the presence of misclassification, covariate error, and time-to-event error. The method assumes a working model for the censored time-to-event that takes the form $U^\star = U + R(\Delta, X, Z)$, where

$R(\Delta, X, Z)$ is an arbitrary function of Δ, X, and Z. Note that if the working error model is misspecified, the raking estimator will still be consistent, albeit with some loss of efficiency.

First, we posit imputation models for Δ, X, and R, as well as non-informative prior distributions for their parameter vectors η, θ, and ω, respectively, to generate posterior distributions. We then draw parameters from their posteriors as follows: $\eta_\star^{(0)} \sim f(\Delta|\Delta^\star, X^\star, U^\star, Z; \eta_V) f(\eta_V)$, $\theta_\star^{(0)} \sim f(X|\Delta^\star, X^\star, U^\star, Z; \theta_V) f(\theta_V)$, and $\omega_\star^{(0)} \sim f(R|\Delta^\star, X^\star, Z; \omega_V) f(\omega_V)$. Then Δ, X, and U are imputed for all N phase one subjects by sampling from the imputation models using the initial parameter draws: $\hat{\Delta}^{(0)} \sim f(\Delta|\Delta^\star, X^\star, U^\star, Z; \eta_\star^{(0)})$, $\hat{X}^{(0)} \sim f(X|\Delta^\star, X^\star, U^\star, Z; \theta_\star^{(0)})$, and $\hat{U}^{(0)} = U^\star - \hat{R}^{(0)}$ where $\hat{R}^{(0)} \sim f(R|\Delta^\star, X^\star, Z; \omega_\star^{(0)})$. Then for iteration $l = 1, \ldots, L$, the algorithm proceeds as follows

$$
\begin{aligned}
\eta_\star^{(l)} &\sim f(\Delta|\Delta^\star, \hat{X}^{(l-1)}, \hat{U}^{(l-1)}, Z; \eta) f(\eta) \\
\hat{\Delta}^{(l)} &\sim f(\Delta|\Delta^\star, \hat{X}^{(l-1)}, \hat{U}^{(l-1)}, Z; \eta_\star^{(l)}) \\
\theta_\star^{(l)} &\sim f(X|\hat{\Delta}^{(l)}, X^\star, \hat{U}^{(l-1)}, Z; \theta) f(\theta) \\
\hat{X}^{(l)} &\sim f(X|\hat{\Delta}^{(l)}, X^\star, \hat{U}^{(l-1)}, Z; \theta_\star^{(l)}) \\
\omega_\star^{(l)} &\sim f(R|\hat{\Delta}^{(l)}, \hat{X}^{(l)}, Z; \omega) f(\omega) \\
\hat{U}^{(l)} = U^\star - \hat{R}^{(l)} \quad &\text{where} \quad \hat{R}^{(l)} \sim f(R|\hat{\Delta}^{(l)}, \hat{X}^{(l)}, Z; \omega_\star^{(l)})
\end{aligned}
$$

The algorithm continues sampling and imputing $\hat{\Delta}$, $\hat{X}$, and $\hat{U}$ for L iterations, after which it is assumed a stationary distribution has been reached. The above steps are repeated for M iterations, where $\hat{\Delta}^{(L)}$, $\hat{X}^{(L)}$, and $\hat{U}^{(L)}$ are taken to be the imputed values of Δ, X, and U, respectively, for each $m = 1, \ldots, M$. $\hat{\Delta}^{(m)}$, $\hat{X}^{(m)}$, and $\hat{U}^{(m)}$ are then augmented with the phase one data to yield M complete imputed datasets. Then for $m = 1, \ldots, M$, the estimating equation $\sum_{i=1}^{N} \psi(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)}; \beta) = 0$ is solved to obtain $\hat{\beta}^{(m)}$. Then the auxiliary variable for each subject, $\hat{A}_i$, is defined as

$$
\hat{A}_i = \frac{1}{M} \sum_{m=1}^{M} \tilde{\ell}_0(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)}; \hat{\beta}^{(m)}),
$$

and $\tilde{\ell}_0(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)})$ can be empirically approximated as

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

where $\hat{S}^{(r)}(\beta, t) = n^{-1} \sum_{j=1}^{n} \hat{Y}_j(t) \left\{ \hat{X}_j^{(m)}, Z_j \right\}^{'\otimes r} \exp(\beta_X' \hat{X}_j^{(m)} + \beta_Z' Z_j)$ ($a^{\otimes 1}$ is the vector a and $a^{\otimes 0}$ is the scalar 1), $\hat{Y}_j(t) = I(\hat{U}_j^{(m)} \geq t)$, and $\hat{N}_i(t) = I(\hat{U}_i^{(m)} \leq t, \hat{\Delta}_i^{(m)} = 1)$.

Lastly, to obtain estimates of the parameter of interest, we solve the raking estimating equation with adjusted weights calculated using $\hat{A}_i$ as auxiliary variables in (3.2).

3.5.3. Model-calibration multiple imputation

We propose a multiple imputation application of the Wu and Sitter (2001) model-calibration approach by specifying a working model for the population influence function and using the fitted values as auxiliary variables for raking in repeated iterations. First, we impute the error-prone variable(s) using MI or FCSMI as described in Sections 3.5.1 and 3.5.2. For the purposes of exposition, assume that FCSMI is used to impute Δ, X, and U to obtain $\hat{\Delta}^{(m)}$, $\hat{X}^{(m)}$, and $\hat{U}^{(m)}$. We posit a working model

$$\mathrm{E}(\tilde{\ell}_0(X_i, Z_i, U_i, \Delta_i) | \tilde{\ell}_0(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)})) = \mu(\tilde{\ell}_0(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)}); \gamma^{(m)}),$$

where $\tilde{\ell}_0(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)})$ is constructed using the empirical approximation given in Section 3.5.2. Here, μ can capture nonlinear relationships and the model is fit on the validation subset to obtain $\hat{\gamma}^{(m)}$. The above steps are repeated $m = 1, \ldots, M$ iterations to obtain $\hat{\gamma}^{(1)}, \ldots, \hat{\gamma}^{(M)}$. The auxiliary variable for each subject, $\hat{A}_i$, is then defined as

$$\hat{A}_i = \frac{1}{M} \sum_{m=1}^{M} \mu(\tilde{\ell}_0(\hat{X}_i^{(m)}, Z_i, \hat{U}_i^{(m)}, \hat{\Delta}_i^{(m)}); \hat{\gamma}^{(m)})$$

Finally, estimates of the parameter of interest are obtained by solving the raking estimating equation with adjusted weights calculated using $\hat{A}_i$ as auxiliary variables in (3.2).

3.5.4. Sampling Design Considerations

In validation study settings, such as those considered in this manuscript, researchers can define the phase two sampling probabilities as functions of the phase one data to select more informative subjects for increased efficiency. For example, researchers may want to oversample cases in rare-event settings or oversample subjects at underrepresented levels of informative covariates. Although generalized raking can easily accommodate such designs, the interplay between sampling designs and raking has not been well studied. We consider the effects of three different sampling designs on the efficiency of raking estimates: simple random sampling (SRS), case-control (CC), and covariate stratified case-control (SCC).

3.6. Simulation Study

In this section, we study the finite sample performance of the proposed raking estimators utilizing multiple imputation in the presence of event indicator misclassification. We compare these estimators to the raking estimator that constructs auxiliary variables using the naive error-prone data (GRN), the HT estimator, and the true estimator, i.e., the Cox proportional hazards model fit with the error-free data for all subjects. We considered three different measurement error scenarios where different variables are observed with error: 1) $(X, Z, U, \Delta^\star)$, 2) $(X, Z, U^\star, \Delta^\star)$, and 3) $(X^\star, Z, U^\star, \Delta^\star)$. For each error scenario, we considered the proposed raking estimator utilizing MI to impute the event indicator only, referred to as Generalized Raking Multiple Imputation (GRMI) hereafter. For error scenarios 2 and 3, which include errors in other variables besides the event indicator, we additionally considered the proposed raking estimator utilizing FCSMI to impute all error-prone variables iteratively, referred to as Generalized Raking Fully Conditional Specification Multiple Imputation (GRFCSMI) hereafter. We refer to these estimators as encompassing the data imputation approach. For all three error scenarios, we also considered the corresponding model-calibration multiple imputation methods described in Section 3.5.3, which we similarly refer to as encompassing the influence function (IF) imputation approach. We present $\%$ biases, average model standard errors (ASE), empirical standard errors (ESE), relative efficiency (RE) calculated with respect to the HT ESE, mean squared errors (MSE), and 95$\%$ coverage probabilities (CP) for varying values of the log hazard ratio β_X, $\%$ censoring, cohort and validation subset sizes, and validation subset sampling designs. We additionally present type 1 error results for $\beta_X = 0$ and $\alpha = 0.05$. All standard errors were calculated using sandwich variance estimators.

3.6.1. Simulation set-up

All simulations were run 2000 times using R version 3.6.2 (R Core Team, 2019). Cohort and validation subset sizes of $\{N, n\} = \{2000, 400\}$ and $\{N, n\} = \{10000, 2000\}$ were considered. Univariate X and Z were considered and were generated as a bivariate normal distribution with means $(\mu_X, \mu_Z) = (0, 2)$, variances $(\sigma_X^2, \sigma_Z^2) = (1, 1)$, and $\rho_{X,Z} = 0.5$. The true log hazard ratios were set to be $\beta_X \in \{\log(1.5), \log(3)\}$ and $\beta_Z = \log(0.5)$. 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 $\lambda_0 = 0.1$. Censoring times were simulated for each β_X and β_Z to yield 50%, 75%, and 90% censoring rates. Specifically, they were generated from Uniform distributions of varying lengths to mimic studies of different lengths.

The error-prone data were generated as follows:

1. Scenario 1: $(X, Z, U, \Delta^\star)$, where

$$\Delta^\star = \text{Bernoulli}\left(\text{expit}\left(-1.1 + 3\Delta - 0.3X - 0.2U + 0.1Z\right)\right)$$

2. Scenario 2: $(X, Z, U^\star, \Delta^\star)$, where

$$\Delta^\star = \text{Bernoulli}\left(\text{expit}\left(-1.1 + 3\Delta - 0.3X - 0.2U + 0.1Z\right)\right)$$
$$U^\star = U + R = U + \sigma_\nu \cdot 3 - 0.2X - 1.05Z + \nu$$

3. Scenario 3: $(X^\star, Z, U^\star, \Delta^\star)$, where

$$\Delta^\star = \text{Bernoulli}\left(\text{expit}\left(-1.1 + 3\Delta - 0.3X - 0.2U + 0.1Z\right)\right)$$
$$U^\star = U + R = U + \sigma_\nu \cdot 3 - 0.2X - 1.05Z + \nu$$
$$X^\star = 0.2 + X - 0.1Z - 0.4\Delta + 0.25U + \epsilon$$

Note that the choice of the intercept term in the event time error model is such that the error-prone time is a valid event time (i.e., 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 scenario 3, the error terms (ϵ, ν) were generated from a bivariate normal distribution with means $(\mu_\epsilon, \mu_\nu) =$

$(0, 0)$, variances $(\sigma^2_\epsilon, \sigma^2_\nu) = (0.5, 0.5)$, and $\rho_{\epsilon,\nu} = 0.5$. ν was generated from a univariate normal distribution for scenario 2 with the same mean and variance as in scenario 3. Table B.1 presents the sensitivity, specificity, positive predictive value, and negative predictive value for the misclassified event indicator across all error scenarios.

For the working imputation models, we fit logistic regression models for Δ and linear regression models for X and R. Under the error generating process considered in this section, analytical expressions for the true imputation models do not exist. Therefore, we considered two types of working imputation models: those including only main effects and those additionally adding all possible interaction effects to potentially specify an imputation model closer to the truth. Specifically, the imputations models including only main effects (referred to as Generalized Raking Multiple Imputation Simple (GRMIS) and Generalized Raking Fully Conditional Specification Multiple Imputation Simple (GRFCSMIS) hereafter) were specified as follows:

1. Scenario 1: $(X, Z, U, \Delta^\star)$

$$\text{logit}(P(\Delta = 1)|\Delta^\star, X, U, Z) = \eta_0 + \eta_1\Delta^\star + \eta_2 X + \eta_3 U + \eta_4 Z$$

2. Scenario 2: $(X, Z, U^\star, \Delta^\star)$

$$\begin{aligned}\text{logit}(P(\Delta = 1)|\Delta^\star, X, U^\star, Z) &= \eta_0 + \eta_1\Delta^\star + \eta_2 X + \eta_3 U^\star + \eta_4 Z \\ \text{E}(R|\Delta^\star, X, Z) &= \omega_0 + \omega_1\Delta^\star + \omega_2 X + \omega_3 Z\end{aligned}$$

3. Scenario 3: $(X^\star, Z, U^\star, \Delta^\star)$

$$\begin{aligned}\text{logit}(P(\Delta = 1)|\Delta^\star, X^\star, U^\star, Z) &= \eta_0 + \eta_1\Delta^\star + \eta_2 X^\star + \eta_3 U^\star + \eta_4 Z \\ \text{E}(R|\Delta^\star, X^\star, Z) &= \omega_0 + \omega_1\Delta^\star + \omega_2 X^\star + \omega_3 Z \\ \text{E}(X|\Delta^\star, X^\star, U^\star, Z) &= \theta_0 + \theta_1\Delta^\star + \theta_2 X^\star + \theta_3 U^\star + \theta_4 Z\end{aligned}$$

The imputation models containing interaction terms (referred to as Generalized Raking Multiple Imputation Complex (GRMIC) and Generalized Raking Fully Conditional Specification Multiple Im-

putation Complex (GRFCSMIC) hereafter) include the same predictors as above as well as all possible interaction terms. For each error scenario and all parameter settings, the number of imputation iterations was set to 50 and the FCSMI estimators performed 500 iterative updates to the imputed variables per imputation iteration. Appendix B.1 provides further detail on the implementation of the multiple imputation procedures. For the IF imputation approach, linear regression models were fit for the working models of the true influence function for each covariate. For example, the following model was fit for error scenario 1:

$$\begin{aligned}\mathrm{E}(\tilde{\ell}_0|\hat{\tilde{\ell}}_0) = \gamma_0 + \gamma_1\hat{\tilde{\ell}}_0 + \gamma_2\hat{\Delta} + \gamma_3 U + \gamma_4 X + \gamma_5 Z \\ + \gamma_6\left(\hat{\tilde{\ell}}_0 \times \hat{\Delta}\right) + \gamma_7\left(\hat{\tilde{\ell}}_0 \times U\right) + \gamma_8\left(\hat{\tilde{\ell}}_0 \times X\right) + \gamma_9\left(\hat{\tilde{\ell}}_0 \times Z\right).\end{aligned}$$

For error scenarios 2 and 3, the same models were fit except U and X were replaced by $\hat{U}$ and $\hat{X}$.

We considered validation subsets selected via simple random sampling for all three error scenarios. For the rare-event setting of 90% censoring in error scenarios 2 and 3, we additionally compared the performance of the estimators using validation subsets selected via case-control sampling and stratified case-control sampling. For these sampling design comparisons, we considered $\{N, n\} = \{4000, 800\}$ and generated the error-prone event indicator according to the model described in Table B.2. The covariate and time-to-event error were generated using the same previous models. To perform case-control sampling, all error-prone cases were selected and a simple random sample of error-prone controls were selected to yield a nearly one-to-one ratio of error-prone cases to controls. To perform stratified case-control sampling, we stratified the continuous covariate X (or $X^\star$ for settings involving covariate error) into four discrete categories by setting cutpoints at the 20^{th}, 50^{th}, and 80^{th} percentiles. We then selected an equal number of subjects from each of the eight strata defined by the combinations of the error-prone case status and the covariate strata (i.e., the balanced sampling design proposed by Breslow and Chatterjee (1999)) . Note that for CC and SCC sampling, the data imputation models and influence function working models for the IF imputation approach were inverse-probability weighted to account for the sampling design of the validation subsets. For the proposed raking estimators utilizing MI or FCSMI for data imputation only, the imputation models were not weighted as we included all stratification variables in the models (Cochran, 2007) and we noticed no empirical differences between including weights or not.

3.6.2. Simulation results

In the scenarios considered, all of the considered estimators were nearly unbiased for all settings, as expected, with the exception of a few specific rare-event settings with $\{N, n\} = \{2000, 400\}$ and simple random sampling, due to relatively few true events (40 on average) in the validation subset. Since the proposed estimators construct improved auxiliary variables to increase efficiency compared to GRN, we focus on the ESE, RE (with respect to the HT estimator), MSE, and CP and how these performance measures differed across settings.

Table 3.1 presents the results under error scenario 1 for estimating $\beta_X \in \{\log(1.5), \log(3)\}$ using the data imputation approach for $\{N, n\} = \{2000, 400\}$, $\{50\%, 75\%, 90\%\}$ censoring, and simple random sampling of the validation subset. GRN had increased efficiency compared to HT with the RE ranging from 1.24 for 50% censoring to 1.06 for 90% censoring. However, GRMIS and GRMIC both had higher RE than GRN for nearly all parameter settings, ranging from 1.41 for 50% censoring to 1.16 for 90% censoring. GRMIS and GRMIC had comparable REs, lower MSE than HT and GRN, and CPs near 95% for all parameter settings.

Table B.3 presents the results under error scenario 2 for estimating $\beta_X \in \{\log(1.5), \log(3)\}$ using the data imputation approach for $\{N, n\} = \{2000, 400\}$, $\{50\%, 75\%, 90\%\}$ censoring, and simple random sampling of the validation subset. GRN again had increased efficiency compared to HT with the RE ranging from 1.21 to 1.07. GRMIS, GRMIC, GRFCSMIS, and GRFCSMIC, however, all had higher RE than GRN for all parameter settings, ranging from 1.43 to 1.14 for GRMI and 1.45 to 1.14 for GRFCSMI. Comparing GRMIS to GRMIC and GRFCSMIS to GRFCSMIC, we observed nearly no difference in efficiency. Comparing GRMI to GRFCSMI, GRFCSMI had higher or equal RE for nearly all settings, although the difference was sometimes small. In addition, GRMI and GRFCSMI had lower MSE than HT and GRN and CPs by $5 - 6\%$ for all settings.

Table 3.2 presents the results under error scenario 3 for estimating $\beta_X \in \{\log(1.5), \log(3)\}$ using the data imputation approach for $\{N, n\} = \{2000, 400\}$, $\{50\%, 75\%, 90\%\}$ censoring, and simple random sampling of the validation subset. In this more complex error scenario, GRN had a small improvement in efficiency over HT, with its RE peaking around 1.05 across all settings. GRMIS and GRMIC similarly showed minor efficiency improvements compared to HT with its RE ranging from 1 to 1.06. However, GRFCSMIS and GRFCSMIC had appreciable gains in efficiency, with RE

Table 3.1: Simulation results for estimating β_x using the data imputation approach for error scenario 1 (error only in event indicator) with $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.03595	0.039644	2.289193	0.039422	0.001572	0.956
			HT	1.228958	0.090753	1	0.087874	0.008261	0.949
			GRN	1.40684	0.07401	1.226214	0.072528	0.00551	0.95
			GRMIS	-0.08937	0.065693	1.381469	0.06386	0.004316	0.948
			GRMIC	-0.42527	0.064598	1.404892	0.063389	0.004176	0.946
		log(3)	True	0.041168	0.041582	2.453957	0.04415	0.001729	0.948
			HT	0.631089	0.10204	1	0.097775	0.01046	0.939
			GRN	0.282312	0.082568	1.235824	0.080447	0.006827	0.942
			GRMIS	0.108883	0.072166	1.413959	0.069818	0.005209	0.948
			GRMIC	0.007399	0.072275	1.411835	0.069152	0.005224	0.948
	75	log(1.5)	True	0.119394	0.051672	2.266392	0.053276	0.00267	0.954
			HT	0.781363	0.117109	1	0.118644	0.013725	0.952
			GRN	0.916624	0.097339	1.203106	0.096548	0.009489	0.945
			GRMIS	0.188371	0.093537	1.252017	0.091773	0.00875	0.944
			GRMIC	0.096736	0.096302	1.216058	0.090939	0.009274	0.94
		log(3)	True	-0.01311	0.06088	2.241353	0.059211	0.003706	0.949
			HT	1.034735	0.136454	1	0.131041	0.018749	0.938
			GRN	0.386125	0.119288	1.143905	0.113786	0.014248	0.934
			GRMIS	0.197862	0.102954	1.325384	0.102518	0.010604	0.943
			GRMIC	0.040924	0.101157	1.348933	0.101394	0.010233	0.944
	90	log(1.5)	True	0.0138	0.084364	2.222885	0.083155	0.007117	0.947
			HT	1.805251	0.187531	1	0.184444	0.035222	0.943
			GRN	0.30929	0.167181	1.121725	0.165789	0.027951	0.94
			GRMIS	0.192308	0.161702	1.159732	0.160033	0.026148	0.944
			GRMIC	-0.55691	0.159657	1.174587	0.158312	0.025495	0.936
		log(3)	True	-0.04654	0.088525	2.315872	0.089229	0.007837	0.95
			HT	1.160558	0.205013	1	0.197598	0.042193	0.938
			GRN	0.945284	0.194363	1.054793	0.187058	0.037885	0.941
			GRMIS	0.26163	0.175215	1.17007	0.16969	0.030708	0.94
			GRMIC	-0.31527	0.17402	1.178102	0.169034	0.030295	0.939

ranging from 1.12 to 1.25 for all settings except for 90% censoring, where the RE was less than 1.1. These efficiency gains suggest that, in the presence of covariate measurement error that depends on the outcome, multiply imputing all error-prone variables was advantageous over only imputing the misclassified event indicator. Overall, GRFCSMI had lower MSE than all other estimators (albeit with some bias for 90% censoring) and CPs that ranged from $94 - 95\%$ for all settings.

Results for $\{N, n\} = \{10000, 2000\}$, keeping all other parameters the same as Table 3.1, Table B.3, and Table 3.2, are presented in Tables B.4, B.5, and B.6, respectively. The conclusions for these large cohort settings were similar to those with $\{N, n\} = \{2000, 400\}$. For error scenario 1, GRMI provided appreciable efficiency gain over GRN. For error scenario 2, both GRMI and GRFCSMI provided comparable and significant efficiency gain over GRN. For error scenario 3, only GRFCSMI yielded appreciable efficiency gain over GRN and both GRMI and GRFCSMI were nearly unbiased even with 90% censoring.

We present the type 1 error results under error scenario 3 for estimating $\beta_X = 0$ using the data imputation approach for $\{N, n\} = \{10000, 2000\}$, $\{50\%, 75\%, 90\%\}$ censoring, and simple random sampling of the validation subset in Table B.7. For the 50% and 75% censoring levels, the type 1 error of the proposed GRMI and GRFCSMI estimators ranged from 0.052 to 0.064. For the 90% censoring setting, the number of cases in the phase two data was very small at 40, and the type 1 error ranged from 0.068 to 0.072 for the proposed methods. However, we note that the type 1 error could likely be improved by using the bootstrap to calculate standard errors instead of the sandwich variance estimators (see Oh et al. (2019) for more detail).

Results for the IF imputation approach under error scenario 3 for $\{N, n\} = \{2000, 400\}$, keeping all other parameters the same as Table 3.2, are presented in Table 3.3. We note that the RE of the proposed estimators cannot be directly compared to those from the data imputation tables due to the HT ESE varying slightly. Overall, the conclusions for this approach were very similar to those of the data imputation approach. We observed that GRFCSMI was more efficient (by RE) and had lower MSE than all other estimators, albeit with some bias. Comparing the IF imputation estimators to the data imputation estimators, the ESE was very similar across all settings; this suggests that in the relatively simple error settings considered, the data imputation improved most of the auxiliary variable nonlinearity issues. Similar tables for error scenarios 1 and 2 are presented in Tables B.8 and B.9 and similar conclusions were reached. Results for the IF imputation approach

Table 3.2: Simulation results for estimating β_x using the data imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	0.07661	0.039569	2.328964	0.039418	0.001566	0.95
			HT	1.342474	0.092155	1	0.088213	0.008522	0.937
			GRN	2.100967	0.093898	0.98143	0.087678	0.008889	0.928
			GRMIS	1.308134	0.092762	0.993454	0.088003	0.008633	0.935
			GRMIC	1.276032	0.092487	0.996408	0.088048	0.008581	0.935
			GRFCSMIS	0.798683	0.075276	1.224217	0.074605	0.005677	0.948
			GRFCSMIC	0.407051	0.073879	1.247364	0.074138	0.005461	0.942
		log(3)	True	-0.00837	0.041674	2.491381	0.04412	0.001737	0.951
			HT	0.777852	0.103825	1	0.097835	0.010853	0.944
			GRN	1.177403	0.101197	1.02597	0.097568	0.010408	0.943
			GRMIS	0.846726	0.103247	1.005597	0.097632	0.010747	0.944
			GRMIC	0.816276	0.103057	1.007453	0.097623	0.010701	0.945
			GRFCSMIS	0.60425	0.088082	1.178736	0.087678	0.007802	0.939
			GRFCSMIC	0.333361	0.088859	1.168426	0.087836	0.007909	0.938
	75	log(1.5)	True	-0.11172	0.050646	2.445003	0.053272	0.002565	0.946
			HT	2.494616	0.12383	1	0.119095	0.015436	0.945
			GRN	3.469044	0.121951	1.015403	0.116576	0.01507	0.944
			GRMIS	3.493033	0.123515	1.002552	0.117963	0.015456	0.94
			GRMIC	3.755253	0.123088	1.006027	0.117855	0.015383	0.938
			GRFCSMIS	1.830123	0.107038	1.156879	0.103193	0.011512	0.946
			GRFCSMIC	1.848374	0.106441	1.163367	0.102455	0.011386	0.947
		log(3)	True	-0.01819	0.05804	2.37266	0.05929	0.003369	0.948
			HT	0.939605	0.13771	1	0.13192	0.019071	0.95
			GRN	1.291495	0.133879	1.028617	0.129447	0.018125	0.947
			GRMIS	1.13204	0.134928	1.020621	0.130678	0.01836	0.948
			GRMIC	1.21211	0.137343	1.002675	0.130285	0.01904	0.947
			GRFCSMIS	0.749482	0.123367	1.116261	0.119853	0.015287	0.946
			GRFCSMIC	0.725826	0.120588	1.14199	0.119671	0.014605	0.944
	90	log(1.5)	True	0.0138	0.084364	2.227607	0.083155	0.007117	0.947
			HT	2.839981	0.18793	1	0.184457	0.03545	0.944
			GRN	4.005694	0.180168	1.043079	0.178185	0.032724	0.94
			GRMIS	4.361114	0.177508	1.05871	0.17808	0.031822	0.937
			GRMIC	4.460343	0.178246	1.054326	0.176558	0.032099	0.936
			GRFCSMIS	1.373064	0.176884	1.062444	0.170686	0.031319	0.943
			GRFCSMIC	2.936905	0.173456	1.08344	0.169147	0.030229	0.938
		log(3)	True	-0.04654	0.088525	2.300257	0.089229	0.007837	0.95
			HT	0.99248	0.203631	1	0.198896	0.041584	0.945
			GRN	1.644558	0.192597	1.05729	0.193718	0.03742	0.942
			GRMIS	1.503862	0.196311	1.037287	0.192159	0.038811	0.945
			GRMIC	1.581827	0.19941	1.021165	0.19132	0.040067	0.943
			GRFCSMIS	1.162629	0.195142	1.043502	0.189566	0.038243	0.947
			GRFCSMIC	1.150689	0.196691	1.035282	0.188478	0.038847	0.946

Table 3.3: Simulation results for estimating β_x using the IF imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.03595	0.039644	2.357942	0.039422	0.001572	0.956
			HT	0.968647	0.093478	1	0.087968	0.008754	0.944
			GRN	2.48905	0.092254	1.013273	0.087589	0.008613	0.942
			GRMIS	1.27812	0.095618	0.977622	0.082415	0.00917	0.904
			GRMIC	0.702605	0.094581	0.988339	0.082057	0.008954	0.912
			GRFCSMIS	1.176668	0.076746	1.218027	0.072894	0.005913	0.932
			GRFCSMIC	0.766525	0.076361	1.224153	0.072638	0.005841	0.938
		log(3)	True	0.041168	0.041582	2.490834	0.04415	0.001729	0.948
			HT	0.313211	0.103573	1	0.097851	0.010739	0.942
			GRN	0.725322	0.104082	0.995114	0.097532	0.010897	0.945
			GRMIS	1.421894	0.102001	1.015417	0.091883	0.010648	0.924
			GRMIC	1.487215	0.102937	1.006184	0.091601	0.010863	0.926
			GRFCSMIS	0.262352	0.096256	1.076016	0.08654	0.009274	0.934
			GRFCSMIC	0.102202	0.095132	1.088739	0.08656	0.009051	0.934
	75	log(1.5)	True	0.119394	0.051672	2.316876	0.053276	0.00267	0.954
			HT	1.004049	0.119718	1	0.118566	0.014349	0.948
			GRN	1.661829	0.119968	0.997919	0.116507	0.014438	0.945
			GRMIS	4.68564	0.122646	0.976125	0.107653	0.015403	0.92
			GRMIC	5.039218	0.121839	0.98259	0.107163	0.015262	0.916
			GRFCSMIS	1.012435	0.104425	1.146449	0.100447	0.010921	0.948
			GRFCSMIC	1.16514	0.108355	1.104869	0.099865	0.011763	0.946
		log(3)	True	-0.01311	0.06088	2.250031	0.059211	0.003706	0.949
			HT	0.836351	0.136982	1	0.131293	0.018849	0.952
			GRN	1.114833	0.133936	1.022745	0.12923	0.018089	0.952
			GRMIS	1.098573	0.134396	1.019243	0.119741	0.018208	0.931
			GRMIC	1.354594	0.135155	1.013522	0.119708	0.018488	0.93
			GRFCSMIS	-0.52327	0.128106	1.069285	0.115569	0.016444	0.928
			GRFCSMIC	-0.46431	0.127535	1.074077	0.115312	0.016291	0.934
	90	log(1.5)	True	0.0138	0.084364	2.251745	0.083155	0.007117	0.947
			HT	1.897751	0.189966	1	0.183082	0.036146	0.94
			GRN	1.897914	0.183304	1.036344	0.176042	0.03366	0.942
			GRMIS	8.193088	0.198884	0.955159	0.163381	0.040658	0.902
			GRMIC	8.29543	0.195141	0.97348	0.162322	0.039211	0.894
			GRFCSMIS	4.745953	0.177903	1.067808	0.159259	0.03202	0.918
			GRFCSMIC	3.798847	0.181029	1.049366	0.157469	0.033009	0.908
		log(3)	True	-0.04654	0.088525	2.348938	0.089229	0.007837	0.95
			HT	0.928622	0.207941	1	0.196985	0.043343	0.939
			GRN	1.061707	0.203441	1.022115	0.192655	0.041524	0.943
			GRMIS	4.095097	0.206598	1.006498	0.181218	0.044707	0.913
			GRMIC	3.94024	0.205562	1.011573	0.180065	0.044129	0.91
			GRFCSMIS	1.614577	0.194645	1.068304	0.175683	0.038201	0.906
			GRFCSMIC	1.290465	0.198216	1.049058	0.174164	0.039491	0.904

Table 3.4: Simulation results for estimating β_x using the data imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 4000$, $n = 800$ comparing simple random sampling (SRS), case-control sampling (CC), and stratified case-control sampling (SCC). The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Design	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	90	log(1.5)	SRS	True	-0.19575	0.056825	2.306052	0.058701	0.00323	0.953
				HT	1.348122	0.131041	1	0.130666	0.017202	0.943
				GRN	0.599619	0.123075	1.064728	0.120298	0.015153	0.942
				GRMIS	1.352229	0.120274	1.089521	0.121238	0.014496	0.942
				GRMIC	1.064436	0.12481	1.049923	0.120657	0.015596	0.938
				GRFCSMIS	0.372602	0.116359	1.126176	0.115015	0.013542	0.938
				GRFCSMIC	0.262224	0.118828	1.102777	0.114345	0.014121	0.936
			CC	True	-0.19575	0.056825	2.307166	0.058701	0.00323	0.953
				HT	1.278795	0.131104	1	0.121309	0.017215	0.938
				GRN	1.295066	0.128054	1.023824	0.120734	0.016425	0.943
				GRMIS	1.925384	0.129153	1.015113	0.122768	0.016741	0.942
				GRMIC	1.665403	0.12981	1.009974	0.123029	0.016896	0.94
				GRFCSMIS	1.221831	0.119855	1.093861	0.11281	0.01439	0.938
				GRFCSMIC	0.804186	0.117846	1.112503	0.112475	0.013898	0.938
			SCC	True	-0.19575	0.056825	1.941845	0.058701	0.00323	0.953
				HT	-0.6459	0.110345	1	0.110845	0.012183	0.957
				GRN	-0.09306	0.109473	1.00797	0.110642	0.011984	0.952
				GRMIS	0.081196	0.110714	0.996668	0.111346	0.012258	0.954
				GRMIC	-0.02695	0.108453	1.017446	0.111431	0.011762	0.954
				GRFCSMIS	-0.16322	0.101909	1.082777	0.105767	0.010386	0.954
				GRFCSMIC	-0.10748	0.100555	1.097364	0.105699	0.010111	0.952
		log(3)	SRS	True	0.1293	0.064842	2.25486	0.06303	0.004206	0.954
				HT	0.974558	0.146209	1	0.140603	0.021492	0.948
				GRN	0.744679	0.129418	1.129747	0.130516	0.016816	0.94
				GRMIS	0.713557	0.131614	1.110893	0.131276	0.017384	0.942
				GRMIC	0.650456	0.131029	1.115852	0.131065	0.01722	0.94
				GRFCSMIS	0.627308	0.127227	1.149195	0.127457	0.016234	0.942
				GRFCSMIC	0.60765	0.128461	1.138158	0.126735	0.016547	0.944
			CC	True	0.1293	0.064842	2.208732	0.06303	0.004206	0.954
				HT	1.422661	0.143218	1	0.130477	0.020756	0.928
				GRN	1.646294	0.141186	1.014393	0.129232	0.020261	0.927
				GRMIS	1.614425	0.1409	1.016452	0.130462	0.020167	0.931
				GRMIC	1.506875	0.139858	1.024024	0.130487	0.019834	0.926
				GRFCSMIS	1.395031	0.13998	1.023132	0.124715	0.019829	0.925
				GRFCSMIC	1.32011	0.137537	1.041307	0.124594	0.019127	0.922
			SCC	True	0.1293	0.064842	1.938671	0.06303	0.004206	0.954
				HT	0.82001	0.125707	1	0.123465	0.015883	0.938
				GRN	0.693561	0.126412	0.99442	0.122793	0.016038	0.94
				GRMIS	0.733702	0.126538	0.99343	0.123577	0.016077	0.94
				GRMIC	0.70857	0.127711	0.984303	0.123601	0.016371	0.936
				GRFCSMIS	0.771774	0.127503	0.985911	0.119766	0.016329	0.944
				GRFCSMIC	0.614896	0.124678	1.008254	0.119554	0.01559	0.946

for $\{N, n\} = \{10000, 2000\}$, keeping all other parameters the same as Tables B.4, B.5, and B.6, are presented in Tables B.10, B.11, and B.12, respectively. The efficiency conclusions were similar to those observed under $\{N, n\} = \{2000, 400\}$, with the larger sample sizes again removing any observed bias.

Table 3.4 presents the relative performance under error scenario 3 for estimating β_X using the data imputation approach comparing simple random sampling to case-control and stratified case-control sampling where $\{N, n\} = \{4000, 800\}$ and censoring was 90%. GRFCSMI had increased efficiency compared to HT and GRN for nearly all designs whereas GRMI did not; however, the absolute gain in efficiency varied by sampling design. The RE for GRFCSMI was higher for SRS than for CC and SCC, ranging from 1.10 to 1.15 for SRS compared to 0.99 to 1.11 for CC and SCC. Although the RE for the proposed estimators was lower for the CC and SCC designs than for SRS, the actual standard errors (ESE and ASE) themselves were lower under these outcome-dependent designs. HT was quite inefficient under SRS, leading to a greater gain in efficiency for GRFCSMI; in contrast, HT under SCC was often nearly as efficient as GRFCSMI under SRS. For instance, the ESE of HT for $\beta_X = \log(3)$ and SCC is 0.126, compared to the ESE of 0.128 for GRFCSMIC for SRS. Similar conclusions were observed for error scenario 2 in Table B.13, with all other parameters the same as Table 3.4, except both GRMI and GRFCSMI had slightly increased efficiency compared to HT and GRN for all designs. The RE for GRMI and GRFCSMI ranged from 1.21 to 1.26 for SRS; for CC and SCC designs, however, the RE ranged from 1.09 to 1.15. The RE for GRFCSMI was higher for SRS than for CC and SCC, ranging from 1.10 to 1.15 for SRS compared to 0.99 to 1.11 for CC and SCC. Thus, we observed less overall efficiency gain in the outcome-dependent sampling designs for the proposed methods but still constructed more efficient estimators generally. Results for the IF imputation approach, keeping all other parameters the same as Table B.13 and Table 3.4, are presented in Tables B.14 and B.15, respectively. The conclusions follow very similarly to those of the data imputation approach.

We considered the relative performance of our proposed methods under error scenario 3 where the misclassification generation process additionally included interaction terms (shown in Table B.16). Results for estimating β_X using the data imputation and IF imputation approaches are shown in Tables B.17 and B.18, respectively, with $\{N, n\} = \{2000, 400\}$ and simple random sampling of the validation subset. While the conclusions regarding the comparisons of GRMI and GRFCSMI to GRN were very similar to previous tables under error scenario 3, the efficiency gains of GRFCSMI were much larger than under the more simple misclassification scenarios. Overall, the RE ranged from 1.03 to 1.34 and the reduction in MSE compared to that of GRN was appreciable across all settings. These results suggest that our methods yield larger efficiency gains with increased nonlinearity. In addition, we observed greater efficiency gains for GRFCSMIC compared to GRFCSMIS,

especially for 75% and 90% censoring where the positive predictive value (PPV) was very low. This high censoring and low PPV setting is common for EHR studies and thus suggests that more complex multiple imputation models to model potential nonlinearity would be helpful. The same set of results for error scenarios 1 and 2, namely with added interaction terms into the error models, were also generated (not presented) and we observed even greater efficiency gains for both GRMI and GRFCSMI with the more complex imputation approaches.

3.7. VCCC Data example

In this section, we applied the proposed raking methods to electronic health records data on 4797 patients from the Vanderbilt Comprehensive Care Clinic (VCCC), a large HIV clinic. Health care providers at the clinic routinely collect and electronically record data on patients, including demographics, laboratory measurements, pharmacy dispensations, opportunistic infections, and vital status. A recent project at the VCCC performed a full chart review for all records to validate important clinical variables, including antiretroviral dispensations and AIDS-defining events (ADEs). Due to the comprehensive chart reviews, two full datasets were available; the first, which we refer to as the unvalidated data, contains the values for all patients prior to chart review and the second, which we refer to as the validated data, contains the true values after chart review. Additional details on the study design and data validation are in Giganti et al. (2020).

In this example, we were interested in estimating the association between the covariates CD4 cell count and age at the time of antiretroviral therapy (ART) and the outcome of time from the start of ART to the first ADE. As is common for studies based on EHR data, the outcome and covariates used in the analysis were derived variables. Specifically, CD4 cell count and age at the time of ART were extracted from tables of laboratory measurements and demographics, respectively, by matching the test date or visit date to the ART start date. In addition, the time from ART start to first ADE is extracted by finding the date of first opportunistic infection and the ART start date and calculating the time elapsed. A comparison of the unvalidated data to the validated data revealed errors in the ART start date in about 41% of subjects, which led to downstream errors in the covariates and outcome of the statistical analysis. In addition, the ADE event was very rare with 93.8% censoring and was subject to appreciable misclassification at 11%, suggesting that raking estimators that ignore the misclassification will be inefficient. The misclassification yielded sensitivity, specificity, positive predictive value, and negative predictive value of 0.879, 0.892, 0.351, and 0.991, respectively. The

exact eligibility criteria used for the analysis and degree of measurement error in the covariates and outcome are given in Appendix B.2.

For this analysis, we considered the validated data to be the "truth" and defined the hazard ratio (HR) estimates calculated using the entire validated dataset to be the true, gold-standard estimates. The naive estimator that calculates the HRs using the entire unvalidated dataset was also considered, along with the HT estimator, the GRN estimator proposed by Oh et al. (2019), and the proposed raking estimators using multiple imputation (GRMI and GRFCSMI) for both the data imputation and IF imputation approaches. Although we had a fully validated dataset, we retrospectively sampled 100 different validation subsets as if we did not have validated data for all records in order to examine the estimators' performance. Due to the rare-event setting, we considered two different validation subset sampling designs: CC and SCC. Two variants of SCC were considered: 1) stratified case-control balanced (SCCB), which is described in Section 3.6.1, and 2) stratified case-control Neyman allocation (SCCN), where the number of subjects sampled in each strata is proportional to the product of the phase one stratum size and the within-stratum (error-prone) influence function standard deviation. In addition, we considered two different validation subset sizes, 340 and 680, representing roughly 21% and 43% of the cohort respectively. For CC, all 248 error-prone cases were selected along with a random sample of 92 (or 432) error-prone controls. For SCCB and SCCN, CD4 count was stratified at cutpoints of 100, 200, and 400 to create four discrete covariate groups for sampling. These cutpoints were selected to strategically oversample more informative subjects. Specifically, given that CD4 count is an important indicator of HIV severity, someone with CD4 count below $200\ \mathrm{cells/mm^3}$ is considered to be at high risk of getting an ADE. Thus, we selected cutpoints at 100 and $200\ \mathrm{cells/mm^3}$ to oversample subjects clinically defined as high risk for an ADE to try to select more true cases and increase efficiency. For each sampling design, the same imputation models (both with and without interaction terms) and influence function working models were fit as described in the simulation section for error scenario 3 with CD4 cell count and age at ART start corresponding to $X^\star$ and Z, respectively.

The median of the 100 HRs and the median of their corresponding 95% confidence interval widths for the proposed methods using the data imputation approach are presented in Table 3.5. For each subset size and sampling design, the naive estimator had significant bias (calculated with respect to the true estimator) for both covariates (31.3% for CD4 and 31.1% for age). In contrast,

Table 3.5: The median hazard ratios (HR) and their corresponding 95% confidence interval widths calculated using the data imputation method from 100 different sampled validation subsets for a 100 cell/mm^3 increase in CD4 count at ART initiation and 10-year increase in age at CD4 count measurement.

Subset size	Sampling	Method	CD4 HR	CD4 CI width	Age HR	Age CI width
340	CC	True	0.693	0.190	0.829	0.361
		Naive	0.910	0.125	1.087	0.275
		HT	0.669	0.313	0.829	0.579
		GRN	0.674	0.274	0.819	0.465
		GRMIS	0.679	0.260	0.824	0.440
		GRMIC	0.678	0.264	0.830	0.438
		GRFCSMIS	0.675	0.265	0.824	0.444
		GRFCSMIC	0.677	0.261	0.824	0.440
	SCCB	True	0.693	0.190	0.829	0.361
		Naive	0.910	0.125	1.087	0.275
		HT	0.686	0.283	0.823	0.573
		GRN	0.687	0.280	0.820	0.494
		GRMIS	0.689	0.272	0.835	0.496
		GRMIC	0.689	0.278	0.826	0.491
		GRFCSMIS	0.687	0.275	0.839	0.498
		GRFCSMIC	0.689	0.276	0.814	0.495
	SCCN	True	0.693	0.190	0.829	0.361
		Naive	0.910	0.125	1.087	0.275
		HT	0.690	0.308	0.779	0.665
		GRN	0.688	0.308	0.807	0.599
		GRMIS	0.684	0.303	0.813	0.608
		GRMIC	0.684	0.299	0.807	0.596
		GRFCSMIS	0.687	0.302	0.818	0.614
		GRFCSMIC	0.690	0.297	0.803	0.598
680	CC	True	0.693	0.190	0.829	0.361
		Naive	0.910	0.125	1.087	0.275
		HT	0.692	0.237	0.826	0.412
		GRN	0.693	0.230	0.825	0.385
		GRMIS	0.693	0.228	0.826	0.380
		GRMIC	0.697	0.228	0.826	0.382
		GRFCSMIS	0.693	0.228	0.826	0.383
		GRFCSMIC	0.696	0.229	0.821	0.382
	SCCB	True	0.693	0.190	0.829	0.361
		Naive	0.910	0.125	1.087	0.275
		HT	0.695	0.234	0.837	0.416
		GRN	0.695	0.233	0.830	0.395
		GRMIS	0.693	0.232	0.829	0.393
		GRMIC	0.697	0.233	0.831	0.393
		GRFCSMIS	0.693	0.231	0.826	0.393
		GRFCSMIC	0.694	0.232	0.832	0.394
	SCCN	True	0.693	0.190	0.829	0.361
		Naive	0.910	0.125	1.087	0.275
		HT	0.690	0.229	0.826	0.430
		GRN	0.689	0.228	0.821	0.406
		GRMIS	0.689	0.226	0.823	0.404
		GRMIC	0.689	0.228	0.825	0.401
		GRFCSMIS	0.689	0.226	0.822	0.403
		GRFCSMIC	0.689	0.228	0.821	0.406

HT and all of the raking estimators yielded nearly unbiased estimates of the true estimates for both covariates. In addition, GRN had narrower 95% confidence interval (CI) widths than that of HT for all sampling designs. For a subset size of 340, GRMI and GRFCSMI both had narrower CI widths than those of GRN for all sampling designs. However, the degree of efficiency gain differed by sampling design; namely, we observed a larger increase in efficiency (around a 5% decrease in CI width) from GRMI and GRFCSMI under CC sampling compared to SCCB or SCCN (at most a 3% decrease in CI width). GRMI and GRFCSMI under CC sampling had the narrowest median CI widths among all estimators for the 340 subset size. When the validation size was 680, the efficiency gain from GRMI and GRFCSMI over GRN was comparable across sampling designs and the median widths of the confidence intervals were similar. The more modest efficiency gains from GRMI and GRFCSMI over GRN compared to those observed in the simulations can likely be attributed to relatively poor imputation models. The small number of cases at phase one and low PPV of the error-prone event indicator made imputation models difficult to build due to the the validation subset containing an extremely small number of true cases. Across the 100 sampled validation subsets, the average ROC-AUC for the imputed event indicator ranged from 0.652 to 0.670 across all sampling designs, suggesting that the imputations of the event indicator were poor. Interestingly, GRMI had comparable, if not narrower, confidence interval widths than GRFCSMI across sampling designs and subset sizes. This is likely due to the fact that the amount of covariate error present was very small, which corresponds to error scenario 2 in the simulations where GRMI and GRFCSMI had comparable efficiency. Table B.19 presents the median HRs and 95% confidence interval widths across the 100 validation subsets for the IF imputation approach. The conclusions about the comparisons of the naive, HT, and GRN estimators are very similar to those of the data imputation approach. For both subset sizes, GRMI and GRFCSMI under CC and SCCB were less efficient than GRN, except for GRMIC under SCCB for the 340 subset size. GRMI and GRFCSMI under SCCN had slightly better performance, with narrower CI widths for the 340 subset size but not the 680 subset size. The lack of efficiency gains observed for the IF imputation approach can be attributed to the very poor influence function imputation working models. Across the 100 sampled validation subsets, the average R-squared for the CD4 influence function working models ranged from 0.099 to 0.194, indicating a lack of predictive accuracy. In small samples, such low correlation between the target and auxiliary variables can limit the improvement over the HT estimator, indicating the need to carefully examine the performance of the imputation working

models, especially under complex error scenarios. In the rare event setting, validation sampling strategies that target missed true cases, such as by stratifying on risk factors that may be less prone to error, will also help efficiency.

3.8. Discussion

The increasing availability of EHR data collected on large patient populations has allowed researchers to study possible associations between a wide array of risk factors and health outcomes rapidly and cost-effectively. However, estimating such associations without bias requires precisely measured data on the variables of interest, an assumption that is often not met with EHR data due to errors in derived variables, error-prone record entry, or other error mechanisms. To address such bias, Oh et al. (2019) proposed validating a subset of records and applying generalized raking estimators, including GRN studied in this manuscript. However, we demonstrated in this manuscript that GRN, which builds the raking variables from the error-prone data, is inefficient in the presence of event indicator misclassification. In addition, we proposed two classes of generalized raking estimators utilizing multiple imputation to estimate the optimal auxiliary variable, one that yields the optimal efficiency. Both MI approaches yield estimates of the expected value of the influence function for the target parameter based on the error-free data. The data imputation estimators impute either the event-indicator or all error-prone variables (if applicable) to construct auxiliary variables with increased degree of linearity with the true population influence functions. The IF imputation estimators take the data imputations and then fit a (potentially flexible, nonlinear) working model of the true population influence functions to construct auxiliary variables. These raking estimators are very appealing for the analysis of EHR data because their validity is not sensitive to the true measurement error structure nor do they require correct specification of the imputation or influence function working models, all of which are generally unknown for such large observational data. These features do, however, affect their efficiency and thus constructing estimators with increased efficiency has been the main focus of this manuscript.

Overall, the proposed raking estimators using multiple imputation performed well, yielding nearly unbiased estimators, the highest RE, and the lowest MSE across all simulation settings. For settings involving misclassification only or misclassification and event-time error, both GRMI and GRFCSMI had large efficiency gains compared to GRN for all parameter settings. For the most complex error setting involving errors in the covariates, event-time, and event indicator, GRFCSMI

had appreciable efficiency gains compared to GRN and GRMI for all parameter settings, which increased when nonlinear error functions were simulated. For all error scenarios, we observed more appreciable efficiency gains under 50% and 75% censoring compared to 90% censoring. It is of note that these simulations involved error settings with very low sensitivity or PPV to mimic real EHR analysis scenarios. In simulations with higher sensitivity or PPV (not presented), larger efficiency gains were realized for GRMI and GRFCSMI, with RE greater than 1.5. The data analysis, which involved an event with over 90% censoring and very low PPV, resulted in similar conclusions. Nevertheless, we observed that GRMI and GRFCSMI yielded around a 5% reduction in CI widths for both covariates, an appreciable gain in a data poor setting. In addition, we considered outcome-dependent sampling designs to select the validation subset to increase efficiency in rare event settings where the number of cases is small. Specifically, we evaluated case-control and stratified case-control sampling designs and found that while the gain in efficiency for GRMI or GRFCSMI over GRN is smaller compared to the efficiency gain under SRS, the overall standard errors are lower, yielding the most efficient estimates across all designs. While good imputation models are difficult to construct in rare events settings, one can still obtain more precise estimates overall by selecting more informative subjects to be validated at phase two.

Another possible estimation approach for the considered settings is the direct multiple imputation estimator, which uses MI to impute the error-prone variables and plug into the Cox model to obtain estimates without the use of raking. Giganti et al. (2020) considered this approach using discrete failure time models but noted challenges with correctly specifying the imputation model. While the MI estimator will be more efficient than raking estimators if the regression and imputation models are correctly specified, Han, Shaw, and Lumley (2019) showed that in the nearly-true model framework of Lumley (2017), even slight misspecification of the models result in bias and worse MSE than raking. This robustness makes raking a very appealing approach in practical settings where the true models are generally unknown.

The two-phase design framework considered in this manuscript is a specific missing data setting where the data are missing by design. This missing data lens allows us to consider the augmented inverse probability weighted (AIPW) estimators proposed by Robins, Rotnitzky, and Zhao (1994), who showed that the class of AIPW estimators contains all regular asymptotically linear estimators consistent for the design-based parameter of interest. There is a close relationship between AIPW

and raking estimators, in that the class of AIPW estimators contains the raking estimators, but the raking estimators include all of the best AIPW estimators (Lumley, Shaw, and Dai, 2011). Thus, raking estimators are asymptotically efficient among design-based estimators and provide simple, easy to compute AIPW estimators. In particular, the raking estimators utilizing multiple imputation proposed in this manuscript yield practical methods to approximate the optimal AIPW estimator in settings involving complex measurement error that is often seen in EHR data. In addition, these estimators are consistent without requiring correct specification of the imputation or working models; however, they yield the most efficient design-based estimator if the models are correctly specified.

In this work we proposed a novel estimation method to improve raking estimators and showed additional efficiency could be gained by pairing these estimators with an efficient two-phase sampling design. While this manuscript considered outcome-dependent sampling designs to improve efficiency in rare-event settings, we believe that more theoretical and empirical work studying efficient sampling designs and their effects on efficiency for failure time outcomes is needed. In particular, constructing multi-phase sampling designs would be a fruitful avenue for future work. See McIsaac and Cook (2015), Chen and Lumley (2020), and Han et al. (2020) for some initial work. These authors considered designs where a pilot sample could initially be selected from the cohort to obtain information on the validated data that can be used to guide follow-up sampling waves. We believe more work is needed to understand how best to take advantage of such strategies for the continuous failure time setting.

CHAPTER 4

CONSIDERATIONS FOR ANALYSIS OF TIME-TO-EVENT OUTCOMES MEASURED WITH ERROR: BIAS AND CORRECTION WITH SIMEX

4.1. Abstract

For time-to-event outcomes, a rich literature exists on the bias introduced by covariate measurement error in regression models, such as the Cox model, and methods of analysis to address this bias. By comparison, less attention has been given to understanding the impact or addressing errors in the failure time outcome. For many diseases, the timing of an event of interest (such as progression-free survival or time to AIDS progression) can be difficult to assess or reliant on self-report and therefore prone to measurement error. For linear models, it is well known that random errors in the outcome variable do not bias regression estimates. With non-linear models, however, even random error or misclassification can introduce bias into estimated parameters. We compare the performance of two common regression models, the Cox and Weibull models, in the setting of measurement error in the failure time outcome. We introduce an extension of the SIMEX method to correct for bias in hazard ratio estimates from the Cox model and discuss other analysis options to address measurement error in the response. A formula to estimate the bias induced into the hazard ratio by classical measurement error in the event time for a log-linear survival model is presented. Detailed numerical studies are presented to examine the performance of the proposed SIMEX method under varying levels and parametric forms of the error in the outcome. We further illustrate the method with observational data on HIV outcomes from the Vanderbilt Comprehensive Care Clinic.

4.2. Introduction

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 some time 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, both of which are subject to error. For example, human biologists and demographers are interested in the variability in the age at menarche (first menstruation) (Marshall and Tanner, 1986). Oftentimes, subjects are enrolled several years after menarche, and so the event-time is based on retrospec-

tive recall and hence subject to error. As Holt, McDonald, and Skinner (1991) observed, studies comparing age at menarche reported retrospectively to those reported in medical records have shown that differences in the two can be attributed to recall error symmetrically distributed around zero. In addition, epidemiological researchers frequently use observational databases, where data accuracy can also be a concern. In observational studies of HIV/AIDS, the event time between antiretroviral therapy (ART) initiation and a disease outcome, such as AIDS-defining illness and associated risk factors, is often of interest. In prior studies using routinely collected health record data, we have observed substantial errors, in as many as 30% of patient records, in the time of ART initiation (Duda et al., 2012; Shepherd and Yu, 2011). Even in studies where a failure time may be measured precisely, such as time to virologic failure defined by an electronically recorded HIV-RNA test exceeding a threshold, the error in the baseline time will create error in the time-to-event outcome. Ignoring these errors can lead to biased estimates of the associations of interest.

There is a rich body of knowledge describing the impact of and methods to correct for covariate measurement error, particularly for time-to-event outcomes (Carroll et al., 2006). For the Cox model, these methods include approximate methods such as regression calibration (Prentice, 1982) and SIMEX (Cook and Stefanski, 1994). They also include methods that have been shown to be unbiased under certain assumptions; including the parametric corrected score (Nakamura, 1992), conditional score (Tsiatis and Davidian, 2001), non-parametric corrected score (Huang and Wang, 2000, 2006) and likelihood methods (Hu, Tsiatis, and Davidian, 1998; Li and Lin, 2000), to name a few.

Much less has been written about the effect of or methods for errors in the failure time outcome itself. For continuous outcomes and linear regression, it is well known that random outcome error does not introduce bias into the regression coefficients. However, for nonlinear models, simple random error in outcomes can bias the coefficients (Carroll et al., 2006). This has been well studied in the case of binary outcomes (Magder and Hughes, 1997; Wang et al., 2016) and discrete failure time data (Meier, Richardson, and Hughes, 2003), where estimates of sensitivity and specificity can be incorporated to adjust estimation for the bias induced by outcome misclassification. Errors in outcome that are correlated with covariates can also be a source of bias in the association between these variables. Some methods, which adjust for covariate-dependent estimates of sensitivity and specificity, have been presented (Edwards et al., 2013; Hunsberger, Albert, and Dodd,

2010; Magaret, 2008).

For uncensored, continuous failure time outcomes, Skinner and Humphreys (1999) found that random multiplicative error has little effect on the acceleration parameter estimated by a Weibull regression model, particularly when there is a relatively small measurement error variance. Korn, Dodd, and Freidlin (2010) noticed that the bias in estimating the hazard ratio is very small with small random multiplicative measurement error in the failure time. Even with larger random error, the bias was small when the hazard ratio was moderate, as commonly seen in clinical trials. However, Hong et al. (2012) noticed in their statistical models for progression-free survival, which involved modeling tumor growth and error-prone detection, that multiplicative error would lead to attenuation of the hazard ratio, with larger measurement error leading to greater attenuation.

The Simulation and Extrapolation method (SIMEX) was developed by Cook and Stefanski (1994) to correct additive measurement error in the covariates. SIMEX has been applied to a wide variety of regression models and is generally implemented as an empirical method (Carroll et al., 2006; Küchenhoff, Mwalili, and Lesaffre, 2006). It has been shown to be a useful tool for estimation in the presence of unbiased covariate measurement error in regression models for time-to-event outcomes, e.g., see Zhang, He, and Li (2014), He, Yi, and Xiong (2007), and Greene and Cai (2004). We extend the SIMEX approach to address random multiplicative error in the event time and study whether this method can be applied to reduce bias in the regression coefficients.

In this manuscript we will present a detailed numerical study of the impact of non-differential outcome measurement error on association analyses of failure time data. We provide an approximate formula to estimate the bias in the association parameters induced by random multiplicative error in the event time and examine performance of our proposed method to correct for the induced bias. In particular, we will consider two popular regression models for survival data, Cox and Weibull regression, and compare the vulnerability of these two regression frameworks to bias from error in the event time. The Weibull model is both an accelerated failure time (AFT) and proportional hazards (PH) model. Thus, within this modeling framework, we will compare the impact of outcome error on estimation of the hazard ratio and acceleration parameters for different measurement error scenarios.

Section 4.3 presents the survival time measurement error framework and develops the extended

SIMEX method. Then Section 4.4 presents numerical studies of the bias on the hazard ratio and the acceleration parameter for different measurement error scenarios and the ability of the adapted SIMEX method to ameliorate this bias. We also discuss estimation options when there is a validation subset available with which to estimate the error structure. In Section 4.5, we apply the SIMEX method in an analysis of HIV outcomes among patients starting ART, where the time-to-event is sometimes recorded incorrectly.

4.3. Survival Time Model

We consider the Cox proportional hazards model and Weibull parametric regression model to study the effects of random error in survival time T. The Cox model is given by

$$\lambda(t) = \lambda_0(t)\text{exp}\left(\mathbf{X}\beta\right)$$

where $\lambda(t)$ is the hazard at time t given the $p \times 1$ covariate vector $\mathbf{X}$, $\lambda_0(t)$ is the baseline hazard, and β is the vector of log hazard ratio parameters. For the Weibull (AFT) model, one has

$$T = \text{exp}\left(\alpha_0 + X\alpha_1 + \sigma\epsilon\right) \tag{4.1}$$

where α_0 and α_1 are regression coefficients, σ is a shape parameter, and ϵ is the error term following an extreme value distribution. The model is also known as a linear transformation model, given by

$$\log(T) = \alpha_0 + X\alpha_1 + \sigma\epsilon.$$

4.3.1. Error Framework

We study the case where there is multiplicative error in the uncensored survival time. Let T' be survival time measured with error such that we observe $T' = T \times \text{exp}(\nu)$, where ν has mean 0, variance σ^2_ν, and is independent of T and X. Then the error prone survival time on the log-scale is given by

$$\log(T') = \alpha_0 + X\alpha_1 + \sigma\epsilon + \nu = \log(T) + \nu. \tag{4.2}$$

The performance of the Cox and Weibull models in the presence of outcome error, namely its ability to capture the true association with X, can be directly compared using the fact that the log hazard ratio from the Cox model can also be estimated from the Weibull model with the following relationship

$$\beta = -\frac{\alpha_1}{\sigma}.$$

We note that with the linear form for $\log(T')$ above, the extra error term ν that is independent of the covariate X will not induce any bias in the acceleration parameter using a typical linear regression model. We also note that the error equation in (2) has the same mathematical form as a log-linear event time model with an added frailty term, ν. Keiding, Andersen, and Klein (1997) considered the AFT model for the setting of heterogeneity due to omitted covariates or frailties and observed that there is bias in the Cox model induced by erroneously ignoring an added frailty term ν, whereas there is no expected bias in the acceleration parameter α_1. These authors also derived an approximate formula for the attenuation factor for the hazard ratio parameter in the Cox model, drawing connections between the log-linear model for the uncensored event time and the theoretical linear regression of $\log(T)$ on X. For further detail, see Keiding, Andersen, and Klein (1997). When adapted to our setting, the bias in $\hat{\beta}_{\text{naive}}$, the estimated hazard ratio from naively applying the Cox model in the presence of the error in (4.2), is given by the approximate attenuation factor

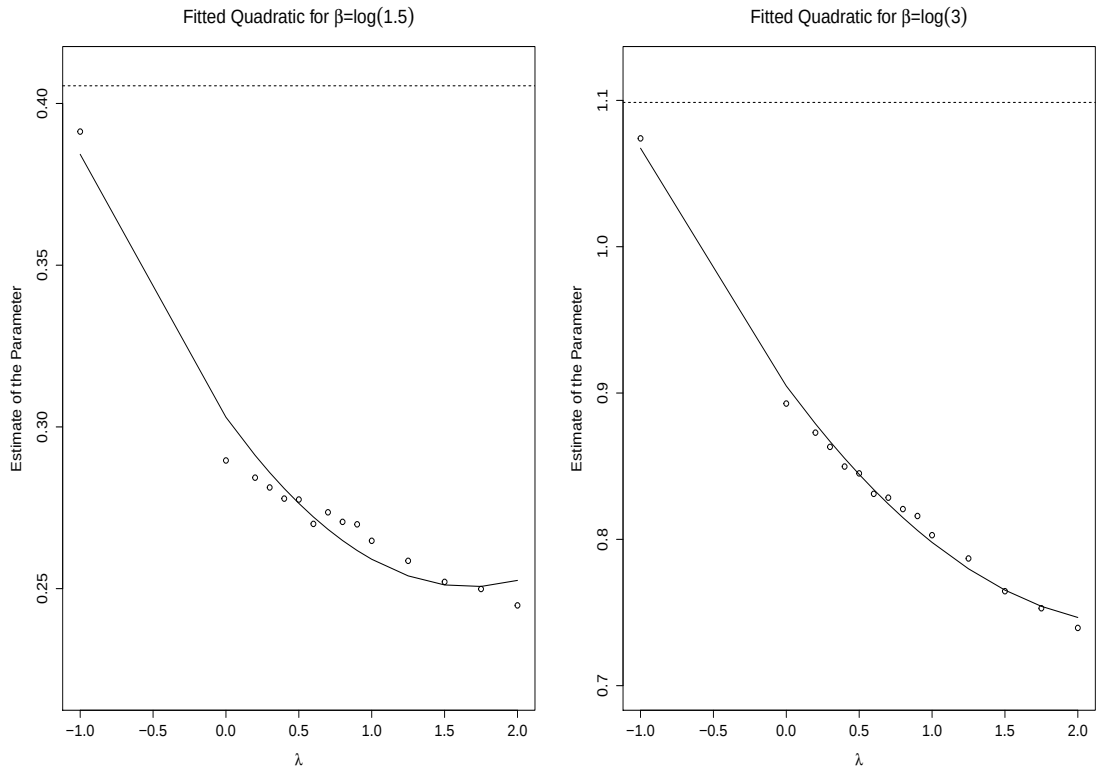
$$\gamma = \frac{1}{\sqrt{1 + \sigma^{-2}\sigma_\nu^2}}. \tag{4.3}$$

That is, $\hat{\beta}_{\text{naive}} \approx \beta \times \gamma$.

4.3.2. SIMEX Method

Given the above framework for the survival model and outcome error, we adapt the SIMEX method to adjust estimation of a regression parameter of interest (e.g. the log hazard ratio β). The SIMEX method was originally developed for additive measurement error in the covariates (Cook and Stefanski, 1994). We adapt the SIMEX method by working with the $\log T$, which converts the assumed multiplicative error to the additive scale.

Figure 4.1: The quadratic approximations of the β parameters as a function of λ, extrapolated to $\lambda = -1$, with the dotted lines denoting the true β for $\beta = \log(1.5)$ (a) and $\beta = \log(3)$ (b)



(a) (b)

We illustrate our method using Figure 4.1. Similar to the original SIMEX method, we estimate the relationship between the size of the measurement error, σ_ν^2, and the bias in the naive estimate of the parameter of interest from an analysis that ignores the error. In the **Simulation** step, we add additional measurement error to each outcome by drawing ω from N$(0, \lambda\sigma_\nu^2)$ and adding the value of this random variable to the already error prone variable $\log T'$ and exponentiating to obtain a new T'_λ. This error addition is repeated B times for a range of values of $\lambda \geq 0$. Then for each iteration of λ and $b = 1, \ldots, B$, we refit the regression model with the new vector of error prone measurement of the survival time $T'_{\lambda b}$ to obtain a new naive log hazard ratio estimate $\beta_{\lambda b}$ (or other parameter of interest, e.g. acceleration parameter from the AFT). The new total measurement error variance in $\log T'_{\lambda b}$ is then given by

$$\sigma_\nu^2 + \lambda\sigma_\nu^2 = (1 + \lambda)\sigma_\nu^2 \tag{4.4}$$

For illustration, we set $B = 1$ and $\lambda \in \{\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}, 1, \frac{5}{4}, \frac{6}{4}, \frac{7}{4}, 2\}$ to estimate new $\beta_{\lambda 1}$, which are shown as small circles in Figure 4.1, and plot these naive $\beta_{\lambda 1}$ versus λ. In the **Extrapolation** step, we then fit a curve to the plot of $\beta_{\lambda 1}$ as a function of the λ's. From this fitted model, we extrapolate back to $\lambda = -1$, which given the new total measurement error variance in Equation (4.4), should approximate the true coefficient value. For the setting with covariate measurement error, Cook and Stefanski recommend a quadratic approximation due to good performance in most cases, but other extrapolation functions such as a linear, loglinear, or nonlinear function are possibilities. We investigated the performance of the quadratic form in our framework using simulations and it outperformed the linear and loglinear approximations. In Figure 4.1, this extrapolation is shown by extending the curve to $\lambda = -1$. Note, in any data application, one could draw a similar figure - increasing the denseness of the vector λ to verify the appropriateness of the chosen extrapolation function (as quadratic or otherwise). This procedure only yields an approximation, since we can only generate curves for which $\lambda \geq 0$, and thus have no estimates on the curve in the region of curve between $[-1, 0)$. To assess the sampling variability of the SIMEX estimates, we utilize the bootstrap to obtain standard errors.

Here, we assume the value of σ_ν^2 is known. In some settings, an estimate of σ_ν^2 may be available from a validation study. In the case that the true value of the measurement error variance, σ_ν^2, is unknown and an estimate is not available, one can apply the method for a variety of possible values for σ_ν^2, and examine the sensitivity of the estimated β. In our data example that follows, we will

illustrate the method with an estimated σ_ν^2 from a validation subsample.

4.4. Simulations and Results

Through simulation, we examined the bias that results from random multiplicative error in the failure time outcome with different distributions of errors and evaluated the estimators from both the Cox and Weibull models. We then applied the proposed SIMEX method to obtain error-corrected estimates of the log-hazard ratio. These values will be compared to those from the true model, a Cox proportional hazards regression model fit with true times T, and a naive Cox model fit with error prone times T'. We present results for varying values of the log hazard ratio β, assumed error distributions, and error variances. From these experiments, we derived means, biases, standard errors (SE), and mean squared errors (MSE). We also compare our observed bias in the hazard ratio with the expected value given by Equation (4.3). We estimate the hazard ratio parameter both parametrically using the Weibull model and semi-parametrically using the Cox model. As we will see from our results below, the multiplicative error introduces no bias in the estimated acceleration parameter, as expected, and so we present results for application of the proposed SIMEX method only for the hazard ratio estimated by the commonly applied Cox model approach.

All simulations were run 2000 times using R version 3.2.1 and assumed that the covariate X followed a standard normal distribution. In addition, we set the true parameters to be $\alpha_0 = 0$ and $\alpha_1 = -\beta$ for $\beta \in \{\log(1.5), \log(3)\}$. The survival outcome T is generated from a Weibull distribution with shape equal to 1 and scale set to $\exp(\alpha_0 + X\alpha_1)$, as defined in Equation (4.1). For the error term ν, as defined in Section 4.3.1, we considered a normal distribution and a shifted gamma distribution with means 0 and variances equal to the varying values of σ_ν^2. These are represented by $\mathrm{N}(0, \sigma_\nu^2)$ and Gamma $\left(\frac{1}{\sigma_\nu^2}, \frac{1}{\sigma_\nu^2}\right) - 1$, respectively, using a parametrization such that if $X \sim$ Gamma(α, β), then $\mathrm{E}(X) = \frac{\alpha}{\beta}$ and $\mathrm{var}(X) = \frac{\alpha}{\beta^2}$. The set of simulations comparing the Cox and Weibull models set the cohort size at $n = 1000$ and varied the measurement error variance to be $\sigma_\nu^2 \in \{0.25, 0.5, 1, 2\}$. We refer to $\sigma_\nu^2 = 0.25$ as small error, $\sigma_\nu^2 = 0.5$ moderate error, $\sigma_\nu^2 = 1$ large error, and $\sigma_\nu^2 = 2$ very large error. For simulations examining the proposed SIMEX method, we set $B = 50$ and $\lambda \in \{0, 0.5, 1, 1.5, 2\}$, following the recommendation by Cook and Stefanski for covariate measurement error. A quadratic function is used in the extrapolation step. These simulations ran 100 bootstrap replications and let the number of subjects be $n \in \{300, 1000\}$, and set the measurement error variance to be $\sigma_\nu^2 \in \{0.25, 0.5, 1\}$ to examine the performance of the SIMEX

method under different amounts of error.

Table C.1 in Appendix C.1 presents summary statistics for $\frac{T'}{T}$ for different values of σ_ν^2 to provide a description of the error in T' as a function of the σ_ν^2 in our simulations. We note that even for our small and normally distributed error, the ratio of $\frac{T'}{T}$ could still be quite appreciable. The error-prone time had an average (SD) multiplicative error factor of $1.15(0.61)$ with an IQR of $(0.71, 1.41)$. For moderate error, the average error (SD) factor was $1.25(0.96)$ with an IQR of $(0.58, 1.59)$. For large error, average (SD) error factor was $1.69(1.95)$ with an IQR $(0.53, 2.08)$. As expected for very large error, the error factor was quite large with an average (SD) of $2.55(5.14)$ and an IQR of $(0.38, 2.52)$. When the error term followed a gamma distribution, we of course observed more extreme skewness in $\frac{T'}{T}$ and much larger standard deviation for this factor compared to those of the normal distribution.

Table 4.1 presents the relative performance of the Cox and Weibull models in the presence of multiplicative error, with ν following a normal distribution. Namely, we present the bias, average model standard error (ASE), empirical standard error (ESE), mean-squared error (MSE), and coverage probabilities for the 95% confidence intervals across the simulations. The ASE is calculated as the mean of the model standard errors and the ESE is calculated as the standard deviation of the parameter estimates. As expected, the estimated acceleration parameter (α_1) using both true time and the error-prone time are extremely close, with small bias in the naive estimate for all settings of the measurement error variance and β. We also notice that for all measurement error variance parameter values, the Weibull and Cox estimates for the hazard ratio parameter are biased and reasonably similar, but the bias from the Weibull estimates is consistently slightly smaller. Overall, the naive intercept, shape, and Cox and Weibull hazard ratio parameters remain largely biased through each value of β and the variance of ν. For moderate error, the percent bias magnitude is greater than 16% and the absolute bias is large for the intercept in the Weibull models. For large to very large error, similar results are observed with percent bias magnitude greater than 30% for estimates of β and large absolute bias for the intercept. In addition, these results in Table 4.1 are consistent with the theoretical amount of attenuation bias in β from Equation (4.3). The expected attenuation for β is approximately 0.816 for $\sigma_\nu^2 = 0.5$, 0.707 for $\sigma_\nu^2 = 1$, and 0.577 for $\sigma_\nu^2 = 2$. For $\sigma_\nu^2 = 0.5$, the observed attenuation was 0.799 and 0.797 for true $\beta = \log(1.5)$ and $\log(3)$, respectively. Similarly for $\sigma_\nu^2 = 1$, the attenuation was 0.689 and 0.687, respectively, and 0.562 and 0.563, respectively, for $\sigma_\nu^2 = 2$.

Table 4.1: The percent (%) bias (absolute bias for intercept α_0), average model standard error (ASE), empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets with $n = 1000$ and a mean zero normal distribution for the error term. For all simulations, the Weibull parameters $\alpha_0 = 0$, $\alpha_1 = -\beta$, and shape equaled 1 (exponential time). Values of β and σ^2_ν are shown below.

			Weibull Model			Cox Model	
β	σ^2_ν		α_0	α_1	shape	β	β
log(1.5)	0.25	Bias	0.056	0.060	-7.200	-11.84	-15.43
		ASE	0.038	0.038	0.020	0.035	0.033
		ESE	0.039	0.040	0.000	0.036	0.035
		MSE	0.069	0.040	0.072	0.060	0.072
		CP	0.676	0.969	0.000	0.708	0.522
	0.5	Bias	0.108	-0.210	-16.04	-20.34	-24.97
		ASE	0.042	0.044	0.016	0.035	0.033
		ESE	0.042	0.044	0.000	0.036	0.033
		MSE	0.116	0.044	0.160	0.090	0.107
		CP	0.275	0.976	0.000	0.351	0.138
	1	Bias	0.199	-0.070	-32.30	-31.26	-36.35
		ASE	0.049	0.053	0.013	0.037	0.033
		ESE	0.049	0.054	0.000	0.038	0.034
		MSE	0.205	0.054	0.323	0.132	0.151
		CP	0.020	0.976	0.000	0.072	0.006
	2	Bias	0.347	0.700	-42.25	-43.33	-48.19
		ASE	0.060	0.067	0.009	0.038	0.032
		ESE	0.060	0.069	0.000	0.040	0.033
		MSE	0.352	0.069	0.423	0.180	0.198
		CP	0.000	0.974	0.000	0.004	0.000
log(3)	0.25	Bias	0.055	0.050	-11.23	-11.86	-13.69
		ASE	0.038	0.038	0.020	0.040	0.041
		ESE	0.037	0.038	0.000	0.042	0.043
		MSE	0.067	0.038	0.112	0.137	0.156
		CP	0.696	0.972	0.000	0.102	0.056
	0.5	Bias	0.106	0.000	-23.89	-20.17	-22.84
		ASE	0.042	0.044	0.016	0.039	0.039
		ESE	0.041	0.045	0.000	0.043	0.043
		MSE	0.114	0.045	0.239	0.226	0.255
		CP	0.278	0.977	0.000	0.000	0.000
	1	Bias	0.198	0.070	-32.36	-31.22	-34.58
		ASE	0.049	0.053	0.013	0.039	0.037
		ESE	0.049	0.055	0.000	0.043	0.042
		MSE	0.204	0.055	0.324	0.346	0.382
		CP	0.020	0.968	0.000	0.000	0.000
	2	Bias	0.350	0.140	-47.17	-43.83	-47.43
		ASE	0.060	0.067	0.009	0.039	0.035
		ESE	0.059	0.068	0.000	0.042	0.039
		MSE	0.355	0.068	0.472	0.483	0.523
		CP	0.000	0.970	0.000	0.000	0.000

Table C.2 in Appendix C.1 presents analogous results to Table 4.1 using the shifted gamma error distribution. Similar results are observed, with the naive intercept, shape, and Cox and Weibull hazard ratio parameters largely biased for each value of σ_ν^2 and β.

Table 4.2 presents the $\%$ bias, coverage probabilities (CP), MSE, ESE, average bootstrap standard errors (ASE) for SIMEX estimates, and average model standard errors (ASE) for naive estimates to compare the performance of our SIMEX method for estimating the hazard ratio with the naive method of ignoring the error, which was simulated as normally distributed. We notice that for small error and both values of n and nonzero β, the bias for the SIMEX method is below 5% with coverage close to 95%. The bias for the naive method $(14-15\%)$ is well over double that of SIMEX, with considerably worse coverage. As σ_ν^2 increases, the results tell a similar story. Overall, the bias for both methods increases, but the bias of the naive method continues to be at least double that of the SIMEX method. The coverage for both methods decreases with σ_ν^2, with naive methods' falling off much more rapidly than that of the SIMEX method. For small error, the SIMEX method performs admirably with small $\%$ bias and near 95% coverage. For moderate error, SIMEX still works reasonably well with under 10% bias and coverage close to 90%. With large amounts of outcome error, SIMEX is noticeably biased, with close to 20% bias, but still outperforms the naive method.

Table 4.2 also presents the type 1 error, coverage probability, MSE, ESE, and ASE measurements for the SIMEX method simulated with the true $\beta = 0$ with a normal error distribution. For all combinations of σ_ν^2 and n, we see that the type 1 error hovers around 0.05.

Table 4.3 presents similar estimates of relative performance of estimating the hazard ratio, comparing our SIMEX method to the naive method when the error in T', ν, follows a gamma distribution. In these scenarios, the average multiplicative error and variance for the error in T' were larger compared to the same scenario for the normally distributed error (Table C.1 in Appendix C.1), and the SIMEX method performed worse overall than with the normally distributed error for a fixed value of σ_ν^2. We notice that for small error and both values of nonzero β and n, the bias for the SIMEX method is 8% or below with coverage close to 95%. The bias for the naive method $(16-18\%)$ is over double that of SIMEX, with considerably worse coverage. As σ_ν^2 increases, the bias for both methods increases, but the bias of the naive method continues to be just under double that of the SIMEX method. As expected, the coverage for both methods decreases with σ_ν^2, with naive methods' de-

Table 4.2: The percent ($\%$) bias, average bootstrap standard error (ASE) for SIMEX, average model standard error (ASE) for naive, empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with exponential time and a mean zero normal distribution for the error term. Type 1 error is shown instead of $\%$ bias for $\beta = 0$.

β	σ_ν^2	n	Method	$\%$ Bias	ASE	ESE	MSE	CP
log(1.5)	0.25	300	SIMEX	-3.234	0.077	0.076	0.077	0.942
			Naive	-15.05	0.062	0.064	0.089	0.810
		1000	SIMEX	-3.432	0.041	0.041	0.043	0.930
			Naive	-15.25	0.033	0.033	0.070	0.535
	0.5	300	SIMEX	-8.815	0.082	0.082	0.089	0.921
			Naive	-23.72	0.060	0.062	0.114	0.633
		1000	SIMEX	-9.232	0.044	0.044	0.058	0.852
			Naive	-25.05	0.033	0.032	0.107	0.128
	1	300	SIMEX	-18.46	0.087	0.086	0.114	0.852
			Naive	-36.08	0.060	0.063	0.159	0.33
		1000	SIMEX	-18.90	0.046	0.047	0.090	0.615
			Naive	-36.64	0.033	0.035	0.153	0.007
log(3)	0.25	300	SIMEX	-2.100	0.101	0.099	0.101	0.945
			Naive	-13.51	0.075	0.079	0.168	0.466
		1000	SIMEX	-2.172	0.054	0.053	0.058	0.922
			Naive	-13.98	0.041	0.043	0.159	0.049
	0.5	300	SIMEX	-6.780	0.107	0.107	0.130	0.889
			Naive	-22.07	0.072	0.076	0.254	0.112
		1000	SIMEX	-7.037	0.058	0.058	0.097	0.72
			Naive	-22.94	0.039	0.043	0.256	0.000
	1	300	SIMEX	-16.03	0.111	0.112	0.209	0.62
			Naive	-34.03	0.069	0.074	0.381	0.002
		1000	SIMEX	-16.40	0.060	0.061	0.190	0.16
			Naive	-34.53	0.037	0.041	0.382	0.000

β	σ_ν^2	n	Method	Type 1 Error	ASE	ESE	MSE
0	0.25	300	SIMEX	0.052	0.071	0.071	0.071
		1000	SIMEX	0.043	0.038	0.038	0.038
	0.5	300	SIMEX	0.053	0.076	0.076	0.076
		1000	SIMEX	0.045	0.040	0.040	0.040
	1	300	SIMEX	0.056	0.081	0.081	0.081
		1000	SIMEX	0.047	0.043	0.043	0.043

Table 4.3: The percent ($\%$) bias, average bootstrap standard error (ASE) for SIMEX, average model standard error (ASE) for naive, empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with exponential time and a shifted gamma distribution (mean 0) for the error term. Type 1 error is shown instead of $\%$ bias for $\beta = 0$.

β	σ_ν^2	n	Method	$\%$ Bias	ASE	ESE	MSE	CP
log(1.5)	0.25	300	SIMEX	-7.270	0.077	0.076	0.082	0.929
			Naive	-17.65	0.061	0.064	0.096	0.766
		1000	SIMEX	-8.025	0.041	0.041	0.053	0.864
			Naive	-18.33	0.033	0.034	0.082	0.379
	0.5	300	SIMEX	-16.98	0.081	0.082	0.107	0.841
			Naive	-30.01	0.060	0.063	0.137	0.483
		1000	SIMEX	-17.87	0.043	0.043	0.084	0.594
			Naive	-29.74	0.033	0.035	0.126	0.055
	1	300	SIMEX	-27.54	0.084	0.083	0.139	0.706
			Naive	-41.07	0.060	0.062	0.178	0.207
		1000	SIMEX	-28.37	0.044	0.045	0.124	0.278
			Naive	-42.45	0.032	0.034	0.176	0.003
log(3)	0.25	300	SIMEX	-5.578	0.105	0.106	0.122	0.898
			Naive	-15.94	0.074	0.084	0.194	0.360
		1000	SIMEX	-6.380	0.057	0.059	0.091	0.758
			Naive	-16.55	0.040	0.047	0.188	0.017
	0.5	300	SIMEX	-15.96	0.113	0.122	0.214	0.636
			Naive	-27.45	0.070	0.089	0.314	0.042
		1000	SIMEX	-17.19	0.062	0.066	0.200	0.146
			Naive	-28.71	0.038	0.049	0.319	0.000
	1	300	SIMEX	-29.28	0.117	0.124	0.345	0.232
			Naive	-41.6	0.066	0.095	0.467	0.000
		1000	SIMEX	-30.91	0.065	0.068	0.346	0.001
			Naive	-42.43	0.035	0.049	0.469	0.000

β	σ_ν^2	n	Method	Type 1 Error	ASE	ESE	MSE
0	0.25	300	SIMEX	0.049	0.071	0.069	0.069
		1000	SIMEX	0.057	0.037	0.037	0.037
	0.5	300	SIMEX	0.052	0.074	0.074	0.074
		1000	SIMEX	0.049	0.039	0.039	0.039
	1	300	SIMEX	0.037	0.077	0.075	0.075
		1000	SIMEX	0.056	0.040	0.041	0.041

creasing much more rapidly than that of the SIMEX method. Even with dramatically skewed error, for small outcome error, the SIMEX method performs well with reasonably small $\%$ bias and decent coverage $(76 - 93\%)$. For moderate error, SIMEX performs less well with bias around $16\% - 17\%$ and relatively weak coverage with the errors from this skewed distribution becoming quite large. Our most extreme skewed error setting led to upwards of 15-fold multiplicative error factors and SIMEX performing poorly with bias around 30%. In a real data setting, this magnitude of error may actually be detected and corrected by usual out of range data quality assurance methods at the data collection level.

For the proposed SIMEX method, Table 4.3 also presents analogous type 1 error results to Table 4.2 using a gamma error distribution. Even for this skewed error, the type 1 error rate was preserved at 0.05 for all σ_ν^2 and n.

Table C.3 in Appendix C.1 presents the performance for the normal error distribution, but with time following a log-normal distribution. In this case, the Weibull model is no longer the correct one even for the true event time. We let ν have variance 1 and let $\beta = \log(3)$. Even for this relatively large β and mis-specified parametric model, we see that the naive acceleration parameter is quite unbiased, as before. However in contrast to the previous results, the Cox and Weibull hazard ratios are quite similar and nearly unbiased, although the bias for the Weibull hazard ratio remains slightly smaller. The naive intercept for the Weibull model continues to display large bias with time distributed log-normally. The lack of bias in the hazard ratio can be attributed to the fact that the naive shape parameter in this model, σ, is estimated without bias, in contrast to all three previous models. These results emphasize that the hazard ratio parameter has unpredictable bias depending on the underlying distribution, but that the acceleration parameter is more robustly estimated in the presence of random multiplicative error in the outcome.

4.4.1. Censoring

The above simulations were done with no censoring of the event time. To examine the impact of censoring, simulations were run in similar parameter settings as described in the beginning of Section 4.4 with the addition that the true survival times were randomly censored and error was added to the censored event times. Specifically, we considered $\beta \in \{\log(1.5), \log(3)\}$, $\sigma_\nu^2 \in \{0.5, 1\}$, and simulated 25%, 50%, 75%, and 90% censoring. True survival times were again generated exponentially, but with the baseline hazard set to 0.1. After the true survival times were generated,

separate random right censoring times were determined for each β to yield the desired $\%$ censored event times. The censoring times were generated uniformly with lengths 4, 4, 2, and 1 for each $\%$ censored time, respectively, to mimic trials of different lengths. The error-prone times were then generated by adding random multiplicative error to the censored times and the rest of the simulation parameters follow as before. We note that this kind of error, ie. error in the censored event time, is consistent with a time-to-event outcome in the HIV/AIDS setting discussed in the introduction, where there may be error only in the start time of the observation period for an event (e.g. time of ART initiation) but the event time (e.g. virologic failure) is determined precisely. As Table 4.4 shows, although the SIMEX method does not handle censoring directly, applying our extension to this setting still improves the bias compared to ignoring the error. Of equal interest is that the amount of censoring seems to have an inverse relationship with the percent bias in the log HR. For example with $\beta = \log(1.5)$ and $\sigma_\nu^2 = 0.5$, as the percent censored increases from 25 to 50 to 75 to 90, the percent bias decreases from -12.72 to -7.090 to -3.130 and to -1.690, respectively. In addition, we observe that the CP increases with increasing censoring. A similar effect is observed for other combinations of β and σ_ν^2 in Table 4.4. Thus our results suggest that for rare events that are exponentially distributed and randomly censored, the effect of random, multiplicative measurement error in the censored failure time outcome has little effect on the estimates of β. In this case, the risk sets in the Cox partial likelihood score at each failure time remained largely the same. In such scenarios, since the event indicator is correct in this setting, the Cox score defined by the error prone event times would be a sum of similar score contributions over the same individuals as the score defined by the true event times, hence why there is little bias. This seems to reconcile the different conclusions that Korn, Dodd, and Freidlin (2010) and Hong et al. (2012) came to regarding this setting described in Section 4.2. Korn et al. considered simulations that approximated data involving outcome error in the evaluation of progression-free survival for breast cancer patients. These authors observed that with the correct hazard ratio and 50% censoring, there is very little bias caused by the random measurement error. In addition, simulating a cancer trial with a very rare event, with 96% censoring, resulted in even less bias. On the other hand, the more appreciable effects of event time error on the estimated hazard ratio that Hong et al. found may be attributed to the studied setting being one where there was a much smaller percentage of censored events (5-25% censoring in simulated tumor progression).

Lack of bias in the naive estimates may not always be the case for rare events. For instance,

Table 4.4: The percent ($\%$) bias, mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with $n = 1000$, exponential time, baseline hazard of 0.1, a normal distribution for the error term, and 25%, 50%, 75%, and 90% censoring for the true event time.

$\%$ Censored	β	σ_ν^2	Method	$\%$ Bias	MSE	CP
25	log(1.5)	0.5	SIMEX	-4.776	0.048	0.931
			Naive	-12.72	0.064	0.712
		1	SIMEX	-9.331	0.059	0.872
			Naive	-18.31	0.084	0.492
	log(3)	0.5	SIMEX	-4.948	0.081	0.836
			Naive	-15.53	0.177	0.054
		1	SIMEX	-10.94	0.135	0.501
			Naive	-23.23	0.260	0.000
50	log(1.5)	0.5	SIMEX	-2.828	0.051	0.939
			Naive	-7.090	0.054	0.902
		1	SIMEX	-5.325	0.056	0.930
			Naive	-10.41	0.063	0.839
	log(3)	0.5	SIMEX	-3.441	0.075	0.904
			Naive	-10.34	0.126	0.408
		1	SIMEX	-7.287	0.105	0.762
			Naive	-15.47	0.179	0.120
75	log(1.5)	0.5	SIMEX	-1.430	0.066	0.948
			Naive	-3.130	0.065	0.939
		1	SIMEX	-2.496	0.068	0.946
			Naive	-4.590	0.067	0.940
	log(3)	0.5	SIMEX	-1.885	0.082	0.947
			Naive	-5.610	0.095	0.848
		1	SIMEX	-3.940	0.093	0.909
			Naive	-8.260	0.117	0.714
90	log(1.5)	0.5	SIMEX	-1.233	0.084	0.956
			Naive	-1.690	0.082	0.946
		1	SIMEX	-1.804	0.085	0.960
			Naive	-2.680	0.084	0.941
	log(3)	0.5	SIMEX	-1.281	0.095	0.950
			Naive	-3.640	0.096	0.918
		1	SIMEX	-2.587	0.101	0.942
			Naive	-5.350	0.106	0.874

there can be appreciable bias in the naive estimate for the rare event setting when the majority of the censoring happens early in the observation time period, creating observed event times that are close together relative to the size of the measurement error. We simulated this scenario by censoring an exponential event time T, with baseline hazard 0.1 and a log hazard coefficient of 1 for a standard normal covariate X, on the interval $(0, 0.15)$ and generating the error-prone T' by adding a random, standard normal error term to the censored event time. In this case, there was approximately 90$\%$ censoring and -13.55$\%$ bias in the naive estimate as Table C.4 in Appendix C.1 shows. SIMEX in this case provides a modest bias reduction, with a bias of -11.6$\%$. We simulated a second scenario with appreciable bias for rare events, which included multiplicative random error and covariate-dependent censoring. For this setting, we simulated the underlying proportional hazards model, only we censored uniformly on (0,0.15) if $X > 0$, and otherwise uniformly on $(0, .05)$. Random log-normal multiplicative error with variance 0.5 was added to the censored event time. In this example, there was approximately 90$\%$ censoring and the naive estimate had a bias of -14.70$\%$, while the SIMEX estimate had a bias of -7.91$\%$ as Table C.5 in Appendix C.1 shows.

4.5. Data Example

For the purposes of illustration, we apply the proposed method 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, including demographics, laboratory measurements, and pharmacy dispensations. In addition, the VCCC has fully validated all key research variables, which revealed extensive errors in the original data. Thus, this observational cohort is ideal for directly assessing the relative performance of the SIMEX estimators compared to naive estimators. Throughout this example, we considered the estimates from the fully validated dataset to be the "truth". For a more detailed description of the cohort, see Lemly et al. (2009).

We analyzed data on 3996 HIV-positive patients who established care at the VCCC between 1998 and 2013. The event time here is considered to be the time from the start of antiretroviral therapy (ART) to the time at virologic failure, which is defined as an HIV-RNA count greater than or equal to 400 copies/mL. The HIV-RNA assay, and hence time at virologic failure is considered to be free of errors, whereas the time at the start of ART is error-prone. We studied the association between CD4 at enrollment (i.e. at first visit to the VCCC), patient sex, age at enrollment and the defined

event time. For each analysis (using validated or unvalidated data), patients were excluded if they had a missing ART start date, did not start ART after enrollment, or had no follow-up after starting ART. In the unvalidated dataset, 3049 patients satisfied the criteria for inclusion whereas 2973 patients satisfied the criteria in the validated dataset. A total of 2923 met the inclusion criteria for the analysis of both the validated and unvalidated datasets and were used in all further analyses to ensure that any differences between estimators are not due to differences in included patients. In this dataset, 28.6% of event times had an error with an average (SD) multiplicative error factor of $2.33(32.06)$. Of the 2923 subjects, 22 did not reach failure in the unvalidated but did reach failure in the validated and 54 failed in the unvalidated but not in the validated. Thus, the number of subjects who, due to the error-prone ART start time, had an incorrect event indicator was small at 3%. While SIMEX does not directly address this kind of inclusion/exclusion error, we were interested in seeing how SIMEX would perform in this real data scenario. Censoring was 23.4% in the validated data and 22.3% in the unvalidated data.

We utilize the method described in Section 4.3 and compare the performance of the SIMEX estimator to that of the naive estimator that ignores the error. The univariate and multivariate Cox models were fit and used to calculate the hazard ratios (HR) for a 100-unit increase in CD4, comparing females to males, and a 10-year increase in age. This was done for both the validated and unvalidated datasets. For our SIMEX approach, we set $B = 50$ and $\lambda = \{0, 0.5, 1, 1.5, 2\}$, as described in Section 4.3.2. Here, the variance of the error in time is not assumed to be known, but rather estimated from a validation subset. From the 2923 subjects in both datasets, a random subsample of 300 was assumed available and σ_ν^2 was estimated. As Table C.6 in Appendix C.1 shows, the amount of error in $T'_{\lambda b}$ compared to T, $\frac{T'_{\lambda b}}{T}$, is substantial, as the IQR increases with λ and the standard deviation stays relatively large. Using a quadratic function, we then extrapolate back to $\lambda = -1$ and obtain our approximation of the true HR using the full cohort. Standard errors for the SIMEX method were then obtained using a bootstrap method, with bootstrap sampling stratified on the validation subset membership and using 100 bootstrap samples.

The HR's and their corresponding confidence intervals are shown in Table 4.5 comparing the true, naive, and SIMEX estimators. The true estimator was calculated using the validated dataset, whereas the naive estimator was calculated using the unvalidated dataset to simulate a scenario in which validated data are not available on any subjects. The SIMEX estimator was also calculated

Table 4.5: The hazard ratios (HR) and their corresponding bootstrap 95% confidence intervals for sex, a 100-unit increase in enrollment CD4, and a 10 year increase in age at enrollment for the time at virologic failure post ART.

	Univariate		
	Sex	100 $\times$ CD4	10 $\times$ Age at Enrollment
True	1.128 (1.024,1.243)	0.883 (0.867,0.900)	0.995 (0.955,1.037)
Naive	1.098 (0.997,1.208)	0.885 (0.869,0.901)	0.990 (0.951,1.032)
SIMEX	1.101 (0.988,1.227)	0.882 (0.863,0.902)	0.985 (0.948,1.024)

	Multivariate		
	Sex	100 $\times$ CD4	10 $\times$ Age at Enrollment
True	1.047 (0.950,1.155)	0.883 (0.867,0.900)	0.975 (0.935,1.017)
Naive	1.020 (0.926,1.123)	0.885 (0.869,0.901)	0.973 (0.933,1.014)
SIMEX	1.025 (0.929,1.131)	0.882 (0.862,0.902)	0.967 (0.926,1.010)

on the unvalidated data, assuming a subset of 300 validated subjects was available to estimate the error variance. For the univariate analyses, the SIMEX estimator appears to slightly improve the bias in the HR compared to the naive estimators for patient sex (-2.39% and -2.66%, respectively) and a 100-unit increase in enrollment CD4 (-0.11% and 0.23%, respectively). However, for a 10 year increase in age at enrollment, the SIMEX estimator does not improve the bias compared to the naive estimator with -1.00% and -0.50% bias, respectively. Overall, there is very little bias in the naive analyses of the unvalidated data. We observe similar results in the multivariate analysis. The SIMEX estimator again appears to slightly improve the bias in the HR compared to the naive estimator for patient sex (-2.10% and -2.58%, respectively) and a 100-unit increase in enrollment CD4 (-0.11% and 0.23%, respectively), but not for a 10 year increase in age at enrollment (-0.82% and -0.20%, respectively). Overall both the SIMEX and naive methods are quite close to the HR from the fully validated data. The performance of our SIMEX extrapolation is presented graphically in Figure C.1 in Appendix C.1. Note that our SIMEX approach assumes random measurement error in the time-to-event outcomes. To test whether this assumption holds for the VCCC data, logistic regression models were run on the full data (N=2923) to estimate the odds of the unvalidated censored event time being incorrect for the covariates sex (OR=0.971, p-value=0.756), enrollment CD4 (OR=0.880, p-value $<$ 0.001 for a 100 unit increase), and enrollment age(OR=0.995, p-value=0.241 for a 10-year increase). Given the odds ratio and significant p-value for CD4, it appears the measurement error in the outcome is not purely random. However, even under some modest violations of the random error assumption in this data example, the method still performed relatively well.

In addition, a similar analysis was run with the event time defined to be the time from the start

of ART to the time of first opportunistic infection (OI). Here, the time at first OI is also error-prone, resulting in 45.0% error in the event time and an average (SD) multiplicative error factor of $1.84(18.8)$. Censoring for this endpoint is 79.4% in the validated data and 69.2% in the unvalidated data. We observe similar results in this scenario - the bias in the naive HRs is small to moderate and the SIMEX estimators stay close to those values. More detailed results are shown in Table C.7 in Appendix C.1. We note that although the average error factor is less than that of the analysis for time to virologic failure, the time at first OI has much greater censoring. This may have contributed to the SIMEX method's success in estimating HRs very close to those of the validated estimates.

The run time for the mulivariate analysis described above was 1.06 hours on a 64-bit PC with an i7 processor. We also provide code and a simulated dataset in the Supplemental files to further demonstrate the ease of application of our SIMEX method.

4.5.1. VCCC Data Simulation

We consider a simulation study that mirrors attributes of the VCCC dataset, excluding any covariate-dependent measurement error, to explore the performance of SIMEX under random multiplicative error in this setting. In particular, the true variance of the error in time is estimated from a random subsample of 300 from the VCCC data similar to above. The true beta parameters were obtained by fitting a parametric Weibull regression model to the fully validated data for time to virologic failure, with CD4 count, patient sex, and age at enrollment as covariates. The sample size for the simulated cohort was set at $n = 2923$ and random right censoring times generated uniformly to average 24% censoring to match those settings of the VCCC data. In addition, all three covariates were generated to be as similar as possible to those observed in the VCCC data. Patient sex was randomly sampled according to the true observed probabilities of males and females. Then stratified on patient sex, we generated bivariate normal distributions for the age at enrollment and the square root of CD4 count, where the square root transformation was applied for normality; the means, SDs, and correlations were matched based on the true covariates. True survival times (T) were generated exponentially using the simulated CD4 count, sex, and age at enrollment variables and censoring applied as described above. For the error-prone T', we applied random multiplicative error to T and matched the measurement error distribution for ν to that of the VCCC data. Specifically, 71.5% of subjects in the VCCC had no error in the time-to-event and the remaining subjects had highly right-skewed error. Thus, a shifted gamma error with shape and rate both equal to the estimated measurement error variance was applied to 28.5% of the simulated subjects. We additionally var-

ied the distribution of ν to be normal and shifted gamma (applied to all simulated subjects) to test the sensitivity of the results to the shape of the error distribution. Similar to the VCCC data analysis, the SIMEX method estimated the measurement error variance using a validation subsample of 300 simulated subjects and obtained standard errors using 100 bootstrap samples stratified on validation subset membership.

Table C.8 in Appendix C.1 presents the $\%$ bias, CP, and SEs for the simulation described above using the gamma mixture, mean zero normal, or shifted gamma error distributions. First, we note that with the assumed random measurement error, there was a modest amount of bias in the naive estimates (10-17$\%$ depending on the coefficient and the error distribution) compared to what we observed in the true data (under 3$\%$). We then observe for our simulation that for all covariates, the SIMEX estimates had at least an 80$\%$ reduction in bias over the naive for the gamma mixture, a 60$\%$ reduction for normal error, and at least a 40$\%$ reduction for the shifted gamma. In addition, the CPs for all covariates are similar or a little higher for SIMEX compared to the naive method for all error distributions and the MSEs are generally similar for the SIMEX and naive methods. Thus, these simulations suggest that the non-random measurement error that seems to be present in the VCCC data counteracted bias that would have been observed in the naive estimates with random, multiplicative error. The random error in our simulation for this setting induces bias in the naive estimates that our SIMEX method is able to correct for a variety of underlying error distributions.

4.6. Discussion

There is no substitute for carefully and accurately collected data. In the event that an error-free outcome cannot be obtained, then ideally one would be able to do a detailed validation study to obtain data on the structure of the outcome measurement error in a subset so that proper statistical models could be formulated to estimate and adjust estimators for this error structure. Without the availability of a validation subset, it is common practice to simply ignore the errors in the outcome and proceed with the same analysis as if there were no measurement error. In this work, we saw that even simple random error in a survival outcome can bias the hazard ratio estimator for continuous time-to-event outcomes. We propose a few analysis options for this setting involving random multiplicative error.

Regression theory and our simulations demonstrate that the log hazard ratio from the Cox model can be quite biased even for relatively small amounts of random measurement error; whereas,

the acceleration parameter of the Weibull model remains unbiased in the presence of random multiplicative measurement error. This is notable, since the addition of the error for the studied settings meant the parametric form for the survival outcome assumed by the Weibull model no longer held for the error prone covariate. The observation also held true in our simulations when the parametric form of the true event time was not Weibull. Given this robustness, and the fact that the AFT model has been advocated as a more intuitive model for treatment effects in clinical settings (Swindell, 2009; Wei, 1992), we recommend consideration of this regression model in place of, or at least performed alongside of, the Cox model, in settings where the censored event time is known to have random error. Due to the Weibull model being both PH and AFT, this is more a change in which summary statistic is chosen for the association between a covariate and outcome, than in the model for how X affects the outcome. Keiding, Andersen, and Klein (1997) and others (Hougaard, Myglegaard, and Borch-Johnsen, 1994) made a similar recommendation for AFT models due to their ability to separate out dispersion from regression parameters.

In addition, we described an extension of the SIMEX algorithm to correct the bias in the hazard ratio induced by non-differential multiplicative outcome error for a continuous event time. Although the proposed method is only an approximate method, with some expected bias, simulations demonstrated that our method corrected multiplicative outcome error and performed much better than the common naive method of ignoring the error, maintaining a smaller MSE in a variety of settings. The method does start to break down with large error variance resulting in bias greater than 15%; however, as Table C.1 in Appendix C.1 illustrates, the amount of multiplicative error that induced large biases was extreme, with the inter-quartile range of the ratio of the error prone to the true outcome ranging from less than half to more than double, and would require analyses beyond that of our approximate method. We also applied the proposed method to a data example where there were both censoring and known associations between the outcome error distribution and important predictors. In this example, SIMEX performed similar to the naive method, and in some cases perhaps made mild improvements. These findings under a non-random error scenario are similar to those of Küchenhoff, Mwalili, and Lesaffre (2006), who studied the use SIMEX for misclassification error in binary response variables and in one simulation assessed effects of differential measurement error. They found that the naive estimates are biased, but in different directions (away from or towards the true parameter). Our results suggest, that like with random error, the hazard may be less subject to bias with moderate systematic error in the outcome when the observed event

time is rare. It is of note that the above simulations and data example all involved right-skewed error, with the mean error-prone T' larger than that of the true T, due to the nature of time-to-event data. We investigated the performance of our method with left-skewed error for a small number of settings and found that SIMEX overestimated the true hazard ratio while still providing similar reduction in the magnitude of the bias for all settings. Detailed results can be found in Table C.9 in Appendix C.1. Thus, our simulations showed that the SIMEX method under random error was an improvement over the naive estimator for a variety of underlying error distributions.

Because the error model under study was observed to cause appreciable bias in the estimated hazard ratio, but imperceptible bias in the acceleration parameter, we applied our method only to the Cox model. However, it could be similarly applied to other regression models. Limitations of our findings include the need to have a known error variance σ_ν^2 or the availability of a validation subset from which to obtain an estimate of this error variance. With a validation subset, our method can incorporate an estimated error variance and the bootstrap can be used to obtain standard error estimates that incorporate the additional uncertainty from the estimated nuisance parameter. It is also noteworthy that all of our simulations were done with independent error terms. Our approximate method is likely not able to easily handle covariate-dependent or differential error in the general setting and correction methods for this error structure is an area for future work. It is also of interest to extend this method to be able to handle data with both outcome and covariate measurement error.

In the setting of random error, SIMEX provides a practical estimation method to adjust the hazard ratio for bias induced by non-differential measurement error in the failure time outcome. Without the availability of a validation subset or known variance for the outcome error, our method provides analysts with a new tool to perform sensitivity analyses that vary assumptions about the underlying measurement error variance, and examine the robustness of results to random error in the event time.

CHAPTER 5

DISCUSSION

Data collected primarily for non-research purposes, such as those from EHR, can be a double-edged sword: they provide novel opportunities for medical discovery but have been observed to be error-prone. As EHR data is increasingly being used as a primary source of data for medical studies, it is imperative that these errors are adjusted for to ensure that associations between risk factors of interest and diseases are estimated without bias. We illustrated the degree of potential bias using EHR data from the VCCC HIV cohort, which contained both a fully validated, "true" dataset as well as a fully unvalidated, error-prone dataset. Using the unvalidated dataset, the estimate of the CD4 hazard ratio from a Cox model was underestimated by 3-fold compared to the corresponding estimate from the fully validated dataset. In addition, the age hazard ratio was overestimated in the wrong direction using the unvalidated dataset such that the null hypothesis of a unit hazard ratio was nearly rejected. Spurious associations driven by such unvalidated outcomes and exposures can misdirect researchers and potentially be harmful to patients down the line. The existing literature does not adequately address the types of complex measurement error observed in EHR data; in particular, errors in the censored time-to-event outcome. In this dissertation, we proposed several methods to address the bias resulting from the correlated failure-time outcome and covariate error often seen in EHR data.

In Chapter 2, we developed four different estimators that combine a validation subset with the full error-prone cohort data to try to obtain unbiased and efficient estimates. The RC and RSRC estimators estimate the true failure-time outcome and/or covariate given all of the unvalidated data and information on the error structure from the validation subset. In settings with moderate true hazard ratios and rare events, the RC and RSRC estimators had the lowest relative MSE; however, they are biased and in some settings, had appreciable bias. The generalized raking estimators GRN and GRRC, in contrast, are consistent whenever the HT estimators yield consistent estimates. In addition, this property is not affected by the true measurement error structure, whereas the RC and RSRC estimators can perform poorly when the error structure is not correctly specified. Overall, the raking estimators were nearly unbiased for all error settings, had lower standard errors than those of the HT estimator, and had the lowest relative MSE for most error settings.

In Chapter 3, we developed generalized raking estimators that improved the GRN estimator from Chapter 2 in the presence of event-indicator misclassification. We demonstrated that the misclassification results in a poor linear association between the variables of interest and the auxiliary variables used for the GRN estimator. Thus, we developed two classes of raking estimators that utilize multiple imputation to impute the true data given the observed error-prone data. The data imputation estimators impute either the event-indicator or all error-prone variables (if applicable) to construct auxiliary variables with increased degree of linearity with the true population influence functions. The model-calibration estimators take the data imputations and then predicts the true population influence functions to construct auxiliary variables. Overall, GRMI and GRFCSMI performed well, yielding nearly unbiased estimators and the lowest MSE across all simulation settings. For error settings involving just misclassification or misclassification and event-time error, both GRMI and GRFCSMI had large efficiency gains compared to GRN. In the most challenging setting involving correlated errors in the event-time and covariate as well as misclassification, GRFCSMI had appreciable efficiency gains compared to GRN and GRMI. Furthermore, we demonstrated that in rare-event settings, the raking estimators can gain additional efficiency by selecting the validation subset using an outcome-dependent sampling design, although the gain in efficiency over GRN was not as appreciable.

In Chapter 4, we studied the effects of random multiplicative error in the failure-time outcome through the lens of the Cox model and Weibull AFT model. We noticed that even for small amounts of error, the estimated hazard ratios from the Cox model were quite biased; whereas, the acceleration parameter from the Weibull model was unbiased. This robustness suggests the Weibull model to be a possible alternative to the Cox model in settings involving error in the censored event time. In addition, we developed an extension of the SIMEX method to correct the bias in the Cox model hazard ratio estimates induced by the multiplicative outcome error. Although the method is only an approximate method with some bias, we reduced bias significantly compared to the naive method of ignoring the error and maintained a lower MSE for many settings.

This dissertation motivates several future areas of research. In Chapter 3, we considered some outcome-dependent sampling designs (case-control and stratified case-control) to improve the efficiency of the generalized raking estimators in rare-event settings. While these offered a clear advantage over simple random sampling in simulations, more theoretical work would be desirable

to better understand the settings and conditions under which these outcome-dependent designs improve efficiency. In addition, it would be useful to develop novel optimal sampling designs for the complex error settings considered in this dissertation. These optimal sampling designs, however, often depend on unknown parameters, such as the true error variance of a variable, and are tailored to a particular analysis as opposed to being broadly applicable. Taking this into consideration, we believe constructing multi-phase samplings designs would be a particularly fruitful avenue for future work (see Holcroft, Rotnitzky, and Robins, 1997; McIsaac and Cook, 2015 for some initial work). For example, if an optimal sampling design was desired, a small random sample could initially be selected from the cohort to obtain validated data and calculate the necessary parameters. These parameters would then be used to define the optimal sampling design to select the phase-three validated data. Another line of future work could be to develop a joint SIMEX method to handle covariate and failure-time outcome error simultaneously. SIMEX is a particularly appealing method when a validation subset is not available and provides a tool to perform sensitivity analyses and robustness checks by varying assumptions about the measurement error structure. We believe developing such an approach for the more challenging error regime of correlated errors in covariates and a failure-time outcome would be worthwhile.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1. Asymptotic theory for RC and RSRC estimators

First, we consider the RC extension for covariate and time-to-event error in Section 3.3. The RC estimator in this setting, $\hat{\beta}_{RC}$, is found by solving the score equation

$$S_{RC}(\beta, \hat{\zeta}) = \sum_{i=1}^{n} \int_0^\tau \left\{ \left\{ \hat{X}_i(\hat{\zeta}_x), Z_i \right\}' - \frac{S^{(1)}(\beta, \hat{\zeta}, t)}{S^{(0)}(\beta, \hat{\zeta}, t)} \right\} d\hat{N}_i(t; \hat{\zeta}_\omega) = 0$$

where $S^{(r)}(\beta, \hat{\zeta}, t) = n^{-1} \sum_{j=1}^{n} \hat{Y}_j(t; \hat{\zeta}_\omega) \left\{ \hat{X}_j(\hat{\zeta}_x), Z_j \right\}^{'\otimes r} \exp(\beta_X' \hat{X}_j(\hat{\zeta}_x) + \beta_Z' Z_j)$ ($a^{\otimes 1}$ is the vector a and $a^{\otimes 0}$ is the scalar 1), and $\left\{ \hat{U}(\hat{\zeta}_\omega), \hat{X}(\hat{\zeta}_x) \right\}$ are as given in Section 3.3. Throughout this section, we assume that (1) (N_i, Y_i, X_i, Z_i) are i.i.d; (2) there exists a finite constant $\tau > 0$ such that $P(U \geq \tau) > 0$; (3) $\int_0^\tau \lambda_0(u) du < \infty$; and (4) $\frac{m}{n} \to p \in (0, 1)$. Define $\beta^\star$ as the solution to $\mathrm{E}\{S_{RC}(\beta, \zeta_0)\} = 0$, which is generally not the same as β. First, we consider consistency for $\beta^\star$ and asymptotic normality for the solution to $S_{RC}(\beta, \zeta_0)$, where $\zeta_0 = (\zeta_{x0}, \zeta_{\omega 0})$ is the true nuisance parameter vector. Then $S_{RC}(\beta, \zeta_0)$, which is based on the standard Cox partial score equation, and thus concave, will have a unique, consistent solution, namely $\beta^\star$, under mild regularity conditions (see Andersen and Gill, 1982). To establish asymptotic normality, we additionally define $\theta^\star = (\beta^\star, \zeta_0)$ and assume that (5) $\frac{\partial}{\partial \theta} S_{RC}(\theta)$ exists and is continuous and bounded for $\theta \in \mathcal{N}(\theta^\star)$, a compact neighborhood of $\theta^\star$; (6) $\frac{\partial}{\partial \theta} S_{RC}(\theta)$ converges to its limit $\mathrm{E}\left\{ \frac{\partial}{\partial \theta} S_{RC}(\theta) \right\}$ uniformly in $\mathcal{N}(\theta^\star)$; (7) $\mathrm{E}\left\{ \frac{\partial}{\partial \theta} S_{RC}(\theta) \right\}$ is nonsingular at $\theta^\star$; and (8) $\mathrm{E}\left[\sup_{\theta \in \mathcal{N}(\theta^\star)} \left\{ \left\{ \hat{X}_j(\hat{\zeta}_x), Z_j \right\} \exp(\beta_X' \hat{X}_j(\hat{\zeta}_x) + \beta_Z' Z_j) \right\}^2 \right] < \infty$. The techniques of Andersen and Gill (1982) can then be used to establish asymptotic normality of the solution to $S_{RC}(\beta, \zeta_0)$. Next, the solution to $S_{RC}(\beta, \hat{\zeta})$, where $\hat{\zeta}$ is our plug-in moment estimator for ζ, can be shown to be consistent and asymptotically normal using Theorem 5.31 in Van der Vaart (1998). The theorem additionally requires that $S_{RC}(\beta, \hat{\zeta})$ be Donsker in $\mathcal{N}(\theta^\star)$. It is well known that the usual Cox score equation is Donsker and given that $\hat{\zeta}$ is a finite dimensional moment estimator, the estimating equations we solve to estimate the nuisance parameters are Donsker as well. $\hat{X}$ and $\hat{U}$ are Lipschitz transformations of X and U involving estimators from a Donsker class of functions, so it follows from Example 19.20 in Van der Vaart (1998) that $S_{RC}(\beta, \hat{\zeta})$ is Donsker.

The arguments above apply to show consistency and asymptotic normality of $\hat{\beta}_{RC}$ from Section 3.2

for time-to-event error only by utilizing the true X instead of $\hat{X}$. Similary, the asymptotic properties of the RSRC estimators from Section 3.4 follow as well due to the fact that we recalibrate a fixed, finite number of times. This results in a finite number of Lipschitz transformations and thus a Donsker class of estimating equations.

A.2. Empirical comparison of sandwich and bootstrap variances for raking estimators

We used the bootstrap to calculate standard errors for the raking estimators due to the fact that we noticed coverage probabilities in some settings under 95% using the sandwich variance estimators. For example, in an independent simulation with settings $\beta_X = \log(3)$, $\sigma^2_\nu = 0.5$, $\sigma^2_\epsilon = 1$, $\sigma_{\nu,\epsilon} = 0.15$, and 25% censoring, the coverage of GRRC was 0.9376 using the sandwich estimator and 0.9524 using the bootstrap. Note that Monte Carlo error cannot explain this undercoverage as the number of simulation runs was 2500, resulting in a 95% confidence interval of $0.95 \pm 1.96\sqrt{\frac{(0.95)(0.05)}{2500}}$, or $(0.9415, 0.9585)$, which does not include 0.9376. The coverage of GRN under the same settings was extremely similar.

A.3. Additive error tables

Table A.1: Simulation results for β_X under correlated, additive measurement error in the outcome and covariate X with $\beta_X = \log 3$, normally distributed error, and 75% censoring for the true event time. 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.

β	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(3)				True	0.146	0.058	0.058	0.003	0.949
	0.5	0.5	0.15	RC	-31.540	0.121	0.121	0.135	0.198
				RSRC	-28.836	0.135	0.134	0.118	0.339
				GRRC	0.819	0.187	0.182	0.033	0.960
				GRN	1.129	0.188	0.183	0.034	0.958
				Naive	-86.163	0.044	0.047	0.898	0.000
				Complete	1.912	0.191	0.197	0.039	0.946
			0.30	RC	-31.567	0.122	0.121	0.135	0.214
				RSRC	-28.627	0.136	0.133	0.117	0.356
				GRRC	0.792	0.188	0.186	0.035	0.955
				GRN	1.329	0.187	0.187	0.035	0.953
				Naive	-102.639	0.046	0.047	1.274	0.000
				Complete	1.766	0.192	0.202	0.041	0.940
		1	0.15	RC	-31.294	0.149	0.148	0.140	0.357
				RSRC	-28.827	0.166	0.164	0.127	0.506
				GRRC	1.283	0.194	0.191	0.037	0.952
				GRN	1.420	0.196	0.192	0.037	0.954
				Naive	-90.669	0.036	0.038	0.994	0.000
				Complete	1.957	0.192	0.200	0.041	0.941
			0.30	RC	-31.431	0.150	0.148	0.141	0.354
				RSRC	-28.754	0.167	0.166	0.127	0.492
				GRRC	1.238	0.194	0.193	0.037	0.958
				GRN	1.611	0.196	0.192	0.037	0.958
				Naive	-101.719	0.038	0.039	1.250	0.000
				Complete	1.839	0.192	0.202	0.041	0.942
	1	0.5	0.15	RC	-33.415	0.123	0.124	0.150	0.178
				RSRC	-31.695	0.137	0.135	0.139	0.288
				GRRC	0.847	0.190	0.187	0.035	0.954
				GRN	1.174	0.190	0.188	0.036	0.950
				Naive	-79.646	0.044	0.046	0.768	0.000
				Complete	1.930	0.193	0.202	0.041	0.946
			0.30	RC	-33.652	0.124	0.123	0.152	0.178
				RSRC	-31.494	0.138	0.138	0.139	0.303
				GRRC	0.874	0.188	0.186	0.034	0.958
				GRN	1.302	0.188	0.186	0.035	0.956
				Naive	-90.541	0.045	0.046	0.992	0.000
				Complete	1.866	0.192	0.201	0.041	0.948
		1	0.15	RC	-33.378	0.152	0.151	0.157	0.328
				RSRC	-31.804	11.643	0.166	0.149	0.438
				GRRC	1.129	0.195	0.193	0.037	0.954
				GRN	1.311	0.196	0.193	0.038	0.952
				Naive	-86.191	0.036	0.038	0.898	0.000
				Complete	1.866	0.192	0.201	0.041	0.946
			0.30	RC	-33.533	0.153	0.151	0.159	0.328
				RSRC	-32.04	3.224	0.164	0.151	0.439
				GRRC	1.202	0.194	0.191	0.037	0.951
				GRN	1.538	0.195	0.192	0.037	0.952
				Naive	-93.700	0.036	0.038	1.061	0.000
				Complete	1.893	0.192	0.200	0.040	0.944

Table A.2: Simulation results for β_Z under additive measurement error only in the outcome with normally distributed error and 25 and 75$\%$ censoring for the true event time. 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.

$\%$ Censoring	β_X	σ^2_ν	Method	$\%$ Bias	ASE	ESE	MSE	CP
25	log(1.5)		True	0.072	0.032	0.033	0.001	0.949
		0.5	RC	-12.523	0.044	0.043	0.009	0.493
			RSRC	-4.891	0.051	0.052	0.004	0.884
			GRRC	0.115	0.066	0.065	0.004	0.956
			GRN	-0.014	0.066	0.065	0.004	0.958
			Naive	12.003	0.033	0.034	0.008	0.294
			Complete	1.428	0.104	0.105	0.011	0.956
		1	RC	-18.495	0.048	0.048	0.019	0.247
			RSRC	-7.617	0.058	0.059	0.006	0.847
			GRRC	0.087	0.074	0.073	0.005	0.957
			GRN	-0.029	0.074	0.072	0.005	0.957
			Naive	2.741	0.032	0.033	0.002	0.902
			Complete	1.385	0.104	0.105	0.011	0.954
	log(3)		True	0.043	0.033	0.033	0.001	0.949
		0.5	RC	-26.719	0.048	0.048	0.037	0.030
			RSRC	-18.712	0.055	0.057	0.020	0.343
			GRRC	-0.851	0.086	0.087	0.008	0.944
			GRN	-1.010	0.084	0.082	0.007	0.948
			Naive	0.144	0.032	0.037	0.001	0.913
			Complete	1.284	0.106	0.108	0.012	0.952
		1	RC	-32.951	0.051	0.051	0.055	0.006
			RSRC	-22.881	0.060	0.062	0.029	0.264
			GRRC	-0.793	0.090	0.088	0.008	0.946
			GRN	-0.866	0.089	0.088	0.008	0.947
			Naive	-10.777	0.032	0.036	0.007	0.362
			Complete	1.298	0.106	0.108	0.012	0.955
75	log(1.5)		True	0.130	0.056	0.056	0.003	0.954
		0.5	RC	-14.874	0.079	0.079	0.017	0.72
			RSRC	-12.248	0.087	0.090	0.015	0.816
			GRRC	-0.101	0.121	0.119	0.014	0.954
			GRN	-0.707	0.129	0.128	0.016	0.952
			Naive	32.244	0.057	0.059	0.053	0.020
			Complete	1.962	0.182	0.190	0.036	0.944
		1	RC	-17.226	0.082	0.082	0.021	0.681
			RSRC	-14.946	0.090	0.094	0.020	0.782
			GRRC	-0.390	0.127	0.124	0.015	0.954
			GRN	-1.010	0.131	0.13	0.017	0.946
			Naive	17.760	0.056	0.058	0.019	0.400
			Complete	1.818	0.182	0.190	0.036	0.944
	log(3)		True	0.188	0.054	0.055	0.003	0.948
		0.5	RC	-30.268	0.083	0.084	0.051	0.288
			RSRC	-27.685	0.092	0.096	0.046	0.443
			GRRC	-1.068	0.148	0.145	0.021	0.944
			GRN	-1.746	0.152	0.149	0.022	0.944
			Naive	20.111	0.055	0.062	0.023	0.297
			Complete	2.265	0.178	0.186	0.035	0.948
		1	RC	-32.691	0.085	0.087	0.059	0.237
			RSRC	-30.628	0.094	0.099	0.055	0.383
			GRRC	-1.096	0.152	0.150	0.022	0.950
			GRN	-1.890	0.154	0.153	0.024	0.943
			Naive	3.982	0.054	0.061	0.004	0.880
			Complete	2.121	0.180	0.188	0.036	0.944

Table A.3: Simulation results for β_Z under additive, general measurement error in the outcome and covariate X with $\beta_X = \log 1.5$, normally distributed error, and 25% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(1.5)				True	0.072	0.032	0.033	0.001	0.949
	0.5	0.5	0.15	RC	-13.936	0.050	0.049	0.012	0.510
				RSRC	-5.554	0.058	0.058	0.005	0.901
				GRRC	0.245	0.071	0.069	0.005	0.964
				GRN	0.115	0.070	0.068	0.005	0.963
				Naive	25.189	0.031	0.032	0.032	0.000
				Complete	1.428	0.104	0.105	0.011	0.956
			0.30	RC	-13.893	0.050	0.050	0.012	0.526
				RSRC	-5.324	0.058	0.059	0.005	0.903
				GRRC	0.245	0.069	0.067	0.004	0.964
				GRN	0.058	0.069	0.066	0.004	0.968
				Naive	27.656	0.031	0.033	0.038	0.000
				Complete	1.486	0.104	0.106	0.011	0.951
		1	0.15	RC	-14.153	0.054	0.053	0.012	0.568
				RSRC	-5.713	0.062	0.062	0.006	0.912
				GRRC	0.346	0.073	0.071	0.005	0.964
				GRN	0.245	0.072	0.070	0.005	0.965
				Naive	26.113	0.031	0.032	0.034	0.000
				Complete	1.457	0.104	0.106	0.011	0.953
			0.30	RC	-14.138	0.055	0.053	0.012	0.578
				RSRC	-5.526	0.062	0.062	0.005	0.917
				GRRC	0.332	0.072	0.069	0.005	0.966
				GRN	0.216	0.071	0.068	0.005	0.966
				Naive	27.786	0.031	0.032	0.038	0.000
				Complete	1.428	0.104	0.106	0.011	0.954
	1	0.5	0.15	RC	-19.563	0.054	0.053	0.021	0.288
				RSRC	-8.094	0.065	0.066	0.008	0.850
				GRRC	0.231	0.078	0.076	0.006	0.962
				GRN	0.115	0.077	0.075	0.006	0.960
				Naive	15.581	0.030	0.031	0.013	0.047
				Complete	1.385	0.104	0.105	0.011	0.954
			0.30	RC	-19.606	0.055	0.054	0.021	0.300
				RSRC	-7.906	0.065	0.066	0.007	0.867
				GRRC	0.216	0.077	0.075	0.006	0.958
				GRN	0.058	0.077	0.074	0.006	0.964
				Naive	17.731	0.030	0.031	0.016	0.020
				Complete	1.443	0.104	0.106	0.011	0.954
		1	0.15	RC	-19.707	0.059	0.058	0.022	0.360
				RSRC	-8.094	0.070	0.070	0.008	0.881
				GRRC	0.317	0.080	0.078	0.006	0.960
				GRN	0.202	0.079	0.077	0.006	0.960
				Naive	16.504	0.030	0.031	0.014	0.030
				Complete	1.472	0.104	0.106	0.011	0.955
			0.30	RC	-19.736	0.060	0.058	0.022	0.359
				RSRC	-7.920	0.070	0.070	0.008	0.886
				GRRC	0.260	0.079	0.077	0.006	0.961
				GRN	0.159	0.078	0.076	0.006	0.962
				Naive	17.99	0.030	0.031	0.016	0.014
				Complete	1.371	0.104	0.106	0.011	0.956

Table A.4: Simulation results for β_Z under correlated, additive measurement error in the outcome and covariate X with $\beta_X = \log 1.5$, normally distributed error, and 75% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(1.5)				True	0.173	0.056	0.056	0.003	0.954
	0.5	0.5	0.15	RC	-15.105	0.088	0.087	0.019	0.785
				RSRC	-12.220	0.098	0.099	0.017	0.853
				GRRC	-0.014	0.131	0.130	0.017	0.954
				GRN	-0.692	0.139	0.139	0.019	0.950
				Naive	45.272	0.051	0.053	0.101	0.000
				Complete	1.962	0.182	0.190	0.036	0.944
			0.30	RC	-14.917	0.088	0.087	0.018	0.802
				RSRC	-11.902	0.098	0.100	0.017	0.858
				GRRC	0.043	0.129	0.128	0.016	0.956
				GRN	-0.808	0.137	0.137	0.019	0.946
				Naive	55.457	0.052	0.055	0.151	0.000
				Complete	1.861	0.182	0.191	0.037	0.940
		1	0.15	RC	-15.163	0.097	0.095	0.020	0.818
				RSRC	-12.119	0.107	0.108	0.019	0.887
				GRRC	-0.014	0.136	0.135	0.018	0.954
				GRN	-0.548	0.143	0.143	0.020	0.950
				Naive	44.103	0.051	0.053	0.096	0.000
				Complete	2.265	0.182	0.191	0.037	0.948
			0.30	RC	-14.975	0.097	0.094	0.020	0.827
				RSRC	-12.047	0.107	0.108	0.018	0.883
				GRRC	-0.144	0.135	0.132	0.018	0.956
				GRN	-0.779	0.142	0.141	0.020	0.950
				Naive	50.595	0.051	0.053	0.126	0.000
				Complete	2.049	0.182	0.187	0.035	0.950
	1	0.5	0.15	RC	-17.543	0.092	0.091	0.023	0.742
				RSRC	-15.018	0.101	0.104	0.022	0.813
				GRRC	-0.274	0.137	0.134	0.018	0.955
				GRN	-0.923	0.141	0.140	0.020	0.950
				Naive	30.701	0.050	0.052	0.048	0.010
				Complete	1.818	0.182	0.190	0.036	0.944
			0.30	RC	-17.586	0.092	0.090	0.023	0.748
				RSRC	-14.903	0.101	0.104	0.021	0.830
				GRRC	-0.115	0.134	0.133	0.018	0.954
				GRN	-0.822	0.140	0.138	0.019	0.948
				Naive	35.894	0.050	0.052	0.065	0.000
				Complete	1.847	0.181	0.185	0.034	0.944
		1	0.15	RC	-17.644	0.100	0.098	0.025	0.780
				RSRC	-14.816	0.111	0.113	0.023	0.846
				GRRC	-0.188	0.139	0.140	0.020	0.944
				GRN	-0.649	0.144	0.145	0.021	0.944
				Naive	30.441	0.049	0.052	0.047	0.010
				Complete	1.760	0.181	0.190	0.036	0.941
			0.30	RC	-17.500	0.101	0.098	0.024	0.789
				RSRC	-14.946	0.110	0.113	0.024	0.849
				GRRC	-0.144	0.140	0.138	0.019	0.955
				GRN	-0.750	0.144	0.145	0.021	0.946
				Naive	34.192	0.050	0.052	0.059	0.001
				Complete	1.746	0.182	0.190	0.036	0.946

Table A.5: Simulation results for β_Z under correlated, additive measurement error in the outcome and covariate X with $\beta_X = \log 3$, normally distributed error, and 25% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(3)				True	0.043	0.033	0.033	0.001	0.945
	0.5	0.5	0.15	RC	-31.667	0.060	0.060	0.052	0.036
				RSRC	-20.789	0.070	0.072	0.026	0.466
				GRRC	-0.433	0.088	0.086	0.007	0.952
				GRN	-0.548	0.087	0.084	0.007	0.952
				Naive	26.892	0.031	0.034	0.036	0.000
				Complete	1.284	0.106	0.108	0.012	0.952
			0.30	RC	-32.432	0.063	0.062	0.054	0.036
				RSRC	-20.602	0.072	0.074	0.026	0.497
				GRRC	-0.303	0.087	0.084	0.007	0.956
				GRN	-0.491	0.087	0.084	0.007	0.957
				Naive	26.661	0.031	0.034	0.035	0.000
				Complete	1.284	0.106	0.107	0.012	0.950
		1	0.15	RC	-32.720	0.067	0.066	0.056	0.050
				RSRC	-20.977	0.076	0.077	0.027	0.522
				GRRC	-0.303	0.088	0.087	0.008	0.955
				GRN	-0.404	0.087	0.085	0.007	0.959
				Naive	29.316	0.031	0.034	0.042	0.000
				Complete	1.298	0.106	0.108	0.012	0.953
			0.30	RC	-33.225	0.069	0.068	0.058	0.050
				RSRC	-20.789	0.077	0.078	0.027	0.550
				GRRC	-0.188	0.088	0.086	0.007	0.956
				GRN	-0.317	0.087	0.085	0.007	0.954
				Naive	29.128	0.031	0.034	0.042	0.000
				Complete	1.313	0.106	0.108	0.012	0.950
	1	0.5	0.15	RC	-36.341	0.062	0.062	0.067	0.010
				RSRC	-23.920	0.073	0.074	0.033	0.371
				GRRC	-0.375	0.092	0.090	0.008	0.954
				GRN	-0.519	0.091	0.089	0.008	0.955
				Naive	16.605	0.030	0.034	0.014	0.040
				Complete	1.298	0.106	0.108	0.012	0.955
			0.30	RC	-37.048	0.064	0.063	0.070	0.008
				RSRC	-23.761	0.074	0.076	0.033	0.398
				GRRC	-0.361	0.091	0.089	0.008	0.956
				GRN	-0.447	0.091	0.089	0.008	0.953
				Naive	16.764	0.030	0.033	0.015	0.038
				Complete	1.356	0.106	0.108	0.012	0.948
		1	0.15	RC	-37.063	0.069	0.068	0.071	0.018
				RSRC	-23.660	0.079	0.080	0.033	0.454
				GRRC	-0.274	0.092	0.090	0.008	0.955
				GRN	-0.361	0.091	0.089	0.008	0.952
				Naive	19.274	0.030	0.033	0.019	0.012
				Complete	1.284	0.106	0.108	0.012	0.950
			0.30	RC	-37.524	0.070	0.069	0.072	0.017
				RSRC	-23.574	0.080	0.081	0.033	0.467
				GRRC	-0.202	0.092	0.090	0.008	0.956
				GRN	-0.289	0.091	0.089	0.008	0.956
				Naive	19.361	0.030	0.033	0.019	0.012
				Complete	1.327	0.106	0.108	0.012	0.950

A.4. Classical measurement error tables

Table A.6: Simulation results for $\beta_X = \log 1.5$ under correlated additive measurement error in the outcome and classical measurement error in the covariate X with normally distributed error and 25% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(1.5)				True	-0.049	0.030	0.031	0.001	0.946
	0.5	0.5	0.15	RC	-13.762	0.056	0.057	0.006	0.800
				RSRC	-6.141	0.066	0.065	0.005	0.920
				GRRC	0.123	0.081	0.082	0.007	0.949
				GRN	0.296	0.081	0.082	0.007	0.947
				Naive	-78.428	0.023	0.023	0.102	0.000
				Complete	0.123	0.098	0.099	0.010	0.952
			0.30	RC	-13.589	0.057	0.057	0.006	0.800
				RSRC	-5.944	0.068	0.066	0.005	0.929
				GRRC	0.222	0.081	0.082	0.007	0.945
				GRN	0.518	0.081	0.082	0.007	0.942
				Naive	-93.621	0.023	0.024	0.145	0.000
				Complete	0.148	0.098	0.099	0.010	0.954
		1	0.15	RC	-13.836	0.067	0.067	0.008	0.832
				RSRC	-6.758	0.079	0.078	0.007	0.918
				GRRC	0.000	0.087	0.088	0.008	0.946
				GRN	0.148	0.087	0.088	0.008	0.946
				Naive	-84.594	0.019	0.020	0.118	0.000
				Complete	0.173	0.098	0.099	0.010	0.952
			0.30	RC	-13.688	0.068	0.068	0.008	0.836
				RSRC	-6.708	0.080	0.079	0.007	0.912
				GRRC	0.247	0.087	0.088	0.008	0.948
				GRN	0.469	0.087	0.088	0.008	0.944
				Naive	-95.471	0.019	0.020	0.150	0.000
				Complete	0.271	0.098	0.098	0.010	0.952
	1	0.5	0.15	RC	-19.286	0.062	0.062	0.010	0.734
				RSRC	-9.224	0.074	0.073	0.007	0.907
				GRRC	0.148	0.083	0.084	0.007	0.942
				GRN	0.271	0.083	0.084	0.007	0.943
				Naive	-78.552	0.023	0.023	0.102	0.000
				Complete	0.247	0.098	0.098	0.010	0.954
			0.30	RC	-19.286	0.062	0.063	0.010	0.732
				RSRC	-9.372	0.076	0.073	0.007	0.904
				GRRC	0.197	0.083	0.084	0.007	0.944
				GRN	0.370	0.083	0.084	0.007	0.946
				Naive	-91.993	0.023	0.023	0.140	0.000
				Complete	0.173	0.098	0.099	0.010	0.948
		1	0.15	RC	-19.139	0.074	0.074	0.012	0.791
				RSRC	-10.013	0.088	0.087	0.009	0.907
				GRRC	0.025	0.089	0.090	0.008	0.942
				GRN	0.123	0.089	0.090	0.008	0.944
				Naive	-84.619	0.019	0.020	0.118	0.000
				Complete	0.271	0.098	0.099	0.010	0.954
			0.30	RC	-19.262	0.074	0.074	0.012	0.779
				RSRC	-10.137	0.089	0.088	0.009	0.902
				GRRC	0.099	0.089	0.090	0.008	0.944
				GRN	0.247	0.089	0.090	0.008	0.943
				Naive	-94.336	0.019	0.020	0.147	0.000
				Complete	0.247	0.098	0.098	0.010	0.953

Table A.7: Simulation results for $\beta_X = \log 1.5$ under correlated additive measurement error in the outcome and classical measurement error in the covariate X with normally distributed error and 75% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(1.5)				True	0.123	0.054	0.054	0.003	0.949
	0.5	0.5	0.15	RC	-14.946	0.104	0.103	0.014	0.901
				RSRC	-12.406	0.114	0.113	0.015	0.918
				GRRC	0.148	0.151	0.149	0.022	0.956
				GRN	0.789	0.153	0.152	0.023	0.952
				Naive	-115.916	0.043	0.043	0.223	0.000
				Complete	0.543	0.177	0.183	0.033	0.951
			0.30	RC	-14.675	0.104	0.103	0.014	0.896
				RSRC	-12.011	0.115	0.112	0.015	0.925
				GRRC	-0.296	0.150	0.147	0.022	0.956
				GRN	1.233	0.149	0.147	0.022	0.953
				Naive	-156.462	0.045	0.045	0.404	0.000
				Complete	0.123	0.176	0.181	0.033	0.952
		1	0.15	RC	-14.970	0.124	0.123	0.019	0.918
				RSRC	-12.677	0.137	0.138	0.022	0.926
				GRRC	-0.370	0.162	0.160	0.026	0.954
				GRN	0.074	0.164	0.163	0.026	0.956
				Naive	-111.082	0.036	0.036	0.204	0.000
				Complete	-0.074	0.176	0.180	0.032	0.947
			0.30	RC	-14.477	0.124	0.123	0.019	0.919
				RSRC	-12.529	0.137	0.137	0.021	0.929
				GRRC	-0.074	0.162	0.160	0.026	0.958
				GRN	0.715	0.164	0.162	0.026	0.955
				Naive	-138.212	0.037	0.038	0.315	0.000
				Complete	0.247	0.176	0.181	0.033	0.948
	1	0.5	0.15	RC	-17.091	0.108	0.107	0.016	0.896
				RSRC	-15.587	0.117	0.118	0.018	0.901
				GRRC	-0.074	0.153	0.151	0.023	0.960
				GRN	0.666	0.154	0.152	0.023	0.956
				Naive	-99.367	0.042	0.042	0.164	0.000
				Complete	0.617	0.177	0.181	0.033	0.952
			0.30	RC	-17.042	0.108	0.108	0.016	0.890
				RSRC	-15.538	0.118	0.119	0.018	0.901
				GRRC	-0.173	0.153	0.151	0.023	0.956
				GRN	0.987	0.154	0.152	0.023	0.950
				Naive	-126.003	0.044	0.044	0.263	0.000
				Complete	0.592	0.178	0.183	0.034	0.950
		1	0.15	RC	-16.993	0.129	0.127	0.021	0.910
				RSRC	-15.784	0.140	0.142	0.024	0.910
				GRRC	0.247	0.165	0.164	0.027	0.959
				GRN	0.765	0.166	0.166	0.028	0.956
				Naive	-99.614	0.036	0.036	0.164	0.000
				Complete	0.518	0.178	0.183	0.034	0.947
			0.30	RC	-17.067	0.129	0.127	0.021	0.908
				RSRC	-15.316	0.141	0.141	0.024	0.914
				GRRC	-0.222	0.164	0.162	0.026	0.962
				GRN	0.567	0.165	0.164	0.027	0.963
				Naive	-118.136	0.036	0.037	0.231	0.000
				Complete	0.222	0.177	0.182	0.033	0.947

Table A.8: Simulation results for $\beta_X = \log 3$ under correlated additive measurement error in the outcome and classical measurement error in the covariate X with normally distributed error and 25% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(3)				True	0.046	0.037	0.036	0.001	0.955
	0.5	0.5	0.15	RC	-30.921	0.073	0.074	0.121	0.018
				RSRC	-22.665	0.088	0.087	0.070	0.216
				GRRC	0.200	0.112	0.110	0.012	0.952
				GRN	0.291	0.112	0.111	0.012	0.954
				Naive	-70.016	0.023	0.026	0.592	0.000
				Complete	0.792	0.118	0.118	0.014	0.957
			0.30	RC	-31.658	0.076	0.076	0.127	0.024
				RSRC	-22.774	0.092	0.091	0.071	0.240
				GRRC	0.300	0.112	0.110	0.012	0.957
				GRN	0.319	0.112	0.110	0.012	0.954
				Naive	-76.032	0.023	0.025	0.698	0.000
				Complete	0.737	0.118	0.117	0.014	0.955
		1	0.15	RC	-31.604	0.087	0.088	0.128	0.062
				RSRC	-23.903	0.104	0.103	0.080	0.304
				GRRC	0.401	0.115	0.113	0.013	0.953
				GRN	0.428	0.115	0.113	0.013	0.952
				Naive	-78.572	0.020	0.021	0.746	0.000
				Complete	0.792	0.118	0.118	0.014	0.960
			0.30	RC	-32.159	0.090	0.089	0.133	0.065
				RSRC	-24.058	0.108	0.107	0.081	0.328
				GRRC	0.437	0.115	0.113	0.013	0.956
				GRN	0.519	0.115	0.113	0.013	0.952
				Naive	-82.713	0.019	0.021	0.826	0.000
				Complete	0.801	0.118	0.118	0.014	0.958
	1	0.5	0.15	RC	-35.681	0.075	0.076	0.159	0.008
				RSRC	-26.488	0.090	0.090	0.093	0.150
				GRRC	0.191	0.113	0.112	0.012	0.950
				GRN	0.218	0.113	0.112	0.012	0.952
				Naive	-71.244	0.023	0.025	0.613	0.000
				Complete	0.746	0.118	0.118	0.014	0.956
			0.30	RC	-36.382	0.077	0.077	0.166	0.009
				RSRC	-26.961	0.093	0.092	0.096	0.156
				GRRC	0.300	0.113	0.111	0.012	0.954
				GRN	0.300	0.113	0.111	0.012	0.956
				Naive	-76.360	0.023	0.025	0.704	0.000
				Complete	0.737	0.118	0.118	0.014	0.956
		1	0.15	RC	-36.055	0.090	0.090	0.165	0.034
				RSRC	-27.835	0.107	0.106	0.105	0.222
				GRRC	0.382	0.116	0.114	0.013	0.948
				GRN	0.428	0.115	0.114	0.013	0.950
				Naive	-79.437	0.020	0.021	0.762	0.000
				Complete	0.801	0.118	0.118	0.014	0.957
			0.30	RC	-36.564	0.091	0.091	0.170	0.039
				RSRC	-28.190	0.110	0.108	0.108	0.231
				GRRC	0.382	0.116	0.114	0.013	0.952
				GRN	0.437	0.115	0.114	0.013	0.954
				Naive	-82.977	0.019	0.021	0.831	0.000
				Complete	0.765	0.118	0.118	0.014	0.957

A.5. Gamma distributed error tables

Table A.9: Simulation results for β_X under additive measurement error only in the outcome with gamma distributed error and 25 and 75% censoring for the true event time. 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.

% Censoring	β_X	σ^2_ν	Method	% Bias	ASE	ESE	MSE	CP
25	log(1.5)		True	0.099	0.030	0.032	0.001	0.942
		0.5	RC	-19.558	0.045	0.045	0.008	0.574
			RSRC	-4.563	0.060	0.059	0.004	0.935
			GRRC	-0.567	0.067	0.067	0.004	0.949
			GRN	-0.567	0.066	0.067	0.004	0.947
			Naive	-31.371	0.030	0.032	0.017	0.018
			Complete	0.543	0.098	0.100	0.010	0.952
		1	RC	-28.905	0.052	0.052	0.016	0.380
			RSRC	-8.879	0.071	0.071	0.006	0.918
			GRRC	-0.617	0.075	0.076	0.006	0.950
			GRN	-0.592	0.075	0.076	0.006	0.945
			Naive	-38.869	0.029	0.032	0.026	0.001
			Complete	0.617	0.098	0.100	0.010	0.949
	log(3)		True	0.155	0.037	0.037	0.001	0.941
		0.5	RC	-33.733	0.055	0.056	0.140	0.000
			RSRC	-23.156	0.067	0.069	0.069	0.041
			GRRC	-1.211	0.113	0.116	0.014	0.923
			GRN	-1.211	0.113	0.119	0.014	0.920
			Naive	-38.166	0.030	0.043	0.178	0.000
			Complete	0.819	0.119	0.121	0.015	0.948
		1	RC	-41.334	0.058	0.059	0.210	0.000
			RSRC	-28.254	0.074	0.076	0.102	0.019
			GRRC	-0.892	0.115	0.116	0.014	0.936
			GRN	-0.856	0.115	0.122	0.015	0.928
			Naive	-44.948	0.030	0.041	0.246	0.000
			Complete	0.874	0.119	0.122	0.015	0.946
75	log(1.5)		True	0.395	0.054	0.056	0.003	0.936
		0.5	RC	-19.829	0.080	0.080	0.013	0.834
			RSRC	-9.989	0.100	0.103	0.012	0.921
			GRRC	0.518	0.118	0.118	0.014	0.954
			GRN	0.543	0.116	0.116	0.014	0.956
			Naive	-40.719	0.054	0.057	0.031	0.156
			Complete	2.318	0.177	0.180	0.032	0.950
		1	RC	-19.903	0.089	0.091	0.015	0.854
			RSRC	-13.762	0.112	0.119	0.017	0.906
			GRRC	0.641	0.121	0.120	0.014	0.958
			GRN	0.641	0.119	0.118	0.014	0.952
			Naive	-36.279	0.054	0.058	0.025	0.242
			Complete	2.738	0.178	0.181	0.033	0.948
	log(3)		True	0.300	0.058	0.059	0.003	0.948
		0.5	RC	-33.187	0.086	0.087	0.140	0.010
			RSRC	-28.527	0.107	0.110	0.110	0.168
			GRRC	-0.692	0.174	0.176	0.031	0.937
			GRN	-0.546	0.173	0.180	0.032	0.940
			Naive	-40.469	0.053	0.068	0.202	0.000
			Complete	2.458	0.193	0.200	0.041	0.946
		1	RC	-33.824	0.097	0.100	0.148	0.022
			RSRC	-30.957	0.121	0.128	0.132	0.201
			GRRC	-0.628	0.176	0.183	0.034	0.938
			GRN	-0.528	0.174	0.183	0.034	0.934
			Naive	-39.186	0.053	0.068	0.190	0.000
			Complete	2.485	0.193	0.204	0.042	0.944

Table A.10: Simulation results for $\beta_X = \log 1.5$ under correlated, additive measurement error in the outcome and covariate X with gamma distributed error and 25% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(1.5)				True	0.099	0.030	0.032	0.001	0.942
	0.5	0.5	0.15	RC	-23.060	0.057	0.057	0.012	0.601
				RSRC	-5.944	0.075	0.076	0.006	0.928
				GRRC	-0.888	0.081	0.082	0.007	0.945
				GRN	-0.814	0.081	0.082	0.007	0.943
				Naive	-56.972	0.025	0.028	0.054	0.000
				Complete	0.543	0.098	0.100	0.010	0.952
			0.30	RC	-25.206	0.058	0.058	0.014	0.547
				RSRC	-4.760	0.077	0.079	0.007	0.925
				GRRC	-1.282	0.082	0.084	0.007	0.941
				GRN	-1.110	0.082	0.083	0.007	0.943
				Naive	-62.718	0.025	0.028	0.066	0.000
				Complete	0.543	0.098	0.099	0.010	0.952
		1	0.15	RC	-25.403	0.068	0.067	0.015	0.607
				RSRC	-8.903	0.086	0.087	0.009	0.906
				GRRC	-1.726	0.087	0.089	0.008	0.938
				GRN	-1.578	0.086	0.088	0.008	0.942
				Naive	-66.689	0.022	0.025	0.074	0.000
				Complete	0.469	0.098	0.100	0.010	0.952
			0.30	RC	-27.499	0.068	0.068	0.017	0.562
				RSRC	-7.941	0.088	0.091	0.009	0.901
				GRRC	-1.899	0.088	0.090	0.008	0.934
				GRN	-1.603	0.087	0.089	0.008	0.938
				Naive	-71.030	0.022	0.026	0.084	0.000
				Complete	0.641	0.098	0.100	0.010	0.946
	1	0.5	0.15	RC	-31.988	0.064	0.063	0.021	0.468
				RSRC	-9.323	0.087	0.090	0.009	0.912
				GRRC	-0.863	0.085	0.086	0.007	0.950
				GRN	-0.789	0.085	0.086	0.007	0.949
				Naive	-61.189	0.025	0.028	0.062	0.000
				Complete	0.617	0.098	0.10	0.010	0.949
			0.30	RC	-33.961	0.064	0.064	0.023	0.417
				RSRC	-7.769	0.088	0.092	0.009	0.910
				GRRC	-1.233	0.086	0.087	0.008	0.944
				GRN	-1.061	0.086	0.086	0.008	0.948
				Naive	-66.023	0.025	0.028	0.072	0.000
				Complete	0.543	0.098	0.100	0.010	0.950
		1	0.15	RC	-33.862	0.074	0.073	0.024	0.506
				RSRC	-11.666	0.099	0.102	0.013	0.899
				GRRC	-1.430	0.090	0.091	0.008	0.942
				GRN	-1.307	0.090	0.090	0.008	0.944
				Naive	-69.870	0.022	0.025	0.081	0.000
				Complete	0.617	0.098	0.100	0.010	0.954
			0.30	RC	-35.737	0.075	0.074	0.026	0.462
				RSRC	-10.432	0.101	0.104	0.013	0.905
				GRRC	-1.554	0.090	0.092	0.009	0.948
				GRN	-1.455	0.090	0.091	0.008	0.948
				Naive	-73.447	0.022	0.025	0.089	0.000
				Complete	0.567	0.098	0.100	0.010	0.952

Table A.11: Simulation results for $\beta_X = \log 1.5$ under correlated, additive measurement error in the outcome and covariate X with gamma distributed error and 75% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(1.5)				True	0.395	0.054	0.056	0.003	0.936
	0.5	0.5	0.15	RC	-25.946	0.097	0.095	0.020	0.807
				RSRC	-7.966	0.124	0.129	0.018	0.919
				GRRC	1.110	0.148	0.146	0.021	0.953
				GRN	1.159	0.148	0.144	0.021	0.952
				Naive	-68.835	0.046	0.049	0.080	0.000
				Complete	2.318	0.177	0.180	0.032	0.950
			0.30	RC	-30.582	0.097	0.096	0.025	0.734
				RSRC	-5.105	0.125	0.132	0.018	0.920
				GRRC	1.061	0.149	0.146	0.021	0.950
				GRN	1.554	0.148	0.144	0.021	0.952
				Naive	-79.292	0.046	0.050	0.106	0.000
				Complete	2.417	0.177	0.182	0.033	0.947
		1	0.15	RC	-27.154	0.111	0.110	0.024	0.806
				RSRC	-9.939	0.140	0.150	0.024	0.912
				GRRC	0.937	0.158	0.153	0.023	0.958
				GRN	1.061	0.157	0.152	0.023	0.954
				Naive	-75.666	0.040	0.043	0.096	0.000
				Complete	2.220	0.177	0.181	0.033	0.947
			0.30	RC	-31.470	0.110	0.109	0.028	0.748
				RSRC	-7.670	0.143	0.155	0.025	0.908
				GRRC	0.913	0.158	0.153	0.024	0.954
				GRN	1.529	0.157	0.153	0.023	0.953
				Naive	-83.287	0.040	0.043	0.116	0.000
				Complete	2.664	0.177	0.179	0.032	0.952
	1	0.5	0.15	RC	-25.107	0.107	0.108	0.022	0.842
				RSRC	-12.110	0.138	0.149	0.025	0.906
				GRRC	1.554	0.150	0.145	0.021	0.954
				GRN	1.603	0.149	0.144	0.021	0.954
				Naive	-63.088	0.046	0.050	0.068	0.001
				Complete	2.738	0.178	0.181	0.033	0.948
			0.30	RC	-27.820	0.106	0.105	0.024	0.810
				RSRC	-8.484	0.138	0.150	0.024	0.917
				GRRC	1.159	0.150	0.149	0.022	0.952
				GRN	1.332	0.149	0.147	0.022	0.949
				Naive	-70.413	0.046	0.049	0.084	0.000
				Complete	2.713	0.177	0.182	0.033	0.949
		1	0.15	RC	-26.439	0.122	0.122	0.026	0.836
				RSRC	-14.675	0.155	0.171	0.033	0.895
				GRRC	0.715	0.158	0.152	0.023	0.954
				GRN	1.061	0.157	0.152	0.023	0.952
				Naive	-71.128	0.040	0.042	0.085	0.000
				Complete	2.220	0.177	0.178	0.032	0.954
			0.30	RC	-29.448	0.121	0.121	0.029	0.810
				RSRC	-11.444	0.156	0.174	0.032	0.899
				GRRC	1.208	0.160	0.154	0.024	0.956
				GRN	1.455	0.158	0.152	0.023	0.954
				Naive	-76.801	0.040	0.043	0.099	0.000
				Complete	3.132	0.178	0.178	0.032	0.955

Table A.12: Simulation results for $\beta_X = \log 3$ under correlated, additive measurement error in the outcome and covariate X with gamma distributed error and 25% censoring for the true event time. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	% Bias	ASE	ESE	MSE	CP
log(3)				True	0.146	0.037	0.038	0.001	0.944
	0.5	0.5	0.15	RC	-39.113	0.071	0.072	0.190	0.002
				RSRC	-28.864	0.089	0.094	0.109	0.106
				GRRC	-0.965	0.118	0.119	0.014	0.937
				GRN	-0.901	0.118	0.119	0.014	0.937
				Naive	-60.376	0.025	0.036	0.441	0.000
				Complete	0.819	0.119	0.121	0.015	0.948
			0.30	RC	-40.879	0.072	0.074	0.207	0.002
				RSRC	-29.710	0.093	0.099	0.116	0.122
				GRRC	-1.047	0.122	0.122	0.015	0.936
				GRN	-0.947	0.120	0.121	0.015	0.934
				Naive	-62.998	0.025	0.038	0.480	0.000
				Complete	0.892	0.119	0.122	0.015	0.948
		1	0.15	RC	-44.438	0.090	0.093	0.247	0.018
				RSRC	-34.726	0.106	0.114	0.159	0.128
				GRRC	-1.265	0.129	0.132	0.018	0.932
				GRN	-1.192	0.127	0.128	0.016	0.931
				Naive	-71.254	0.020	0.035	0.614	0.000
				Complete	0.856	0.119	0.121	0.015	0.948
			0.30	RC	-45.912	0.090	0.094	0.263	0.015
				RSRC	-35.208	0.110	0.119	0.164	0.145
				GRRC	-1.338	0.131	0.131	0.017	0.930
				GRN	-1.320	0.128	0.129	0.017	0.930
				Naive	-73.056	0.020	0.036	0.646	0.000
				Complete	0.819	0.119	0.121	0.015	0.947
	1	0.5	0.15	RC	-45.066	0.074	0.074	0.251	0.000
				RSRC	-32.204	0.095	0.100	0.135	0.079
				GRRC	-0.664	0.119	0.120	0.014	0.941
				GRN	-0.674	0.118	0.119	0.014	0.937
				Naive	-63.871	0.025	0.034	0.494	0.000
				Complete	0.874	0.119	0.122	0.015	0.946
			0.30	RC	-46.322	0.074	0.074	0.264	0.000
				RSRC	-32.668	0.097	0.103	0.139	0.095
				GRRC	-0.819	0.121	0.120	0.014	0.938
				GRN	-0.755	0.119	0.120	0.014	0.937
				Naive	-65.883	0.025	0.035	0.525	0.000
				Complete	0.847	0.119	0.121	0.015	0.950
		1	0.15	RC	-49.171	0.091	0.093	0.300	0.008
				RSRC	-36.755	0.112	0.118	0.177	0.124
				GRRC	-0.992	0.126	0.127	0.016	0.938
				GRN	-0.956	0.125	0.126	0.016	0.937
				Naive	-73.393	0.020	0.033	0.651	0.000
				Complete	0.828	0.119	0.122	0.015	0.949
			0.30	RC	-50.254	0.091	0.093	0.314	0.006
				RSRC	-37.029	0.114	0.122	0.180	0.130
				GRRC	-1.001	0.128	0.128	0.016	0.936
				GRN	-0.956	0.126	0.128	0.016	0.935
				Naive	-74.831	0.020	0.034	0.677	0.000
				Complete	0.856	0.119	0.121	0.015	0.949

A.6. Misclassification table

Table A.13: Simulation results for $\beta_X = \log 1.5$ under misspecification and correlated, additive measurement error in the outcome and covariate X with normally distributed error, 75% censoring for the true event time, 90% sensitivity, and 90% specificity. 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.

β_X	σ^2_ν	σ^2_ϵ	$\sigma_{\nu,\epsilon}$	Method	$\%$ Bias	ASE	ESE	MSE	CP
log(1.5)				True	-0.099	0.055	0.054	0.003	0.953
	0.5	0.5	0.15	RC	-43.111	0.106	0.101	0.041	0.611
				RSRC	-40.842	0.118	0.117	0.041	0.681
				GRRC	-0.049	0.170	0.163	0.027	0.952
				GRN	0.641	0.172	0.164	0.027	0.954
				Naive	-141.097	0.042	0.042	0.329	0.000
				Complete	-0.025	0.177	0.178	0.032	0.953

A.7. VCCC eligibility criteria

We analyzed data on 4797 HIV-positive patients that established care at the VCCC between 1998 and 2013. For the virologic failure outcome, patients were excluded if they had an indeterminate ART start date, started ART prior to enrollment, had no CD4 count measurement between 180 days before or 30 days after starting ART, or had no follow-up after starting ART. Using the unvalidated data, 2143 patients met the criteria for inclusion, of which 1863 met the criteria using the validated data. These 1863 patients were used in all further analyses to ensure that any differences between estimators are not due to the differences in included patients. For the ADE outcome, the exclusion criteria was similar to that of the former analysis except we additionally excluded patients that had an ADE before ART initiation and those with indeterminate ADE dates. Using the unvalidated data, 1995 patients met the ADE analysis criteria, of which 1595 met the criteria using the validated data. Again, these 1595 were used in all further ADE analyses. Note that for both analyses, failures within 6 months of ART start were not considered a true failure due to the time required by the regimen to be efficacious. In addition, we made some further simplifying assumptions for the purpose of this data example for ease of exposition. Specifically, we removed subjects from the analyses that were not in both the unvalidated and validated datasets for ease of interpretation and selected validation subsets as if we did not validate all subjects. This was done to highlight the application of our methods and be able to effectively compare their relative performance.

Of the 1863 patients in the analysis of the virologic failure outcome, 20 were incorrectly classified as having failed, resulting in a 1% misclassification rate. There were 386 incorrectly recorded event

times, with the error having mean and standard deviation of -0.13 and 1.1 years, respectively. CD4 count at ART start was incorrect for 125 patients, with the error having mean and standard deviation of 21 and 164 cell/mm^3, respectively. The correlation between the error in the failure times and CD4 count at ART initiation for subjects with both types of error was -0.17.

Of the 1595 patients in the analysis of the ADE outcome, 161 were incorrectly classified as having had an ADE and 12 were incorrectly classified as having been censored, resulting in an appreciable misclassification rate of 11%. There were 551 incorrectly recorded event times, with the error having mean and standard deviation of -0.75 and 2.89 years, respectively. CD4 count at ART start was incorrect for 107 patients, with the error having mean and standard deviation of 10 and 154 cell/mm^3, respectively. The correlation between the error in the failure times and CD4 count at ART initiation for subjects with both types of error was -0.10.

A.8. VCCC tables

Table A.14: The hazard ratios (HR) 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. The CIs are calculated using the bootstrap for the RC, RSRC, GRRC, and GRN estimators.

Outcome	Method	$100\times$ CD4	$10\times$ Age
Time to virologic failure	True	0.902 (0.869, 0.935)	0.860 (0.806, 0.916)
	RC	0.920 (0.888, 0.953)	0.880 (0.825, 0.939)
	RSRC	0.918 (0.885, 0.953)	0.879 (0.821, 0.942)
	GRRC	0.918 (0.883, 0.954)	0.869 (0.811, 0.932)
	GRN	0.918 (0.882, 0.956)	0.869 (0.802, 0.942)
	Naive	0.918 (0.885, 0.953)	0.878 (0.824, 0.936)
	HT	0.929 (0.852, 1.012)	0.790 (0.679, 0.919)
Time to ADE	True	0.693 (0.593, 0.809)	0.829 (0.671, 1.023)
	RC	0.899 (0.832, 0.971)	1.071 (0.940, 1.221)
	RSRC	0.895 (0.827, 0.969)	1.073 (0.938, 1.226)
	GRRC	0.694 (0.565, 0.852)	0.883 (0.632, 1.234)
	GRN	0.693 (0.564, 0.853)	0.883 (0.622, 1.253)
	Naive	0.910 (0.841, 0.986)	1.087 (0.957, 1.235)
	HT	0.748 (0.597, 0.939)	1.114 (0.757, 1.640)

Table A.15: The mean of 10 hazard ratios (HR) from 10 different case-cohort sampled validation subsets for a 100 cell/mm^3 increase in CD4 count at ART initiation and 10 year increase in age at CD4 count measurement.

Outcome	Method	$100\times$ CD4	$10\times$ Age
Time to ADE	True	0.693	0.829
	RC	0.909	1.088
	RSRC	0.900	1.086
	GRRC	0.673	0.817
	GRN	0.673	0.817
	Naive	0.910	1.087
	HT	0.689	0.802

A.9. Example R code

The code below demonstrates how to implement the Regression Calibration and Generalized Raking Naive methods for example datasets. This example assumes the validation subset was selected as a simple random sample and that there are two covariates (X is error prone and Z is error free). Note that the code only demonstrates how to obtain estimates; standard errors must be calculated using the stratified bootstrap as described in the paper. Full code implementing all methods discussed in the paper (including standard errors) is available at https://github.com/ericoh17/RRCME.

```
library(dplyr)
library(survival)
library(survey)


# Example datasets
full_dat <- read.csv("example_dat.csv", row.names = 1)
valid_subset <- read.csv("example_valid_subset.csv", row.names = 1)


full_dat$time <- full_dat$delta <- full_dat$x <- NA


full_dat$time[full_dat$randomized == TRUE] <- valid_subset$time
full_dat$delta[full_dat$randomized == TRUE] <- valid_subset$delta
full_dat$x[full_dat$randomized == TRUE] <- valid_subset$x


### Regression Calibration ###
# Calibrate the covariate
```

```
x_calib_model <- lm(x ~ x_star + z, data = valid_subset)
x_hat <- predict(x_calib_model, data = full_dat)

# Calibrate the outcome
w_calib_model <- lm(total_y_err ~ x_star + z, data = valid_subset)
w_hat <- predict(w_calib_model, data = full_dat)
time_hat <- full_dat$time_star - w_hat

# Fit RC model
rc_mod <- coxph(Surv(time_hat, full_dat$delta_star) ~ x_hat + full_dat$z)

# Extract RC coefficients
beta_x_RC <- rc_mod$coef[1]
beta_z_RC <- rc_mod$coef[2]

### Generalized Raking Naive ###
# Fit naive model
naive_mod <- coxph(Surv(time_star, delta_star) ~ x_star + z, data = full_dat)

# Extract influence functions from naive model
IF_naive <- data.frame(resid(naive_mod, "dfbeta"))
colnames(IF_naive) <- paste("if", 1:2, sep = "")
full_IF_dat <- dplyr::bind_cols(full_dat, IF_naive)

# Calculate raking weights
IF_design <- twophase(id = list(~id, ~id), subset = ~randomized,
                      data = full_IF_dat)
IF_raking <- calibrate(IF_design, phase = 2, formula = ~if1+if2,
                       calfun = "raking")

# Fit raking model
```

```
raking_mod <- svycoxph(Surv(time, delta) ~ x + z, design = IF_raking)


# Extract raking coefficients
beta_x_GRN <- raking_mod$coef[1]
beta_z_GRN <- raking_mod$coef[2]
```

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Table B.1: The sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and negative predictive value (NPV) for the event indicator generated for error scenarios 1, 2, and 3 in the simple random sampling simulations.

β_z	% Cens	β_x	Sens	Spec	PPV	NPV
$\log(0.5)$	50	log(1.5)	0.465	0.947	0.878	0.684
		log(3)	0.479	0.948	0.893	0.669
	75	log(1.5)	0.672	0.905	0.693	0.897
		log(3)	0.705	0.889	0.659	0.908
	90	log(1.5)	0.822	0.820	0.330	0.977
		log(3)	0.819	0.796	0.294	0.977

Table B.2: Misclassification generation process for the sampling design comparison simulations. The sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and negative predictive value (NPV) for the event indicator are presented.

β_z	% Cens	β_x	$\Delta^\star$	Sens	Spec	PPV	NPV
$\log(0.5)$	90	log(1.5)	$\text{Bernoulli}(\text{expit}(-1 + 4 * \Delta + 0.5 * X - 0.5 * U - 0.5 * Z))$	0.718	0.961	0.665	0.969
		log(3)	$\text{Bernoulli}(\text{expit}(-1.5 + 4 * \Delta + 0.5 * X - 0.5 * U - 0.5 * Z))$	0.715	0.970	0.710	0.971

B.1. Multiple Imputation Details

We explicate the multiple imputation implementation details below for imputation models without interaction terms. Define

- $V_i = (1_i, \Delta_i^\star, X_i^\star, U_i^\star, Z_i)'$
- $V_{-U,i} = (1_i, \Delta_i^\star, X_i^\star, Z_i)'$
- $V_{\Delta,i}^{(l)} = (1_i, \Delta_i^\star, \hat{X}_i^{(l-1)}, \hat{U}_i^{(l-1)}, Z_i)$
- $V_{X,i}^{(l)} = (1_i, \hat{\Delta}_i^{(l)}, X_i^\star, \hat{U}_i^{(l-1)}, Z_i)$
- $V_{U,i}^{(l)} = (1_i, \hat{\Delta}_i^{(l)}, \hat{X}_i^{(l)}, Z_i)$,

and let the lower case versions denote their observed counterparts. MI with interaction terms follows exactly the same except the terms defined above contain all possible interaction terms.

Multiple imputation for Δ only

1. Fit the logistic regression model $\text{logit}(P(\Delta_i = 1)|V_i = v_i) = v_i'\eta$ using the validation subset to obtain $\hat{\eta}$. This corresponds to characterizing a posterior distribution for η given the phase two data under a non-informative prior distribution.
2. For $m = 1, \ldots, M$ iterations:
3. Generate $\eta_\star^{(m)} \sim \text{N}(\hat{\eta}, \tau_{\Delta,\star}^2 (V'V)^{-1})$, where $\tau_{\Delta,\star}^2 \sim \hat{\tau}_\Delta^2 \frac{n-p_\eta}{\chi^2_{n-p_\eta}}$, $\hat{\tau}_\Delta^2$ is the squared sum of the working residuals from the logistic regression model, and p_η is the dimension of η.
4. Sample and impute $\hat{\Delta}_i^{(m)} \sim \text{Bernoulli}(\text{expit}(v_i'\eta_\star^{(m)}))$ for all phase one subjects.
5. Stop after M iterations

Fully conditional specification multiple imputation

1. Fit the logistic regression model $\text{logit}(P(\Delta_i = 1)|V_i = v_i) = v_i'\eta_V$ and linear regression models $\text{E}(X_i|V_i = v_i) = v_i'\theta_V$ and $\text{E}(R_i|V_{-U,i} = v_{-U,i}) = v'_{-U,i}\omega_V$ using the validation subset to obtain $\hat{\eta}_V$, $\hat{\theta}_V$, and $\hat{\omega}_V$.
2. For $m = 1, \ldots, M$ iterations:
3. Generate

- $\eta_\star^{(0)} \sim \mathrm{N}(\hat{\eta}_V, \tau^2_{\Delta,V,\star}(V'V)^{-1})$
- $\theta_\star^{(0)} \sim \mathrm{N}(\hat{\theta}_V, \tau^2_{X,V,\star}(V'V)^{-1})$
- $\omega_\star^{(0)} \sim \mathrm{N}(\hat{\omega}_V, \tau^2_{U,V,\star}(V'_{-U}V_{-U})^{-1})$,

where $\tau^2_{\Delta,V,\star} \sim \hat{\tau}^2_{\Delta,V}\frac{n-p_{\eta_V}}{\chi^2_{n-p_{\eta_V}}}$, $\tau^2_{X,V,\star} \sim \hat{\tau}^2_{X,V}\frac{n-p_{\theta_V}}{\chi^2_{n-p_{\theta_V}}}$, $\tau^2_{U,V,\star} \sim \hat{\tau}^2_{U,V}\frac{n-p_{\omega_V}}{\chi^2_{n-p_{\omega_V}}}$, $\hat{\tau}^2_{\Delta,V}$, $\hat{\tau}^2_{X,V}$, and $\hat{\tau}^2_{U,V}$ are the squared sum of working residuals/residual sum of squares from their respective regression models, and p_{η_V}, p_{θ_V}, and p_{ω_V} are the dimensions of their respective parameters.

4. Sample and impute $\hat{\Delta}_i^{(0)} \sim \mathrm{Bernoulli}(\mathrm{expit}(v_i'\eta_\star^{(0)}))$ and $\hat{X}_i^{(0)} \sim \mathrm{N}(v_i'\theta_\star^{(0)}, \tau^2_{X,V,\star})$ for all phase one subjects. Sample $\hat{R}_i^{(0)} \sim \mathrm{N}(v'_{-U,i}\omega_\star^{(0)}, \tau^2_{U,V,\star})$ and impute $\hat{U}_i^{(0)} = U_i^\star - \hat{R}_i^{(0)}$ for all phase one subjects.
5. For $l = 1, \ldots, L$ iterations:
6. Fit the logistic regression model $\mathrm{logit}(P(\Delta_i = 1)|V^{(l)}_{\Delta,i} = v^{(l)}_{\Delta,i}) = v'^{(l)}_{\Delta,i}\eta$ on the validation subset to obtain $\hat{\eta}^{(l)}$.
7. Generate $\eta_\star^{(l)} \sim \mathrm{N}(\hat{\eta}^{(l)}, \tau^2_{\Delta,\star}(V_\Delta^{(l)'}V_\Delta^{(l)})^{-1})$ where $\tau^2_{\Delta,\star} \sim \hat{\tau}^2_\Delta\frac{n-p_\eta}{\chi^2_{n-p_\eta}}$.
8. Sample and impute $\hat{\Delta}_i^{(l)} \sim \mathrm{Bernoulli}(\mathrm{expit}(v_{\Delta,i}^{(l)'}\eta_\star^{(l)}))$ for all phase one subjects.
9. Fit the linear regression model $\mathrm{E}(X_i|V^{(l)}_{X,i} = v^{(l)}_{X,i}) = v'^{(l)}_{X,i}\theta$ on the validation subset to obtain $\hat{\theta}^{(l)}$.
10. Generate $\theta_\star^{(l)} \sim \mathrm{N}(\hat{\theta}^{(l)}, \tau^2_{X,\star}(V_X^{(l)'}V_X^{(l)})^{-1})$ where $\tau^2_{X,\star} \sim \hat{\tau}^2_X\frac{n-p_\theta}{\chi^2_{n-p_\theta}}$.
11. Sample and impute $\hat{X}_i^{(l)} \sim \mathrm{N}(v_{X,i}^{(l)'}\theta_\star^{(l)}, \tau^2_{X,\star})$ for all phase one subjects.
12. Fit the linear regression model $\mathrm{E}(R_i|V^{(l)}_{U,i} = v^{(l)}_{U,i}) = v'^{(l)}_{U,i}\omega$ on the validation subset to obtain $\hat{\omega}^{(l)}$.
13. Generate $\omega_\star^{(l)} \sim \mathrm{N}(\hat{\omega}^{(l)}, \tau^2_{U,\star}(V_U^{(l)'}V_U^{(l)})^{-1})$ where $\tau^2_{U,\star} \sim \hat{\tau}^2_U\frac{n-p_\omega}{\chi^2_{n-p_\omega}}$.
14. Sample $\hat{R}_i^{(l)} \sim \mathrm{N}(v_{U,i}^{(l)'}\omega_\star^{(l)}, \tau^2_{U,\star})$ and impute $\hat{U}_i^{(l)} = U_i^\star - \hat{R}_i^{(l)}$ for all phase one subjects.
15. Stop after L iterations

16. Stop after M iterations

Table B.3: Simulation results for estimating β_x using the data imputation approach for error scenario 2 (errors in event indicator and failure time) with $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	0.07661	0.039569	2.322461	0.039418	0.001566	0.95
			HT	0.862615	0.091897	1	0.087962	0.008457	0.937
			GRN	0.858763	0.076007	1.209062	0.072554	0.005789	0.946
			GRMIS	1.031894	0.06415	1.432527	0.063852	0.004133	0.948
			GRMIC	0.855469	0.064549	1.423681	0.063311	0.004179	0.948
			GRFCSMIS	0.85294	0.063187	1.454371	0.063714	0.004005	0.95
			GRFCSMIC	0.876251	0.064668	1.421072	0.063193	0.004195	0.947
		log(3)	True	-0.00837	0.041674	2.365635	0.04412	0.001737	0.951
			HT	0.44906	0.098585	1	0.097536	0.009743	0.942
			GRN	0.21076	0.081199	1.214119	0.080367	0.006599	0.944
			GRMIS	0.017758	0.070751	1.393417	0.070419	0.005006	0.944
			GRMIC	0.054635	0.070446	1.399443	0.069713	0.004963	0.946
			GRFCSMIS	0.097153	0.068036	1.449007	0.070358	0.00463	0.948
			GRFCSMIC	-0.0385	0.0696	1.416459	0.069488	0.004844	0.944
	75	log(1.5)	True	-0.11172	0.050646	2.511236	0.053272	0.002565	0.946
			HT	1.592927	0.127184	1	0.118928	0.016218	0.938
			GRN	0.404828	0.099182	1.282334	0.096779	0.00984	0.946
			GRMIS	0.15642	0.091196	1.394624	0.091856	0.008317	0.941
			GRMIC	-0.60126	0.093375	1.362087	0.090987	0.008725	0.945
			GRFCSMIS	-0.27014	0.091629	1.388037	0.091734	0.008397	0.942
			GRFCSMIC	-0.9402	0.090891	1.399307	0.090986	0.008276	0.945
		log(3)	True	-0.01819	0.05804	2.371545	0.05929	0.003369	0.948
			HT	0.563709	0.137646	1	0.131856	0.018985	0.938
			GRN	0.513363	0.122738	1.121461	0.114183	0.015096	0.938
			GRMIS	0.130314	0.111008	1.239958	0.103685	0.012325	0.946
			GRMIC	-0.18315	0.110439	1.246354	0.10261	0.012201	0.941
			GRFCSMIS	-0.0051	0.109314	1.259179	0.103718	0.01195	0.942
			GRFCSMIC	-0.31075	0.109638	1.255454	0.102901	0.012032	0.939
	90	log(1.5)	True	0.0138	0.084364	2.231975	0.083155	0.007117	0.947
			HT	1.89313	0.188298	1	0.184107	0.035515	0.944
			GRN	0.747537	0.167743	1.122541	0.166189	0.028147	0.94
			GRMIS	0.392568	0.160833	1.170769	0.159171	0.02587	0.929
			GRMIC	0.50219	0.162774	1.156806	0.156793	0.0265	0.928
			GRFCSMIS	0.250184	0.160999	1.169562	0.159623	0.025922	0.933
			GRFCSMIC	0.860638	0.163186	1.153889	0.157477	0.026642	0.93
		log(3)	True	-0.04654	0.088525	2.286886	0.089229	0.007837	0.95
			HT	1.420837	0.202447	1	0.199373	0.041229	0.944
			GRN	1.282639	0.188611	1.073361	0.188717	0.035773	0.946
			GRMIS	0.11986	0.177052	1.143433	0.175632	0.031349	0.944
			GRMIC	-0.27151	0.175575	1.153052	0.173819	0.030836	0.938
			GRFCSMIS	0.852398	0.177915	1.137887	0.176882	0.031741	0.945
			GRFCSMIC	-0.01923	0.176458	1.147285	0.173343	0.031137	0.935

Table B.4: Simulation results for estimating β_x using the data imputation approach for error scenario 1 (error only in event indicator) with $N = 10000$, $n = 2000$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	0.054808	0.017284	2.362416	0.017613	0.000299	0.951
			HT	-0.02749	0.040833	1	0.039395	0.001667	0.958
			GRN	-0.17062	0.032145	1.270254	0.032467	0.001034	0.946
			GRMIS	0.11896	0.028221	1.446867	0.02839	0.000797	0.952
			GRMIC	0.171039	0.028357	1.43994	0.02831	0.000805	0.95
		log(3)	True	-0.05218	0.020492	2.155264	0.019701	0.00042	0.942
			HT	-0.06065	0.044165	1	0.043853	0.001951	0.946
			GRN	-0.09902	0.038639	1.143009	0.035966	0.001494	0.945
			GRMIS	-0.08065	0.031253	1.413158	0.030816	0.000978	0.946
			GRMIC	-0.13329	0.030914	1.428659	0.030735	0.000958	0.948
	75	log(1.5)	True	-0.34616	0.023745	2.238788	0.023804	0.000566	0.956
			HT	-0.29482	0.053159	1	0.053226	0.002827	0.944
			GRN	-0.23871	0.043041	1.235098	0.043158	0.001853	0.943
			GRMIS	-0.30717	0.042088	1.263042	0.041024	0.001773	0.954
			GRMIC	-0.11051	0.043443	1.223657	0.040882	0.001888	0.949
		log(3)	True	-0.03541	0.027097	2.085792	0.026437	0.000734	0.942
			HT	0.006602	0.056519	1	0.058911	0.003194	0.948
			GRN	-0.03041	0.052345	1.079745	0.050762	0.00274	0.948
			GRMIS	-0.14913	0.045914	1.230966	0.045108	0.002111	0.95
			GRMIC	-0.17122	0.045346	1.246394	0.044947	0.00206	0.948
	90	log(1.5)	True	0.300231	0.038368	2.189938	0.037121	0.001474	0.946
			HT	-0.27875	0.084024	1	0.082716	0.007061	0.95
			GRN	-0.15046	0.076347	1.100555	0.074226	0.005829	0.951
			GRMIS	-0.12485	0.07321	1.147715	0.070996	0.00536	0.946
			GRMIC	0.114554	0.072765	1.154728	0.070949	0.005295	0.946
		log(3)	True	-0.10845	0.040031	2.20517	0.039781	0.001604	0.948
			HT	-0.03262	0.088274	1	0.08865	0.007792	0.952
			GRN	-0.06333	0.085414	1.033488	0.083241	0.007296	0.953
			GRMIS	-0.05366	0.076953	1.147114	0.07441	0.005922	0.954
			GRMIC	-0.1297	0.076059	1.160603	0.074555	0.005787	0.948

Table B.5: Simulation results for estimating β_x using the data imputation approach for error scenario 2 (errors in event indicator and failure time) with $N = 10000$, $n = 2000$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.19065	0.017589	2.302689	0.017612	0.00031	0.948
			HT	0.56596	0.040501	1	0.039375	0.001646	0.949
			GRN	-0.00021	0.033095	1.223769	0.032559	0.001095	0.958
			GRMIS	-0.09951	0.028698	1.411272	0.028497	0.000824	0.952
			GRMIC	0.057628	0.028235	1.434443	0.028443	0.000797	0.95
			GRFCSMIS	-0.16005	0.028514	1.420372	0.028459	0.000813	0.954
			GRFCSMIC	-0.09481	0.028744	1.40905	0.028368	0.000826	0.953
		log(3)	True	0.007371	0.020095	2.216554	0.019705	0.000404	0.954
			HT	0.14059	0.044541	1	0.043984	0.001986	0.958
			GRN	0.120252	0.037165	1.198481	0.036194	0.001383	0.953
			GRMIS	-0.00645	0.032569	1.367607	0.031431	0.001061	0.95
			GRMIC	0.003242	0.032407	1.374455	0.031323	0.00105	0.954
			GRFCSMIS	-0.10256	0.031067	1.433718	0.031312	0.000966	0.958
			GRFCSMIC	0.021591	0.031842	1.398842	0.031165	0.001014	0.95
	75	log(1.5)	True	0.014001	0.025018	2.11733	0.023801	0.000626	0.949
			HT	0.244099	0.052971	1	0.053189	0.002807	0.951
			GRN	0.107019	0.043513	1.217349	0.043265	0.001894	0.951
			GRMIS	0.342023	0.042648	1.242038	0.041066	0.001821	0.946
			GRMIC	0.264359	0.041911	1.263871	0.040974	0.001758	0.944
			GRFCSMIS	0.310782	0.043342	1.22217	0.041012	0.00188	0.948
			GRFCSMIC	0.286289	0.042452	1.24779	0.040907	0.001803	0.944
		log(3)	True	-0.05745	0.028097	2.082046	0.02643	0.00079	0.946
			HT	0.080356	0.0585	1	0.059039	0.003423	0.958
			GRN	-0.01744	0.050943	1.148345	0.051062	0.002595	0.954
			GRMIS	-0.10403	0.047242	1.238287	0.04617	0.002233	0.95
			GRMIC	-0.09856	0.047631	1.228176	0.046065	0.00227	0.954
			GRFCSMIS	-0.13389	0.046445	1.259538	0.046111	0.002159	0.954
			GRFCSMIC	-0.11486	0.04664	1.254269	0.045871	0.002177	0.95
	90	log(1.5)	True	0.300231	0.038368	2.093359	0.037121	0.001474	0.946
			HT	0.614175	0.080318	1	0.082808	0.006457	0.944
			GRN	0.516905	0.075704	1.060956	0.074574	0.005735	0.947
			GRMIS	0.190219	0.071749	1.119431	0.071582	0.005149	0.941
			GRMIC	0.410931	0.071522	1.122986	0.071456	0.005118	0.942
			GRFCSMIS	0.539946	0.070147	1.145004	0.071652	0.004925	0.942
			GRFCSMIC	0.64252	0.070001	1.147396	0.071463	0.004907	0.942
		log(3)	True	-0.10845	0.040031	2.109783	0.039781	0.001604	0.948
			HT	0.207349	0.084456	1	0.088877	0.007138	0.949
			GRN	0.114892	0.079391	1.063797	0.083919	0.006304	0.942
			GRMIS	-0.15029	0.077382	1.091408	0.07822	0.005991	0.948
			GRMIC	-0.2245	0.075666	1.116166	0.077916	0.005731	0.948
			GRFCSMIS	-0.14806	0.075271	1.122027	0.077797	0.005668	0.948
			GRFCSMIC	-0.29095	0.074947	1.126878	0.077503	0.005627	0.946

Table B.6: Simulation results for estimating β_x using the data imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 10000$, $n = 2000$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.19065	0.017589	2.305954	0.017612	0.00031	0.948
			HT	-0.38798	0.040559	1	0.039363	0.001647	0.947
			GRN	-0.0649	0.040211	1.008631	0.039278	0.001617	0.949
			GRMIS	-0.40733	0.040069	1.01222	0.039306	0.001608	0.943
			GRMIC	-0.42348	0.040052	1.012644	0.039306	0.001607	0.944
			GRFCSMIS	-0.19395	0.032359	1.253383	0.033263	0.001048	0.953
			GRFCSMIC	-0.20848	0.032963	1.230438	0.03321	0.001087	0.95
		log(3)	True	0.007371	0.020095	2.176311	0.019705	0.000404	0.954
			HT	0.152582	0.043733	1	0.043902	0.001915	0.944
			GRN	0.182688	0.04374	0.999833	0.043789	0.001917	0.938
			GRMIS	0.140377	0.043289	1.010252	0.043781	0.001876	0.944
			GRMIC	0.13122	0.043248	1.01122	0.043759	0.001872	0.945
			GRFCSMIS	0.08376	0.039471	1.107966	0.039375	0.001559	0.946
			GRFCSMIC	0.082497	0.03993	1.095225	0.03949	0.001595	0.944
	75	log(1.5)	True	0.014001	0.025018	2.070413	0.023801	0.000626	0.949
			HT	-0.42736	0.051797	1	0.053204	0.002686	0.953
			GRN	-0.08905	0.050398	1.027756	0.052399	0.00254	0.952
			GRMIS	-0.19243	0.051672	1.002418	0.052889	0.002671	0.95
			GRMIC	-0.14175	0.051877	0.998465	0.052852	0.002692	0.95
			GRFCSMIS	-0.1983	0.046177	1.121696	0.046052	0.002133	0.946
			GRFCSMIC	0.056028	0.045931	1.127705	0.045945	0.00211	0.942
		log(3)	True	-0.05745	0.028097	2.160778	0.02643	0.00079	0.946
			HT	0.257044	0.060712	1	0.058916	0.003694	0.94
			GRN	0.304253	0.060606	1.001741	0.058098	0.003684	0.939
			GRMIS	0.304018	0.061782	0.982683	0.058593	0.003828	0.937
			GRMIC	0.279433	0.061726	0.983575	0.058552	0.003819	0.938
			GRFCSMIS	0.247202	0.055262	1.098615	0.05378	0.003061	0.947
			GRFCSMIC	0.162435	0.054445	1.115095	0.053735	0.002967	0.942
	90	log(1.5)	True	0.300231	0.038368	2.135321	0.037121	0.001474	0.946
			HT	-0.13418	0.081929	1	0.082664	0.006713	0.944
			GRN	0.403334	0.081083	1.010423	0.079681	0.006577	0.949
			GRMIS	0.539478	0.081062	1.010686	0.079814	0.006576	0.948
			GRMIC	0.397156	0.08106	1.010708	0.079601	0.006573	0.949
			GRFCSMIS	0.113621	0.078127	1.048663	0.076604	0.006104	0.948
			GRFCSMIC	0.324641	0.076865	1.065877	0.076394	0.00591	0.947
		log(3)	True	-0.10845	0.040031	2.365923	0.039781	0.001604	0.948
			HT	0.106186	0.094709	1	0.08875	0.008971	0.944
			GRN	0.368331	0.090655	1.044722	0.086697	0.008235	0.946
			GRMIS	0.397741	0.089394	1.059461	0.086097	0.00801	0.944
			GRMIC	0.384319	0.091384	1.036391	0.085945	0.008369	0.944
			GRFCSMIS	0.205223	0.087776	1.078985	0.084764	0.00771	0.943
			GRFCSMIC	0.16134	0.090074	1.051463	0.084645	0.008116	0.946

Table B.7: Type 1 error results for $\beta_x = 0$ using the data imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 10000$, $n = 2000$, and simple random sampling. The absolute bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error (MSE), and type 1 error are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	Bias	ESE	RE	ASE	MSE	Type 1 error
log(0.5)	50	0	True	0.00091	0.04430	2.19292	0.04309	0.00196	0.052
			HT	0.00171	0.09715	1.00000	0.09609	0.00944	0.052
			GRN	0.00282	0.09357	1.03825	0.09300	0.00876	0.052
			GRMIS	0.00557	0.09364	1.03750	0.09291	0.00880	0.052
			GRMIC	0.00609	0.09248	1.05050	0.09270	0.00859	0.055
			GRFCSMIS	0.00383	0.08964	1.08383	0.09020	0.00805	0.052
			GRFCSMIC	0.00468	0.09041	1.07457	0.09000	0.00820	0.056
	75	0	True	-0.00268	0.06448	2.42275	0.06675	0.00417	0.044
			HT	-0.00357	0.15623	1.00000	0.14633	0.02442	0.057
			GRN	0.00064	0.15408	1.01394	0.14484	0.02374	0.06
			GRMIS	-0.00021	0.15210	1.02713	0.14405	0.02314	0.06
			GRMIC	-0.00194	0.15477	1.00944	0.14364	0.02396	0.064
			GRFCSMIS	-0.00429	0.15389	1.01521	0.14389	0.02370	0.057
			GRFCSMIC	-0.00005	0.14833	1.05325	0.14309	0.02200	0.059
	90	0	True	0.00094	0.11292	2.24950	0.11122	0.01275	0.057
			HT	0.00193	0.25401	1.00000	0.23831	0.06453	0.066
			GRN	-0.00086	0.25437	0.99859	0.23712	0.06470	0.069
			GRMIS	0.00193	0.25698	0.98846	0.23662	0.06604	0.068
			GRMIC	-0.00063	0.26162	0.97090	0.23583	0.06845	0.07
			GRFCSMIS	0.00212	0.25010	1.01563	0.23629	0.06256	0.071
			GRFCSMIC	-0.00458	0.24366	1.04246	0.23422	0.05939	0.072

Table B.8: Simulation results for estimating β_x using the IF imputation approach for error scenario 1 (error only in event indicator) with $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.03595	0.039644	2.289193	0.039422	0.001572	0.956
			HT	1.228958	0.090753	1	0.087874	0.008261	0.949
			GRN	1.40684	0.07401	1.226214	0.072528	0.00551	0.95
			GRMIS	-0.90195	0.064965	1.396945	0.064525	0.004234	0.946
			GRMIC	-1.01073	0.065105	1.393946	0.064286	0.004255	0.95
		log(3)	True	0.041168	0.041582	2.453957	0.04415	0.001729	0.948
			HT	0.631089	0.10204	1	0.097775	0.01046	0.939
			GRN	0.282312	0.082568	1.235824	0.080447	0.006827	0.942
			GRMIS	-0.22609	0.07478	1.364544	0.071255	0.005598	0.952
			GRMIC	-0.22782	0.073727	1.384023	0.070891	0.005442	0.954
	75	log(1.5)	True	0.119394	0.051672	2.266392	0.053276	0.00267	0.954
			HT	0.781363	0.117109	1	0.118644	0.013725	0.952
			GRN	0.916624	0.097339	1.203106	0.096548	0.009489	0.945
			GRMIS	-0.96486	0.094689	1.236776	0.090898	0.008981	0.94
			GRMIC	-0.55219	0.095677	1.224011	0.090511	0.009159	0.94
		log(3)	True	-0.01311	0.06088	2.241353	0.059211	0.003706	0.949
			HT	1.034735	0.136454	1	0.131041	0.018749	0.938
			GRN	0.386125	0.119288	1.143905	0.113786	0.014248	0.934
			GRMIS	-0.24879	0.104165	1.309981	0.102151	0.010858	0.945
			GRMIC	-0.1159	0.102101	1.336464	0.101521	0.010426	0.942
	90	log(1.5)	True	0.0138	0.084364	2.222885	0.083155	0.007117	0.947
			HT	1.805251	0.187531	1	0.184444	0.035222	0.943
			GRN	0.30929	0.167181	1.121725	0.165789	0.027951	0.94
			GRMIS	-2.53059	0.16078	1.166381	0.154901	0.025956	0.942
			GRMIC	-1.37636	0.161713	1.159658	0.152769	0.026182	0.933
		log(3)	True	-0.04654	0.088525	2.315872	0.089229	0.007837	0.95
			HT	1.160558	0.205013	1	0.197598	0.042193	0.938
			GRN	0.945284	0.194363	1.054793	0.187058	0.037885	0.941
			GRMIS	-1.02709	0.182758	1.121773	0.164852	0.033528	0.931
			GRMIC	-0.94924	0.178754	1.146899	0.163399	0.032062	0.924

Table B.9: Simulation results for estimating β_x using the IF imputation approach for error scenario 2 (errors in event indicator and failure time) with $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.03595	0.039644	2.259058	0.039422	0.001572	0.956
			HT	1.193166	0.089558	1	0.088263	0.008044	0.944
			GRN	0.36901	0.073756	1.21425	0.072739	0.005442	0.94
			GRMIS	-0.66916	0.067803	1.320857	0.064383	0.004605	0.95
			GRMIC	-1.17444	0.067482	1.327148	0.064126	0.004576	0.947
			GRFCSMIS	-0.6887	0.066363	1.349513	0.064489	0.004412	0.948
			GRFCSMIC	-1.04496	0.066929	1.338113	0.064188	0.004497	0.948
		log(3)	True	0.041168	0.041582	2.387503	0.04415	0.001729	0.948
			HT	0.159486	0.099277	1	0.097769	0.009859	0.942
			GRN	-0.05843	0.08363	1.187094	0.080697	0.006994	0.94
			GRMIS	-0.33902	0.074376	1.334794	0.072157	0.005546	0.942
			GRMIC	-0.33888	0.074533	1.331979	0.071715	0.005569	0.942
			GRFCSMIS	-0.33134	0.074262	1.336838	0.071762	0.005528	0.94
			GRFCSMIC	-0.41767	0.072415	1.370951	0.071454	0.005265	0.941
	75	log(1.5)	True	0.119394	0.051672	2.277078	0.053276	0.00267	0.954
			HT	0.106275	0.117662	1	0.118679	0.013844	0.946
			GRN	0.073904	0.099255	1.185442	0.096827	0.009852	0.938
			GRMIS	-1.39746	0.095574	1.2311	0.091069	0.009167	0.934
			GRMIC	-2.02299	0.098117	1.199203	0.090413	0.009694	0.928
			GRFCSMIS	-1.13488	0.095578	1.231054	0.09105	0.009156	0.938
			GRFCSMIC	-1.8046	0.096456	1.21985	0.090576	0.009357	0.929
		log(3)	True	-0.01311	0.06088	2.182885	0.059211	0.003706	0.949
			HT	0.358656	0.132894	1	0.131964	0.017676	0.946
			GRN	-0.08903	0.115453	1.151068	0.114087	0.01333	0.939
			GRMIS	-1.04825	0.101828	1.305086	0.103487	0.010502	0.944
			GRMIC	-1.14319	0.103824	1.279992	0.103089	0.010937	0.948
			GRFCSMIS	-1.18049	0.100957	1.31635	0.102971	0.01036	0.943
			GRFCSMIC	-1.16145	0.101611	1.307871	0.102171	0.010488	0.944
	90	log(1.5)	True	0.0138	0.084364	2.241325	0.083155	0.007117	0.947
			HT	-0.12238	0.189087	1	0.184527	0.035754	0.94
			GRN	-0.47182	0.167907	1.126137	0.166491	0.028197	0.937
			GRMIS	-4.90776	0.165713	1.141049	0.155956	0.027857	0.925
			GRMIC	-3.43053	0.166734	1.134061	0.153326	0.027994	0.923
			GRFCSMIS	-4.20924	0.162172	1.165962	0.154132	0.026591	0.931
			GRFCSMIC	-2.69717	0.166436	1.136093	0.152002	0.027821	0.927
		log(3)	True	-0.04654	0.088525	2.307755	0.089229	0.007837	0.95
			HT	1.212813	0.204295	1	0.200054	0.041914	0.946
			GRN	1.110413	0.195368	1.045689	0.188311	0.038318	0.942
			GRMIS	-1.13955	0.177245	1.152611	0.171462	0.031573	0.929
			GRMIC	-1.08415	0.179691	1.13692	0.16897	0.032431	0.923
			GRFCSMIS	-1.55686	0.172123	1.186911	0.168801	0.029919	0.928
			GRFCSMIC	-0.90004	0.176818	1.155395	0.166461	0.031362	0.926

Table B.10: Simulation results for estimating β_x using the IF imputation approach for error scenario 1 (error only in event indicator) with $N = 10000$, $n = 2000$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	0.054808	0.017284	2.362416	0.017613	0.000299	0.951
			HT	-0.02749	0.040833	1	0.039395	0.001667	0.958
			GRN	-0.17062	0.032145	1.270254	0.032467	0.001034	0.946
			GRMIS	0.100479	0.029726	1.373624	0.029255	0.000884	0.949
			GRMIC	0.088663	0.029435	1.387224	0.02926	0.000867	0.95
		log(3)	True	-0.05218	0.020492	2.155264	0.019701	0.00042	0.942
			HT	-0.06065	0.044165	1	0.043853	0.001951	0.946
			GRN	-0.09902	0.038639	1.143009	0.035966	0.001494	0.945
			GRMIS	-0.09059	0.032648	1.352761	0.032008	0.001067	0.952
			GRMIC	-0.0665	0.032999	1.338383	0.032021	0.001089	0.95
	75	log(1.5)	True	-0.34616	0.023745	2.238788	0.023804	0.000566	0.956
			HT	-0.29482	0.053159	1	0.053226	0.002827	0.944
			GRN	-0.23871	0.043041	1.235098	0.043158	0.001853	0.943
			GRMIS	-0.32534	0.043819	1.213156	0.041613	0.001922	0.95
			GRMIC	-0.46772	0.04413	1.204598	0.041607	0.001951	0.948
		log(3)	True	-0.03541	0.027097	2.085792	0.026437	0.000734	0.942
			HT	0.006602	0.056519	1	0.058911	0.003194	0.948
			GRN	-0.03041	0.052345	1.079745	0.050762	0.00274	0.948
			GRMIS	-0.28792	0.047163	1.198373	0.046367	0.002234	0.948
			GRMIC	-0.23723	0.046623	1.212257	0.04631	0.00218	0.948
	90	log(1.5)	True	0.300231	0.038368	2.189938	0.037121	0.001474	0.946
			HT	-0.27875	0.084024	1	0.082716	0.007061	0.95
			GRN	-0.15046	0.076347	1.100555	0.074226	0.005829	0.951
			GRMIS	-0.38222	0.075817	1.108242	0.072511	0.005751	0.947
			GRMIC	-0.47893	0.074779	1.123631	0.072106	0.005596	0.946
		log(3)	True	-0.10845	0.040031	2.20517	0.039781	0.001604	0.948
			HT	-0.03262	0.088274	1	0.08865	0.007792	0.952
			GRN	-0.06333	0.085414	1.033488	0.083241	0.007296	0.953
			GRMIS	-0.1807	0.078796	1.120288	0.07501	0.006213	0.947
			GRMIC	-0.13735	0.078226	1.128451	0.074792	0.006122	0.948

Table B.11: Simulation results for estimating β_x using the IF imputation approach for error scenario 2 (errors in event indicator and failure time) with $N = 10000$, $n = 2000$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	0.054808	0.017284	2.34789	0.017613	0.000299	0.951
			HT	0.439737	0.040582	1	0.03938	0.00165	0.953
			GRN	-0.05916	0.033695	1.204375	0.032522	0.001135	0.955
			GRMIS	0.002278	0.030202	1.343668	0.029332	0.000912	0.953
			GRMIC	0.040389	0.030626	1.325078	0.029336	0.000938	0.951
			GRFCSMIS	0.028835	0.030745	1.319925	0.029295	0.000945	0.955
			GRFCSMIC	0.001368	0.030748	1.319817	0.029306	0.000945	0.956
		log(3)	True	-0.05218	0.020492	2.161159	0.019701	0.00042	0.942
			HT	0.061535	0.044286	1	0.043865	0.001962	0.951
			GRN	-0.14009	0.035509	1.247183	0.036138	0.001263	0.95
			GRMIS	-0.17072	0.031834	1.391139	0.032509	0.001017	0.95
			GRMIC	-0.20784	0.031965	1.385436	0.032543	0.001027	0.95
			GRFCSMIS	-0.26753	0.031827	1.391444	0.032351	0.001022	0.945
			GRFCSMIC	-0.24417	0.031706	1.396765	0.032372	0.001012	0.946
	75	log(1.5)	True	-0.34616	0.023745	2.312845	0.023804	0.000566	0.956
			HT	0.621078	0.054918	1	0.053213	0.003022	0.949
			GRN	-0.17351	0.043689	1.257005	0.043311	0.001909	0.943
			GRMIS	-0.27204	0.041358	1.327858	0.04161	0.001712	0.95
			GRMIC	-0.32999	0.040478	1.356729	0.041675	0.00164	0.949
			GRFCSMIS	-0.238	0.041502	1.323273	0.041606	0.001723	0.952
			GRFCSMIC	-0.35271	0.040672	1.350258	0.041595	0.001656	0.951
		log(3)	True	-0.03541	0.027097	2.136482	0.026437	0.000734	0.942
			HT	0.174341	0.057892	1	0.059047	0.003355	0.948
			GRN	0.101953	0.050843	1.138661	0.051027	0.002586	0.936
			GRMIS	-0.26643	0.046069	1.256639	0.047328	0.002131	0.949
			GRMIC	-0.23906	0.046182	1.253583	0.047276	0.00214	0.952
			GRFCSMIS	-0.38985	0.0466	1.242337	0.047083	0.00219	0.946
			GRFCSMIC	-0.43044	0.046508	1.244774	0.046946	0.002185	0.95
	90	log(1.5)	True	0.300231	0.038368	2.138466	0.037121	0.001474	0.946
			HT	0.681825	0.082049	1	0.082724	0.00674	0.945
			GRN	-0.02971	0.075265	1.090132	0.074518	0.005665	0.942
			GRMIS	-0.51917	0.072877	1.125852	0.073122	0.005316	0.948
			GRMIC	-0.53355	0.073562	1.115369	0.072863	0.005416	0.948
			GRFCSMIS	-0.5727	0.072729	1.128148	0.07256	0.005295	0.95
			GRFCSMIC	-0.72049	0.073653	1.114003	0.072181	0.005433	0.948
		log(3)	True	-0.10845	0.040031	2.21953	0.039781	0.001604	0.948
			HT	0.172047	0.088849	1	0.088705	0.007898	0.946
			GRN	-0.0031	0.082597	1.075695	0.08387	0.006822	0.94
			GRMIS	-0.57558	0.078538	1.131282	0.07818	0.006208	0.945
			GRMIC	-0.62865	0.077182	1.151156	0.078037	0.006005	0.945
			GRFCSMIS	-0.68889	0.077174	1.151274	0.077335	0.006013	0.942
			GRFCSMIC	-0.7213	0.07917	1.122251	0.077035	0.006331	0.941

Table B.12: Simulation results for estimating β_x using the IF imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 10000$, $n = 2000$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	0.054808	0.017284	2.283291	0.017613	0.000299	0.951
			HT	-0.00955	0.039465	1	0.0394	0.001557	0.952
			GRN	0.521031	0.039278	1.004761	0.039304	0.001547	0.95
			GRMIS	-0.23187	0.039863	0.990027	0.038274	0.00159	0.939
			GRMIC	-0.34967	0.03956	0.997613	0.038145	0.001567	0.938
			GRFCSMIS	0.158378	0.034421	1.146544	0.033399	0.001185	0.944
			GRFCSMIC	0.04174	0.033688	1.171498	0.033346	0.001135	0.94
		log(3)	True	-0.05218	0.020492	2.194225	0.019701	0.00042	0.942
			HT	0.033411	0.044964	1	0.043911	0.002022	0.95
			GRN	0.10533	0.044923	1.000899	0.043838	0.002019	0.953
			GRMIS	0.104264	0.043356	1.037085	0.042279	0.001881	0.946
			GRMIC	0.101906	0.043544	1.032593	0.042242	0.001897	0.946
			GRFCSMIS	-0.08375	0.041255	1.089891	0.040115	0.001703	0.948
			GRFCSMIC	-0.09068	0.04112	1.093472	0.040135	0.001692	0.945
	75	log(1.5)	True	-0.34616	0.023745	2.223618	0.023804	0.000566	0.956
			HT	-0.14543	0.052799	1	0.053227	0.002788	0.952
			GRN	0.48857	0.052468	1.006304	0.052421	0.002757	0.95
			GRMIS	1.162651	0.051602	1.023206	0.050093	0.002685	0.948
			GRMIC	1.205129	0.052288	1.009773	0.049919	0.002758	0.946
			GRFCSMIS	0.014211	0.044344	1.19068	0.046387	0.001966	0.955
			GRFCSMIC	-0.05963	0.043994	1.200136	0.046299	0.001936	0.954
		log(3)	True	-0.03541	0.027097	2.195267	0.026437	0.000734	0.942
			HT	0.210299	0.059485	1	0.058999	0.003544	0.955
			GRN	0.343256	0.058068	1.024415	0.058181	0.003386	0.952
			GRMIS	0.179398	0.054387	1.093735	0.05498	0.002962	0.948
			GRMIC	0.178737	0.055385	1.074034	0.05504	0.003071	0.944
			GRFCSMIS	0.118899	0.053781	1.106067	0.053714	0.002894	0.947
			GRFCSMIC	0.082925	0.054461	1.092252	0.053663	0.002967	0.944
	90	log(1.5)	True	0.300231	0.038368	2.272879	0.037121	0.001474	0.946
			HT	0.093613	0.087206	1	0.08286	0.007605	0.95
			GRN	0.390952	0.083979	1.038428	0.079908	0.007055	0.95
			GRMIS	1.612847	0.083257	1.047438	0.077403	0.006974	0.94
			GRMIC	1.605483	0.083066	1.049849	0.077219	0.006942	0.938
			GRFCSMIS	1.232738	0.078398	1.11235	0.076839	0.006171	0.946
			GRFCSMIC	0.715285	0.08002	1.089812	0.076573	0.006412	0.944
		log(3)	True	-0.10845	0.040031	2.176706	0.039781	0.001604	0.948
			HT	-0.21345	0.087135	1	0.088947	0.007598	0.956
			GRN	-0.02096	0.086341	1.009188	0.086725	0.007455	0.954
			GRMIS	0.446222	0.087491	0.995923	0.085366	0.007679	0.948
			GRMIC	0.344767	0.088426	0.985395	0.085189	0.007834	0.948
			GRFCSMIS	-0.18669	0.08665	1.005597	0.083511	0.007512	0.952
			GRFCSMIC	-0.12586	0.087487	0.995971	0.083299	0.007656	0.945

Table B.13: Simulation results for estimating β_x using the data imputation approach for error scenario 2 (errors in event indicator and failure time) with $N = 4000$, $n = 800$ comparing simple random sampling (SRS), case-control sampling (CC), and stratified case-control sampling (SCC). The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Design	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	90	log(1.5)	SRS	True	-0.19575	0.056825	2.316178	0.058701	0.00323	0.953
				HT	-0.09687	0.131616	1	0.130847	0.017323	0.946
				GRN	-0.21918	0.104644	1.257759	0.104506	0.010951	0.948
				GRMIS	-0.45879	0.105389	1.248862	0.103657	0.01111	0.946
				GRMIC	-0.09587	0.104913	1.254526	0.102638	0.011007	0.94
				GRFCSMIS	-0.28308	0.104808	1.255791	0.10352	0.010986	0.946
				GRFCSMIC	0.056619	0.106483	1.236032	0.102959	0.011339	0.944
			CC	True	-0.19575	0.056825	2.173331	0.058701	0.00323	0.953
				HT	1.443152	0.123499	1	0.120076	0.015286	0.936
				GRN	1.055416	0.108371	1.139601	0.105822	0.011763	0.934
				GRMIS	-0.3051	0.110035	1.122364	0.106001	0.012109	0.935
				GRMIC	1.012974	0.107239	1.151629	0.105819	0.011517	0.928
				GRFCSMIS	0.350989	0.107726	1.146416	0.105584	0.011607	0.939
				GRFCSMIC	-0.05063	0.110085	1.12185	0.105693	0.012119	0.93
			SCC	True	-0.19575	0.056825	2.024597	0.058701	0.00323	0.953
				HT	1.541675	0.115047	1	0.109119	0.013275	0.942
				GRN	1.3337	0.103861	1.107711	0.099885	0.010816	0.934
				GRMIS	1.34774	0.104582	1.100072	0.099689	0.010967	0.936
				GRMIC	1.120339	0.104766	1.098134	0.099614	0.010997	0.935
				GRFCSMIS	1.454201	0.105636	1.089092	0.099647	0.011194	0.939
				GRFCSMIC	0.826524	0.104132	1.104821	0.099525	0.010855	0.934
		log(3)	SRS	True	0.1293	0.064842	2.198686	0.06303	0.004206	0.954
				HT	0.1639	0.142567	1	0.139971	0.020328	0.946
				GRN	0.046035	0.120143	1.186638	0.115264	0.014435	0.943
				GRMIS	-0.33881	0.116575	1.222958	0.114212	0.013604	0.951
				GRMIC	-0.36136	0.117864	1.209587	0.113062	0.013908	0.941
				GRFCSMIS	-0.35298	0.117358	1.214796	0.113453	0.013788	0.946
				GRFCSMIC	-0.29133	0.117659	1.211692	0.11276	0.013854	0.945
			CC	True	0.1293	0.064842	2.034957	0.06303	0.004206	0.954
				HT	0.863599	0.13195	1	0.130468	0.017501	0.93
				GRN	0.906118	0.118367	1.114757	0.113327	0.01411	0.93
				GRMIS	0.354587	0.116178	1.135762	0.113018	0.013512	0.934
				GRMIC	0.490041	0.115837	1.1391	0.112904	0.013447	0.93
				GRFCSMIS	0.275773	0.115996	1.137543	0.112389	0.013464	0.931
				GRFCSMIC	0.28242	0.116929	1.128468	0.112369	0.013682	0.93
			SCC	True	0.1293	0.064842	1.918313	0.06303	0.004206	0.954
				HT	0.744288	0.124387	1	0.119663	0.015539	0.938
				GRN	0.856953	0.111821	1.112379	0.108713	0.012592	0.94
				GRMIS	0.448073	0.11011	1.129659	0.108819	0.012148	0.944
				GRMIC	0.589966	0.111043	1.120168	0.108831	0.012373	0.944
				GRFCSMIS	0.529959	0.109927	1.131544	0.108549	0.012118	0.942
				GRFCSMIC	0.279283	0.110541	1.125258	0.108364	0.012229	0.94

Table B.14: Simulation results for estimating β_x using the IF imputation approach for error scenario 2 (errors in event indicator and failure time) with $N = 4000$, $n = 800$ comparing simple random sampling (SRS), case-control sampling (CC), and stratified case-control sampling (SCC). The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Design	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	90	log(1.5)	SRS	True	0.013181	0.059384	2.257952	0.05879	0.003526	0.951
				HT	0.590898	0.134085	1	0.130847	0.017985	0.94
				GRN	0.764177	0.109177	1.228152	0.104129	0.011929	0.94
				GRMIS	-1.38011	0.108987	1.230285	0.103116	0.01191	0.93
				GRMIC	-1.16293	0.109731	1.221942	0.102465	0.012063	0.924
				GRFCSMIS	-1.4325	0.110198	1.216764	0.1025	0.012177	0.929
				GRFCSMIC	-0.97002	0.112221	1.194832	0.102177	0.012609	0.926
			CC	True	0.013181	0.059384	2.085539	0.05879	0.003526	0.951
				HT	3.005687	0.123847	1	0.121097	0.015487	0.94
				GRN	2.412196	0.111991	1.105862	0.106782	0.012638	0.934
				GRMIS	-0.87831	0.114389	1.082684	0.10729	0.013097	0.935
				GRMIC	-0.09236	0.12005	1.031628	0.10668	0.014412	0.922
				GRFCSMIS	-1.99286	0.112352	1.102314	0.106793	0.012688	0.931
				GRFCSMIC	-0.97418	0.116849	1.059884	0.106057	0.013669	0.916
			SCC	True	0.013181	0.059384	1.908976	0.05879	0.003526	0.951
				HT	0.351249	0.113362	1	0.109603	0.012853	0.946
				GRN	0.059211	0.102666	1.104187	0.10025	0.01054	0.945
				GRMIS	-2.49652	0.104083	1.089149	0.099592	0.010936	0.934
				GRMIC	-2.39435	0.102572	1.105198	0.099262	0.010615	0.931
				GRFCSMIS	-2.94422	0.102822	1.102511	0.099375	0.010715	0.937
				GRFCSMIC	-2.92328	0.10429	1.086985	0.099046	0.011017	0.933
		log(3)	SRS	True	0.090194	0.065292	2.228051	0.06311	0.004264	0.948
				HT	0.98782	0.145474	1	0.140929	0.02128	0.942
				GRN	0.462722	0.114252	1.273267	0.115252	0.013079	0.947
				GRMIS	-0.88636	0.113281	1.284186	0.112866	0.012927	0.94
				GRMIC	-1.00917	0.116412	1.249647	0.111686	0.013675	0.933
				GRFCSMIS	-1.29521	0.11259	1.292064	0.111025	0.012879	0.94
				GRFCSMIC	-1.21127	0.113696	1.279502	0.109949	0.013104	0.936
			CC	True	0.090194	0.065292	2.178578	0.06311	0.004264	0.948
				HT	1.522176	0.142244	1	0.129912	0.020513	0.918
				GRN	1.018669	0.121551	1.170234	0.113392	0.0149	0.922
				GRMIS	-0.35563	0.125685	1.13175	0.111195	0.015812	0.911
				GRMIC	0.023545	0.129196	1.100987	0.110825	0.016692	0.907
				GRFCSMIS	-0.50886	0.120788	1.177634	0.110295	0.014621	0.919
				GRFCSMIC	-0.13024	0.128134	1.110117	0.109476	0.01642	0.898
			SCC	True	0.090194	0.065292	2.036797	0.06311	0.004264	0.948
				HT	0.601277	0.132986	1	0.119273	0.017729	0.934
				GRN	0.594733	0.115309	1.153306	0.108376	0.013339	0.938
				GRMIS	-0.58776	0.11268	1.180213	0.106481	0.012738	0.942
				GRMIC	-0.44739	0.115677	1.14964	0.106537	0.013405	0.935
				GRFCSMIS	-0.97012	0.113241	1.174364	0.105788	0.012937	0.934
				GRFCSMIC	-0.87356	0.110695	1.201377	0.105449	0.012345	0.927

Table B.15: Simulation results for estimating β_x using the IF imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with $N = 4000$, $n = 800$ comparing simple random sampling (SRS), case-control sampling (CC), and stratified case-control sampling (SCC). The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Design	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	90	log(1.5)	SRS	True	0.013181	0.059384	2.333618	0.05879	0.003526	0.951
				HT	0.644516	0.138579	1	0.130362	0.019211	0.948
				GRN	1.081391	0.125327	1.10574	0.120056	0.015726	0.951
				GRMIS	3.093863	0.126174	1.098315	0.115137	0.016077	0.928
				GRMIC	2.868721	0.124744	1.110902	0.115237	0.015696	0.926
				GRFCSMIS	0.401896	0.119263	1.161954	0.111873	0.014226	0.94
				GRFCSMIC	0.594765	0.120779	1.147378	0.111389	0.014593	0.939
			CC	True	0.013181	0.059384	2.112828	0.05879	0.003526	0.951
				HT	1.421846	0.125467	1	0.120629	0.015775	0.944
				GRN	1.616807	0.124611	1.00687	0.120092	0.015571	0.942
				GRMIS	2.025107	0.125633	0.998682	0.115642	0.015851	0.926
				GRMIC	2.742054	0.130046	0.964796	0.115097	0.017035	0.919
				GRFCSMIS	-0.48853	0.114874	1.092217	0.111443	0.0132	0.936
				GRFCSMIC	-0.44037	0.121669	1.031218	0.110755	0.014807	0.93
			SCC	True	0.013181	0.059384	1.850378	0.05879	0.003526	0.951
				HT	0.544971	0.109882	1	0.110417	0.012079	0.944
				GRN	1.084536	0.110315	0.996079	0.110318	0.012189	0.946
				GRMIS	1.545407	0.112827	0.973897	0.107852	0.012769	0.93
				GRMIC	1.325522	0.114064	0.963336	0.107612	0.01304	0.93
				GRFCSMIS	-1.44068	0.105472	1.041814	0.104069	0.011158	0.947
				GRFCSMIC	-0.42746	0.107909	1.018281	0.103619	0.011647	0.94
		log(3)	SRS	True	0.090194	0.065292	2.211227	0.06311	0.004264	0.948
				HT	0.407488	0.144375	1	0.14084	0.020864	0.94
				GRN	0.385225	0.136955	1.054183	0.130108	0.018774	0.942
				GRMIS	1.24374	0.138733	1.040671	0.127417	0.019434	0.93
				GRMIC	1.30716	0.140311	1.028966	0.127054	0.019893	0.93
				GRFCSMIS	-0.73771	0.135317	1.066941	0.122208	0.018376	0.934
				GRFCSMIC	-0.66206	0.135598	1.064733	0.121757	0.01844	0.925
			CC	True	0.090194	0.065292	2.015433	0.06311	0.004264	0.948
				HT	1.264558	0.131591	1	0.130024	0.017509	0.935
				GRN	1.1846	0.134597	0.977673	0.128859	0.018286	0.932
				GRMIS	1.728262	0.134406	0.979058	0.126532	0.018426	0.924
				GRMIC	1.827018	0.136783	0.962044	0.126234	0.019113	0.919
				GRFCSMIS	-0.35803	0.131977	0.997079	0.120769	0.017433	0.921
				GRFCSMIC	0.105282	0.129613	1.015268	0.120185	0.016801	0.915
			SCC	True	0.090194	0.065292	1.900692	0.06311	0.004264	0.948
				HT	1.19953	0.1241	1	0.122704	0.015574	0.934
				GRN	1.281997	0.12254	1.01273	0.121992	0.015214	0.934
				GRMIS	1.168752	0.125257	0.99076	0.120536	0.015854	0.932
				GRMIC	1.34881	0.124024	1.000608	0.120272	0.015602	0.928
				GRFCSMIS	-0.73769	0.122563	1.012538	0.116614	0.015087	0.93
				GRFCSMIC	-0.36084	0.1256	0.988055	0.116154	0.015791	0.928

Table B.16: Misclassification generation process for the simulations testing misclassification generation with interactions. The sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and negative predictive value (NPV) for the event indicator are presented.

$\Delta^\star$	% Cens	β_x	β_z	Sens	Spec	PPV	NPV
$\text{Bernoulli}(\text{expit}(-1.1 + 0.5 * \Delta - 0.25 * X - 0.1 * U + 0.2 * Z + 0.85 * \Delta * X + 0.2 * \Delta * U + 0.8 * \Delta * z))$	50	log(1.5)	log(0.5)	0.833	0.889	0.860	0.867
		log(3)	log(0.5)	0.874	0.892	0.880	0.887
	75	log(1.5)	log(0.5)	0.768	0.818	0.573	0.917
		log(3)	log(0.5)	0.826	0.797	0.553	0.938
	90	log(1.5)	log(0.5)	0.709	0.734	0.224	0.959
		log(3)	log(0.5)	0.797	0.717	0.226	0.972

Table B.17: Simulation results for estimating β_x using the data imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with interaction terms in the misclassification generation, $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.03595	0.039644	2.357942	0.039422	0.001572	0.956
			HT	0.968647	0.093478	1	0.087968	0.008754	0.944
			GRN	2.065881	0.093214	1.002836	0.087707	0.008759	0.942
			GRMIS	2.037858	0.092802	1.007285	0.087701	0.00868	0.944
			GRMIC	2.068534	0.092468	1.010926	0.087639	0.008621	0.942
			GRFCSMIS	1.111458	0.06979	1.339419	0.07076	0.004891	0.954
			GRFCSMIC	0.806618	0.069849	1.338291	0.069445	0.00489	0.947
		log(3)	True	0.041168	0.041582	2.490834	0.04415	0.001729	0.948
			HT	0.313211	0.103573	1	0.097851	0.010739	0.942
			GRN	0.533937	0.1046	0.990187	0.097608	0.010976	0.946
			GRMIS	0.450849	0.10535	0.983136	0.097619	0.011123	0.945
			GRMIC	0.524909	0.104816	0.988144	0.097605	0.01102	0.947
			GRFCSMIS	0.28924	0.088922	1.164762	0.085733	0.007917	0.943
			GRFCSMIC	0.231675	0.088009	1.17685	0.08482	0.007752	0.938
	75	log(1.5)	True	0.119394	0.051672	2.316876	0.053276	0.00267	0.954
			HT	1.004049	0.119718	1	0.118566	0.014349	0.948
			GRN	1.895293	0.11939	1.002747	0.11707	0.014313	0.949
			GRMIS	2.027832	0.120512	0.993414	0.117998	0.014591	0.95
			GRMIC	2.267477	0.121239	0.987452	0.117415	0.014784	0.949
			GRFCSMIS	0.287842	0.099554	1.202546	0.104128	0.009912	0.952
			GRFCSMIC	0.811668	0.099056	1.208593	0.102347	0.009823	0.946
		log(3)	True	-0.01311	0.06088	2.250031	0.059211	0.003706	0.949
			HT	0.836351	0.136982	1	0.131293	0.018849	0.952
			GRN	1.165105	0.134278	1.020136	0.130112	0.018195	0.95
			GRMIS	1.064519	0.136716	1.001951	0.130604	0.018828	0.95
			GRMIC	1.028011	0.135967	1.007466	0.130518	0.018615	0.952
			GRFCSMIS	0.639607	0.123603	1.108244	0.121525	0.015327	0.949
			GRFCSMIC	0.474697	0.121686	1.125707	0.12072	0.014835	0.944
	90	log(1.5)	True	0.0138	0.084364	2.251745	0.083155	0.007117	0.947
			HT	1.897751	0.189966	1	0.183082	0.036146	0.94
			GRN	1.971642	0.18939	1.00304	0.179804	0.035933	0.94
			GRMIS	2.311279	0.189845	1.000635	0.179986	0.036129	0.948
			GRMIC	2.802321	0.186701	1.017489	0.177676	0.034986	0.94
			GRFCSMIS	-0.05938	0.184643	1.02883	0.174435	0.034093	0.94
			GRFCSMIC	-0.06001	0.18401	1.032369	0.171773	0.03386	0.934
		log(3)	True	-0.04654	0.088525	2.348938	0.089229	0.007837	0.95
			HT	0.928622	0.207941	1	0.196985	0.043343	0.939
			GRN	0.855951	0.204852	1.015079	0.19409	0.042053	0.938
			GRMIS	1.023834	0.205516	1.011799	0.193344	0.042363	0.939
			GRMIC	1.051391	0.203195	1.023355	0.190818	0.041422	0.937
			GRFCSMIS	0.819666	0.20135	1.032732	0.190641	0.040623	0.933
			GRFCSMIC	0.471738	0.197965	1.050389	0.1891	0.039217	0.935

Table B.18: Simulation results for estimating β_x using the IF imputation approach for error scenario 3 (errors in event indicator, failure time, and X) with interaction terms in the misclassification generation, $N = 2000$, $n = 400$, and simple random sampling. The $\%$ bias, empirical standard error (ESE), relative efficiency (RE), average standard error (ASE), mean squared error, and coverage probabilities (CP) are presented for 2000 simulated datasets.

β_z	% Cens	β_x	Method	% Bias	ESE	RE	ASE	MSE	CP
log(0.5)	50	log(1.5)	True	-0.03595	0.039644	2.357942	0.039422	0.001572	0.956
			HT	0.968647	0.093478	1	0.087968	0.008754	0.944
			GRN	2.065881	0.093214	1.002836	0.087707	0.008759	0.942
			GRMIS	1.469612	0.093884	0.995677	0.08226	0.00885	0.914
			GRMIC	0.960839	0.093115	1.003903	0.081983	0.008686	0.92
			GRFCSMIS	-0.01847	0.07019	1.331791	0.06906	0.004927	0.944
			GRFCSMIC	-0.27818	0.069723	1.340707	0.068681	0.004863	0.942
		log(3)	True	0.041168	0.041582	2.490834	0.04415	0.001729	0.948
			HT	0.313211	0.103573	1	0.097851	0.010739	0.942
			GRN	0.533937	0.1046	0.990187	0.097608	0.010976	0.946
			GRMIS	1.79989	0.103708	0.998707	0.092364	0.011146	0.927
			GRMIC	1.777451	0.102679	1.008709	0.09228	0.010924	0.925
			GRFCSMIS	-0.13418	0.095758	1.081613	0.084204	0.009172	0.934
			GRFCSMIC	-0.28799	0.094592	1.094954	0.083523	0.008958	0.932
	75	log(1.5)	True	0.119394	0.051672	2.316876	0.053276	0.00267	0.954
			HT	1.004049	0.119718	1	0.118566	0.014349	0.948
			GRN	1.895293	0.11939	1.002747	0.11707	0.014313	0.949
			GRMIS	3.457688	0.120251	0.99557	0.107183	0.014657	0.926
			GRMIC	3.698102	0.119798	0.999333	0.106444	0.014576	0.924
			GRFCSMIS	0.700943	0.104505	1.145569	0.101122	0.010929	0.947
			GRFCSMIC	0.953667	0.103681	1.154675	0.100249	0.010765	0.943
		log(3)	True	-0.01311	0.06088	2.250031	0.059211	0.003706	0.949
			HT	0.836351	0.136982	1	0.131293	0.018849	0.952
			GRN	1.165105	0.134278	1.020136	0.130112	0.018195	0.95
			GRMIS	1.141388	0.133559	1.025635	0.121304	0.017995	0.931
			GRMIC	1.081814	0.134695	1.016982	0.121179	0.018284	0.933
			GRFCSMIS	-0.38405	0.127498	1.074387	0.11702	0.016274	0.934
			GRFCSMIC	-0.25338	0.125074	1.095207	0.11659	0.015651	0.93
	90	log(1.5)	True	0.0138	0.084364	2.251745	0.083155	0.007117	0.947
			HT	1.897751	0.189966	1	0.183082	0.036146	0.94
			GRN	1.971642	0.18939	1.00304	0.179804	0.035933	0.94
			GRMIS	8.575347	0.207643	0.914869	0.168291	0.044324	0.902
			GRMIC	8.465425	0.199277	0.953278	0.165431	0.040889	0.892
			GRFCSMIS	4.650762	0.183287	1.036438	0.165352	0.03395	0.925
			GRFCSMIC	4.944207	0.18214	1.042967	0.161798	0.033577	0.91
		log(3)	True	-0.04654	0.088525	2.348938	0.089229	0.007837	0.95
			HT	0.928622	0.207941	1	0.196985	0.043343	0.939
			GRN	0.855951	0.204852	1.015079	0.19409	0.042053	0.938
			GRMIS	4.226444	0.204751	1.015579	0.183868	0.044079	0.916
			GRMIC	4.045888	0.205878	1.010016	0.181946	0.044362	0.915
			GRFCSMIS	1.592917	0.195551	1.063359	0.178513	0.038546	0.911
			GRFCSMIC	1.238013	0.201628	1.031308	0.176939	0.040839	0.908

Table B.19: The median hazard ratios (HR) and their corresponding 95% confidence interval widths calculated using the IF imputation method from 100 different sampled validation subsets for a 100 cell/mm^3 increase in CD4 count at ART initiation and 10-year increase in age at CD4 count measurement.

Subset size	Sampling	Method	CD4 HR	CD4 CI width	Age HR	Age CI width
340	CC	True	0.693	0.19	0.829	0.361
		Naive	0.91	0.125	1.087	0.275
		HT	0.677	0.323	0.805	0.576
		GRN	0.68	0.284	0.821	0.477
		GRMIS	0.704	0.323	0.807	0.526
		GRMIC	0.695	0.296	0.804	0.492
		GRFCSMIS	0.69	0.307	0.813	0.488
		GRFCSMIC	0.684	0.299	0.813	0.463
	SCCB	True	0.693	0.19	0.829	0.361
		Naive	0.91	0.125	1.087	0.275
		HT	0.682	0.283	0.855	0.571
		GRN	0.682	0.278	0.835	0.497
		GRMIS	0.691	0.284	0.851	0.515
		GRMIC	0.691	0.277	0.861	0.499
		GRFCSMIS	0.7	0.289	0.846	0.506
		GRFCSMIC	0.702	0.282	0.848	0.49
	SCCN	True	0.693	0.19	0.829	0.361
		Naive	0.91	0.125	1.087	0.275
		HT	0.694	0.31	0.829	0.702
		GRN	0.69	0.304	0.813	0.609
		GRMIS	0.711	0.303	0.82	0.583
		GRMIC	0.715	0.298	0.824	0.57
		GRFCSMIS	0.708	0.301	0.838	0.566
		GRFCSMIC	0.723	0.298	0.826	0.561
680	CC	True	0.693	0.19	0.829	0.361
		Naive	0.91	0.125	1.087	0.275
		HT	0.691	0.237	0.839	0.411
		GRN	0.69	0.227	0.83	0.386
		GRMIS	0.696	0.234	0.829	0.391
		GRMIC	0.7	0.232	0.834	0.385
		GRFCSMIS	0.696	0.232	0.832	0.388
		GRFCSMIC	0.702	0.23	0.83	0.386
	SCCB	True	0.693	0.19	0.829	0.361
		Naive	0.91	0.125	1.087	0.275
		HT	0.688	0.228	0.828	0.413
		GRN	0.69	0.227	0.821	0.387
		GRMIS	0.694	0.23	0.831	0.398
		GRMIC	0.697	0.229	0.83	0.39
		GRFCSMIS	0.698	0.23	0.826	0.393
		GRFCSMIC	0.7	0.231	0.824	0.388
	SCCN	True	0.693	0.19	0.829	0.361
		Naive	0.91	0.125	1.087	0.275
		HT	0.688	0.231	0.832	0.438
		GRN	0.687	0.231	0.832	0.409
		GRMIS	0.693	0.232	0.825	0.407
		GRMIC	0.694	0.231	0.825	0.402
		GRFCSMIS	0.694	0.233	0.823	0.402
		GRFCSMIC	0.698	0.233	0.828	0.4

B.2. VCCC analysis details

For this study, we analyzed data on 4797 HIV-positive patients that had been fully validated and applied some common inclusion/exclusion criteria used in HIV studies to obtain the final analysis dataset. Specifically, any patients that had an indeterminate ART start date, no CD4 count measurement between 180 days before or 30 days after starting ART, no follow-up visits in the clinic after starting ART, an ADE before starting ART, or an indeterminate ADE date were excluded. In addition, patients must have been at least 18 years of age at ART start and not started ART prior to enrollment. Lastly, any ADE within 6 months of starting ART were not considered a true failure due to the time required for ART to be efficacious. After application of these criteria, the unvalidated and validated data contained 1995 and 1595 patients, respectively. The 1595 patients that met the criteria in the validated dataset were used for the analysis of the ADE outcome.

The censoring rate among the 1595 patients was very high at 93.8%, suggesting that an outcome-dependent sampling design that oversamples cases would be necessary. Of the 1595 patients, 11% had a misclassified ADE; specifically, 161 were incorrectly classified as having an ADE and 12 were incorrectly classified as having been censored. For the failure times, 34.5% were incorrect, with the errors having mean and standard deviation of -0.75 and 2.89 years, respectively. There were errors in the CD4 count at ART start for only 6.7% of the patients; however, the errors were right skewed, having mean and standard deviation of 10 and $154\ \mathrm{cell}/\mathrm{mm}^3$, respectively. In addition, the errors in the failure times and CD4 count at ART start had a correlation of -0.10.

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

C.1. Supplementary tables and figures

Table C.1: The quantiles, mean, and standard deviation (SD) for the error-prone event time divided by true event time $\left(\frac{T'}{T}\right)$ for $\beta = \log(1.5)$ and $n = 1000$.

Error Distribution	σ_ν^2	5^{th}	25^{th}	50^{th}	Mean	75^{th}	95^{th}	SD
Normal	0.25	0.460	0.715	1.013	1.146	1.410	2.407	0.608
	0.5	0.337	0.575	0.973	1.254	1.586	3.009	0.964
	1	0.190	0.531	1.100	1.687	2.078	5.021	1.949
	2	0.091	0.375	0.948	2.550	2.521	10.15	5.139
Shifted Gamma	0.25	0.518	0.700	0.927	1.208	1.367	2.753	0.938
	0.5	0.437	0.588	0.863	1.484	1.444	4.143	2.810
	1	0.392	0.493	0.772	2.970	1.609	8.798	14.76
	2	0.370	0.405	0.564	35.99	1.329	23.49	657.0

Table C.2: The percent (%) bias (absolute bias for intercept α_0), average model standard error (ASE), empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets with $n = 1000$, exponential time, and shifted gamma error.

			Weibull Model				Cox Model
β	σ_ν^2		α_0	α_1	shape	β	β
log(1.5)	0.25	Bias	0.059	0.120	-12.51	-14.23	-18.39
		ASE	0.039	0.041	0.021	0.036	0.033
		ESE	0.040	0.042	0.000	0.038	0.034
		MSE	0.071	0.042	0.125	0.069	0.082
		CP	0.664	0.973	0.000	0.640	0.390
	0.5	Bias	0.119	0.020	-25.17	-26.58	-29.89
		ASE	0.046	0.055	0.020	0.041	0.033
		ESE	0.046	0.061	0.000	0.046	0.035
		MSE	0.127	0.061	0.252	0.117	0.126
		CP	0.252	0.964	0.000	0.262	0.058
	1	Bias	0.233	-0.220	-47.33	-43.70	-41.78
		ASE	0.061	0.087	0.016	0.049	0.032
		ESE	0.061	0.105	0.000	0.060	0.035
		MSE	0.240	0.105	0.473	0.187	0.173
		CP	0.020	0.959	0.000	0.074	0.001
	2	Bias	0.441	1.110	-61.09	-60.65	-50.90
		ASE	0.171	0.210	0.028	0.089	0.032
		ESE	0.095	0.203	0.000	0.077	0.035
		MSE	0.451	0.203	0.611	0.258	0.209
		CP	0.005	0.939	0.004	0.056	0.000
log(3)	0.25	Bias	0.059	0.040	-15.62	-14.34	-16.53
		ASE	0.039	0.042	0.021	0.042	0.040
		ESE	0.039	0.041	0.000	0.046	0.046
		MSE	0.073	0.041	0.156	0.164	0.187
		CP	0.636	0.978	0.000	0.052	0.021
	0.5	Bias	0.118	-0.030	-24.78	-26.43	-28.70
		ASE	0.046	0.055	0.020	0.046	0.038
		ESE	0.045	0.059	0.000	0.056	0.049
		MSE	0.126	0.059	0.248	0.296	0.319
		CP	0.264	0.974	0.000	0.000	0.000
	1	Bias	0.230	-0.050	-45.67	-43.38	-42.55
		ASE	0.061	0.088	0.017	0.052	0.035
		ESE	0.060	0.105	0.000	0.074	0.049
		MSE	0.238	0.105	0.457	0.482	0.470
		CP	0.024	0.968	0.000	0.000	0.000
	2	Bias	0.437	0.280	-59.46	-60.81	-53.65
		ASE	0.157	0.211	0.027	0.069	0.034
		ESE	0.094	0.200	0.000	0.089	0.047
		MSE	0.447	0.200	0.595	0.674	0.591
		CP	0.001	0.941	0.000	0.000	0.000

Table C.3: The percent ($\%$) bias (absolute bias for intercept α_0), average model standard error (ASE), empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets with $n = 1000$, log-normal time, and mean zero normal error.

			Weibull Model				Cox Model
β	σ_ν^2		α_0	α_1	shape	β	β
log(3)	1	Bias	0.499	0.020	3.59	0.520	-3.510
		ASE	0.034	0.040	0.030	0.052	0.041
		ESE	0.034	0.042	0.000	0.055	0.053
		MSE	0.500	0.042	0.036	0.055	0.065
		CP	0.000	0.975	1.000	0.931	0.772

Table C.4: The percent ($\%$) bias, average bootstrap standard error (ASE) for SIMEX, average model standard error (ASE) for naive, empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with $n = 1000$, exponential time, baseline hazard of 0.1, a normal distribution for the additive error term, and 90% uniform censoring for the true event time.

$\%$ Censored	σ_ν^2	Method	$\%$ Bias	ASE	ESE	MSE	CP
90	1	SIMEX	-11.59	0.125	0.122	0.168	0.834
		Naive	-13.55	0.101	0.101	0.169	1.000

Table C.5: The percent ($\%$) bias, average bootstrap standard error (ASE) for SIMEX, average model standard error (ASE) for naive, empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with $n = 1000$, exponential time, baseline hazard of 0.1, a normal distribution for the multiplicative error term, and 90% covariate-dependent censoring for the true event time.

$\%$ Censored	σ_ν^2	Method	$\%$ Bias	ASE	ESE	MSE	CP
90	0.5	SIMEX	-7.912	0.118	0.115	0.139	0.894
		Naive	-14.70	0.104	0.101	0.179	1.000

Table C.6: The quantiles, interquartile range (IQR), and standard deviation (SD) for the ratio of the error-prone simulated event time and the true event time for virological failure $\left(\frac{T'_{\lambda b}}{T}\right)$ in the VCCC example.

λ	25^{th}	50^{th}	75^{th}	IQR	SD
0	1	1	1	0	32.06
0.5	0.669	1.026	1.543	0.874	38.57
1	0.570	1.002	1.784	1.214	48.40
1.5	0.499	1.020	2.064	1.565	68.25
2	0.452	0.991	2.239	1.787	56.29

Table C.7: The hazard ratios (HR) and their corresponding bootstrap 95% confidence intervals for sex, a 100-unit increase in enrollment CD4, and a 10 year increase in age at enrollment for the time at first opportunistic infection post ART.

	Univariate		
	Sex	100 $\times$ CD4	10 $\times$ Age at Enrollment
True	0.951 (0.790,1.146)	0.781 (0.748,0.816)	1.146 (1.057,1.242)
Naive	1.053 (0.903,1.229)	0.840 (0.813,0.868)	1.153 (1.079,1.232)
SIMEX	1.078 (0.914,1.270)	0.846 (0.808,0.885)	1.177 (1.101,1.259)

	Multivariate		
	Sex	100 $\times$ CD4	10 $\times$ Age at Enrollment
True	0.822 (0.682,0.991)	0.782 (0.749,0.817)	1.113 (1.025,1.209)
Naive	0.928 (0.795,1.084)	0.843 (0.815,0.871)	1.121 (1.048,1.200)
SIMEX	0.908 (0.754,1.095)	0.845 (0.806,0.886)	1.145 (1.061,1.235)

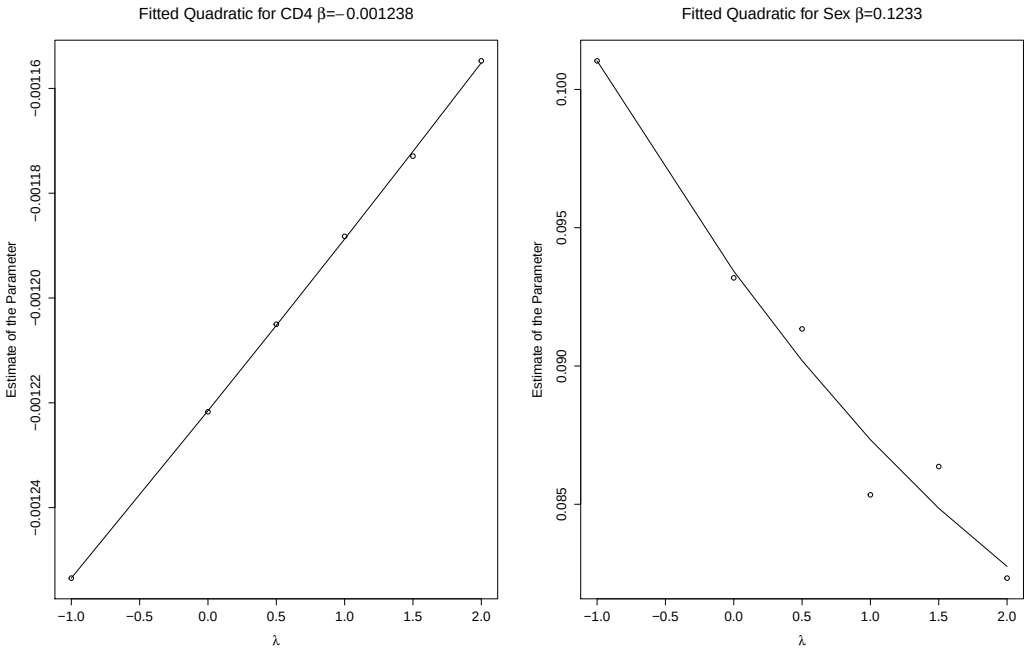
Table C.8: The percent ($\%$) bias, average bootstrap standard error (ASE) for SIMEX, average model standard error (ASE) for naive, empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with exponential time and mixture gamma, mean zero normal, and shifted gamma error distributions.

Error Distribution	Covariate	Method	$\%$ Bias	ASE	ESE	MSE	CP
Mixed	CD4	SIMEX	-1.760	0.0001	0.0001	0.0001	0.316
		Naive	-12.40	0.0001	0.0001	0.0002	0.198
	Gender	SIMEX	-1.520	0.058	0.059	0.059	0.880
		Naive	-13.16	0.051	0.052	0.052	0.878
	Age	SIMEX	-3.050	0.003	0.003	0.003	0.624
		Naive	-14.99	0.002	0.002	0.002	0.623
Normal	CD4	SIMEX	-4.930	0.0001	0.0001	0.0001	0.299
		Naive	-11.82	0.0001	0.0001	0.0002	0.204
	Gender	SIMEX	-3.170	0.059	0.060	0.060	0.875
		Naive	-11.51	0.051	0.052	0.052	0.878
	Age	SIMEX	-2.580	0.003	0.003	0.003	0.634
		Naive	-10.51	0.002	0.002	0.002	0.636
Gamma	CD4	SIMEX	-6.620	0.0001	0.0001	0.0001	0.288
		Naive	-13.23	0.0001	0.0001	0.0002	0.180
	Gender	SIMEX	-7.880	0.058	0.059	0.059	0.896
		Naive	-14.94	0.051	0.052	0.053	0.892
	Age	SIMEX	-10.62	0.002	0.002	0.002	0.630
		Naive	-17.33	0.002	0.002	0.002	0.633

Table C.9: The percent ($\%$) bias, average bootstrap standard error (ASE) for SIMEX, average model standard error (ASE) for naive, empirical standard error (ESE), mean squared error (MSE), and coverage probabilities (CP) are given for 2000 simulated data sets for the SIMEX and naive methods with $n = 1000$, exponential time, and a left-skewed gamma error distribution.

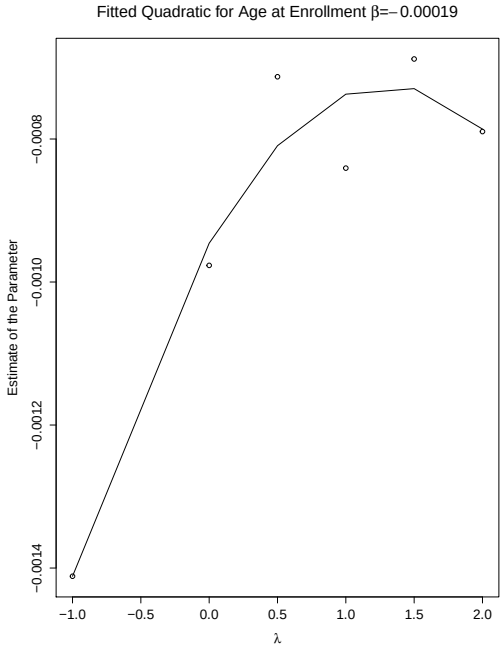
β	σ_ν^2	Method	$\%$ Bias	ASE	ESE	MSE	CP
log(1.5)	0.5	SIMEX	4.838	0.045	0.046	0.050	0.928
		Naive	-17.79	0.033	0.033	0.079	0.399
	1	SIMEX	12.38	0.051	0.051	0.072	0.828
		Naive	-22.07	0.033	0.033	0.095	0.22
log(3)	0.5	SIMEX	3.762	0.058	0.059	0.072	0.891
		Naive	-17.08	0.040	0.041	0.192	0.008
	1	SIMEX	9.189	0.064	0.066	0.121	0.656
		Naive	-22.24	0.040	0.039	0.247	0.000

Figure C.1: The quadratic approximations of the β parameters as a function of λ for CD4 (a), sex (b), and age at enrollment (c), extrapolated to $\lambda = -1$.



(a)

(b)



(c)

BIBLIOGRAPHY

Adler-Milstein, J and Jha, AK (2017). HITECH Act drove large gains in hospital electronic health record adoption. *Health Affairs* 36.8, 1416–1422.

Andersen, PK and Gill, RD (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 1100–1120.

Barnard, J and Rubin, DB (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* 86.4, 948–955.

Beresniak, A, Schmidt, A, Proeve, J, Bolanos, E, Patel, N, Ammour, N, Sundgren, M, Ericson, M, Karakoyun, T, Coorevits, P, et al. (2016). Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. *Contemporary Clinical Trials* 46, 85–91.

Boe, LA, Tinker, LF, and Shaw, PA (2020). An Approximate Quasi-Likelihood Approach for Error-Prone Failure Time Outcomes and Exposures. *arXiv preprint arXiv:2004.01112*.

Botsis, T, Hartvigsen, G, Chen, F, and Weng, C (2010). Secondary use of EHR: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics* 2010, 1.

Breslow, NE and Chatterjee, N (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48.4, 457–468.

Breslow, NE, Lumley, T, Ballantyne, CM, Chambless, LE, and Kulich, M (2009). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. *Statistics in Biosciences* 1.1, 32–49.

Breslow, NE and Wellner, JA (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* 34.1, 86–102.

Carroll, RJ, Ruppert, D, Stefanski, LA, and Crainiceanu, CM (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press.

Chen, T and Lumley, T (2020). Optimal multi-wave sampling for regression modelling in two-phase designs. *arXiv preprint arXiv:2005.13739*.

Cochran, WG (2007). *Sampling techniques*. John Wiley & Sons.

Cook, JR and Stefanski, LA (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 89.428, 1314–1328.

Deville, JC and Särndal, CE (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87.418, 376–382.

Deville, JC, Särndal, CE, and Sautory, O (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88.423.

Duda, SN, Shepherd, BE, Gadd, CS, Masys, DR, and McGowan, CC (2012). Measuring the quality of observational study data in an international HIV research network. *PloS ONE* 7.4, e33908.

Edwards, JK, Cole, SR, Troester, MA, and Richardson, DB (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American Journal of Epidemiology* 177.9, 904–912.

Floyd, JS, Heckbert, SR, Weiss, NS, Carrell, DS, and Psaty, BM (2012). Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. *Journal of the American Medical Association* 307.15, 1580–1582.

Giganti, MJ, A., SP, Chen, G, Bebaway, SS, Turner, MM, Sterling, TR, and Shepherd, BE (2020). Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples, and multiple imputation. *Annals of Applied Statistics, in press*.

Gravel, CA, Dewanji, A, Farrell, PJ, and Krewski, D (2018). A validation sampling approach for consistent estimation of adverse drug reaction risk with misclassified right-censored survival data. *Statistics in Medicine* 37.27, 3887–3903.

Greene, WF and Cai, J (2004). Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics* 60.4, 987–996.

Han, K, Lumley, T, Shepherd, BE, and Shaw, PA (2020). Two-phase analysis and study design for survival models with error-prone exposures. *arXiv preprint arXiv:2005.05511*.

Han, K, Shaw, PA, and Lumley, T (2019). Combining multiple imputation with raking of weights in the setting of nearly-true models. *arXiv preprint arXiv:1910.01162*.

Han, P (2016). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics* 43.1, 246–260.

He, W, Yi, GY, and Xiong, J (2007). Accelerated failure time models with covariates subject to measurement error. *Statistics in Medicine* 26.26, 4817–4832.

Hillestad, R, Bigelow, J, Bower, A, Girosi, F, Meili, R, Scoville, R, and Taylor, R (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs* 24.5, 1103–1117.

Holcroft, CA, Rotnitzky, A, and Robins, JM (1997). Efficient estimation of regression parameters from multistage studies with validation of outcome and covariates. *Journal of Statistical Planning and Inference* 65.2, 349–374.

Holt, D, McDonald, J, and Skinner, C (1991). The effect of measurement error on event history analysis. In: *Measurement Errors in Surveys*. Ed. by PP Biemer, RM Groves, LE Lyberg, NA Mathiowetz, and S Sudman. Wiley, 665–685.

Hong, S, Schmitt, N, Stone, A, and Denne, J (2012). Attenuation of treatment effect due to measurement variability in assessment of progression-free survival. *Pharmaceutical Statistics* 11.5, 394–402.

Hougaard, P, Myglegaard, P, and Borch-Johnsen, K (1994). Heterogeneity models of disease susceptibility, with application to diabetic nephropathy. *Biometrics*, 1178–1188.

Hu, P, Tsiatis, AA, and Davidian, M (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics*, 1407–1419.

Huang, Y and Wang, C (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association* 95.452, 1209–1219.

Huang, Y and Wang, C (2006). Errors-in-covariates effect on estimating functions: Additivity in limit and nonparametric correction. *Statistica Sinica*, 861–881.

Hunsberger, S, Albert, PS, and Dodd, L (2010). Analysis of progression-free survival data using a discrete time survival model that incorporates measurements with and without diagnostic error. *Clinical Trials*, 1740774510384887.

Jensen, PB, Jensen, LJ, and Brunak, S (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13.6, 395–405.

Keiding, N, Andersen, PK, and Klein, JP (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* 16.2, 215–224.

Kiragga, AN, Castelnuovo, B, Schaefer, P, Muwonge, T, and Easterbrook, PJ (2011). Quality of data collection in a large HIV observational clinic database in sub-Saharan Africa: implications for clinical research and audit of care. *Journal of the International AIDS Society* 14.1, 3–3.

Korn, EL, Dodd, LE, and Freidlin, B (2010). Measurement error in the timing of events: effect on survival analyses in randomized clinical trials. *Clinical Trials*.

Küchenhoff, H, Mwalili, SM, and Lesaffre, E (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* 62.1, 85–96.

Kulich, M and Lin, DY (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* 99.467, 832–844.

Lemly, DC, Shepherd, BE, Hulgan, T, Rebeiro, P, Stinnette, S, Blackwell, RB, Bebawy, S, Kheshti, A, Sterling, TR, and Raffanti, SP (2009). Race and sex differences in antiretroviral therapy use and mortality among HIV-infected persons in care. *Journal of Infectious Diseases* 199.7, 991–998.

Li, Y and Lin, X (2000). Covariate measurement errors in frailty models for clustered survival data. *Biometrika* 87.4, 849–866.

Lumley, T (2011). *Complex Surveys: A Guide to Analysis Using R*. Vol. 565. John Wiley & Sons.

Lumley, T (2016). *Survey: Analysis of Complex Survey Samples*. R package version 3.32.

Lumley, T (2017). Robustness of semiparametric efficiency in nearly-true models for two-phase samples. *arXiv preprint arXiv:1707.05924*.

Lumley, T, Shaw, PA, and Dai, JY (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* 79.2, 200–220.

Magaret, AS (2008). Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Statistics in Medicine* 27.26, 5456–5470.

Magder, LS and Hughes, JP (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* 146.2, 195–203.

Marshall, WA and Tanner, JM (1986). Puberty. In: *Postnatal Growth Neurobiology*. Springer, 171–209.

McIsaac, MA and Cook, RJ (2015). Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Statistics in Medicine* 34.21, 2899–2912.

Meier, AS, Richardson, BA, and Hughes, JP (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics* 59.4, 947–954.

Nakamura, T (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, 829–838.

Oh, EJ, Shepherd, BE, Lumley, T, and Shaw, PA (2018). Considerations for analysis of time-to-event outcomes measured with error: bias and correction with SIMEX. *Statistics in Medicine* 37.8, 1276–1289.

Oh, EJ, Shepherd, BE, Lumley, T, and Shaw, PA (2019). Raking and Regression Calibration: Methods to Address Bias from Correlated Covariate and Time-to-Event Error. *arXiv preprint arXiv:1905.08330*.

Prentice, R (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69.2, 331–342.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robins, JM, Rotnitzky, A, and Zhao, LP (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89.427, 846–866.

Robins, JM and Wang, N (2000). Inference for imputation estimators. *Biometrika* 87.1, 113–124.

Rubin, DB (2004). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

Saegusa, T and Wellner, JA (2013). Weighted likelihood estimation under two-phase sampling. *Annals of Statistics* 41.1, 269–295.

Shaw, PA, He, J, and Shepherd, BE (Nov. 2018). Regression calibration to correct correlated errors in outcome and exposure. *arXiv preprint arXiv:1811.10147*. arXiv: arXiv:1811.10147 [stat.ME].

Shaw, PA and Prentice, RL (2012). Hazard Ratio Estimation for Biomarker-Calibrated Dietary Exposures. *Biometrics* 68.2, 397–407.

Shepherd, BE and Yu, C (2011). Accounting for data errors discovered from an audit in multiple linear regression. *Biometrics* 67.3, 1083–1091.

Skinner, CJ and Humphreys, K (1999). Weibull regression for lifetimes measured with error. *Lifetime Data Analysis* 5.1, 23–37.

Staa, TP van, Dyson, L, McCann, G, Padmanabhan, S, Belatri, R, Goldacre, B, Cassell, J, Pirmohamed, M, Torgerson, D, Ronaldson, et al. (2014). The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technology Assessment* 18.43, 1–146.

Swindell, WR (2009). Accelerated failure time models provide a useful statistical framework for aging research. *Experimental Gerontology* 44.3, 190–200.

Tsiatis, AA and Davidian, M (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* 88.2, 447–458.

Van Buuren, S (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16.3, 219–242.

Van der Vaart, AW (1998). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

Wang, CY, Hsu, L, Feng, ZD, and Prentice, RL (1997). Regression calibration in failure time regression. *Biometrics*, 131–145.

Wang, L, Shaw, P, Mathelier, H, Kimmel, S, and French, B (2016). Evaluating Risk-Prediction Models Using Data From Electronic Health Records. *Annals of Applied Statistics*.

Wei, L (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* 11.14-15, 1871–1879.

Weiskopf, NG and Weng, C (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* 20.1, 144–151.

Wu, C and Sitter, RR (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* 96.453, 185–193.

Xie, SX, Wang, CY, and Prentice, RL (2001). A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.4, 855–870.

Zhang, J, He, W, and Li, H (2014). A semiparametric approach for accelerated failure time models with covariates subject to measurement error. *Communications in Statistics-Simulation and Computation* 43.2, 329–341.