Using observational data in healthcare research: New methods to design, conduct, and analyze efficient two-phase designs

By

Sarah Camilla Lotspeich

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

May 31, 2021

Nashville, Tennessee

Approved:

Jonathan Schildcrout, Ph.D.

Bryan Shepherd, Ph.D.

Ran Tao, Ph.D.

Peter Rebeiro, Ph.D.

ACKNOWLEDGEMENTS

With the fullest heart, I want to begin by thanking my parents, Rob and Kyndra, whose hard work and support have given me everything. I would also like to thank my brother, Ross, for keeping me grounded and giving me plenty to brag about (I tell everyone that "my brother is a tugboat captain"), and my grandmother, Bobbie, for always, always, always answering the phone to chat with me as I run all over town! Of course, I really have to thank my entire family for their constant support. You were all so wonderful, understanding, and happy to share in this journey with me. I am so blessed to have such a fabulous group of cheerleaders.

There are so many people at Vanderbilt to thank. My advisers Bryan Shepherd and Ran Tao were so supportive, patient, and enthusiastic about this work; they always encouraged me to travel when I could and remember my work-life balance, and I am appreciative of how their mentorship so wonderfully shaped my PhD experience. I am also thankful to my other committee members, Jonathan Schildcrout and Peter Rebeiro, for their insights and suggestions on my work. And to my unofficial fifth committee member, Gustavo Amorim, who was sometimes more of a life coach - thank you for taking the time to convince me I was right, even when I couldn't see it. This group really made for an enjoyable dissertation. I am grateful to the Biostatistics graduate program, especially my cohort members and friends Nathan James, Elizabeth Sigworth, and Valerie Welty, our "honorary" cohort member/my travel partner, Hannah Weeks, and Mark Giganti, who helped kickstart my research and introduced me to the magic of international conferences. From Baltimore to Vancouver, thank you to all of my fellow students who followed me to some famous donut shop while away at a conference. Thank you to the fantastic collaborators who became valuable mentors, Rameela Raman and Staci Sudenga, and to my CCASAnet colleagues, especially Stephany Duda and Cathy McGowan, whose work, support, and passion were the backbone of this dissertation. Last, but certainly not least, thank you to my collaborators from the Penn-Vanderbilt-Auckland Research Collaborative, especially Pamela Shaw who provided valuable feedback on my work.

Thank you to the people who choose to be my friend even though I've been ranting about error-prone data for as long as I can remember now. Bridget, thank you for proofreading all of my emails, for our annual "Bridget–Sarah messarounds," and for being such a constant in my life for the last seven (!!) years. Amber, thank you for all of the quality afternoons typing away in coffee shops and for always encouraging

"second coffee." And to the friends who sent words of encouragement, met me on trips, or had virtual movie nights, especially Alyssa, Claudia, Courtney, Dalal, and Sarah - thank you for getting me to stop working and enjoy myself.

Finally, thank you Matthew for seeing me through the finish line and supporting me no matter where in the world you are. I can't wait to start our next chapter together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

$\sqrt{CD4}$ Square root transformed CD4 count

ADE AIDS-defining event

ART Antiretroviral therapy

BCC* Unvalidated balanced case-control sampling

CC Case-control sampling

CC* Unvalidated case-control sampling

CCASAnet Caribbean, Central, and South America network for HIV epidemiology

CI Confidence interval

CoG Country grouping

CP Coverage probability of the 95% confidence interval

EHR Electronic health records

EM Expectation-Maximization

FNR False negative rate

FPR False positive rate

FPR_0 Baseline false positive rate

HT Horvitz–Thompson estimator

IPW Inverse probability weighted estimator

IQR Interquartile range

MAR Missing at random

MLE Maximum likelihood estimator

OR Odds ratio

PLWH People living with HIV/AIDS

RC Regression calibration

RE Relative efficiency

RI Relative interquartile range

SDV Source document verification

SE Empirical standard error

SEE Average standard error estimates

SMLE Sieve maximum likelihood estimator

SRS Simple random sampling

TB Tuberculosis

TPR_0 Baseline true positive rate

VCCC Vanderbilt Comprehensive Care Clinic

VDCC CCASAnet Data Coordinating Center at Vanderbilt

CHAPTER 1

INTRODUCTION

The volume of routinely collected data, like those in observational databases such as electronic health records (EHR), is steadily climbing, increasing the accessibility to clinically meaningful variables. Perhaps it is to be expected, then, that the uptake of repurposing observational data, e.g., for analysis, research, or to inform policy, has grown accordingly (Safran et al., 2007; Kim et al., 2019). Of particular appeal to biomedical researchers is the ability to analyze data that come at little to no additional cost for collection. Thus, we have seen the utilization of EHR data surge in many clinical domains, including HIV and AIDS (Zaniewski et al., 2018), genetics (Wei and Denny, 2015), emergency medicine (Green, 2013), and therapeutic effectiveness (Tannen et al., 2009) to name a few. While the benefits might be clear, there are a number of obstacles to responsibly analyzing EHR data that need to be addressed, chief among them the error-prone nature of routinely collected data (Safran et al., 2007).

So-called "secondary use" data like those extracted from the EHR or other observational databases are expected to be error-prone since their primary purpose was in direct support of patient care, rather than analysis (Nordo et al., 2019). Many papers have noted quality concerns with observational data, particularly in the EHR, as a major hurdle to large-scale adoption in healthcare research (Hersh et al., 2013; Kim et al., 2019). Others have described the impacts of error-prone data on clinical analyses (Green, 2013; Chen et al., 2019; Giganti et al., 2019). In the statistics literature, errors in variables are described as either measurement error or misclassification, with the former applying to continuous variables and the latter to categorical ones. Common sources of continuous measurement error include instrument error or errors due to self-report; examples of these errors include mismeasured lab values and dietary intake, respectively (Keogh et al., 2020). Misclassification can be the result of imperfect diagnostic testing, i.e., due to low sensitivity or specificity, but in secondary use data, like EHR, there is particular concern of misclassification in derived outcomes such as disease status or phenotype (e.g., Sinnott et al. (2014); Beesley and Mukherjee (2020)). Errors in observational data can also be quite complicated. They can be the result of complex relationships between error-prone and error-free variables which can be difficult to identify in practice.

To ensure the integrity of observational data, complete data validation would be

ideal. However, this is an expensive undertaking that is often unattainable in practice, particularly for large databases like EHR. For example, the Vanderbilt Comprehensive Care Clinic (VCCC) in Nashville, Tennessee spends >$60,000 US annually to sustain full-data validation, i.e., ongoing review of all patients and variables, in their EHR. A cost-effective alternative to complete data validation is a two-phase design (White, 1982) or validation study. Phase I consists of the error-prone variables which are available or inexpensive to obtain for all subjects, e.g., from the EHR; this information can be used to select the Phase II subsample. Then, Phase II involves reviewing the records for a subset of people, e.g., through chart review, to collect the validated outcome and predictor(s). Many statistical methods have been proposed which use all data from both Phases I and II, thus attaining high-powered inference based on an audited subset (e.g., Tang et al. (2015) and Tao et al. (2021)).

While two-phase designs are promising alternatives to full-data validation, they remain resource-intensive undertakings. Even auditing subsets of patients or variables can be costly and time-consuming. For instance, a survey of the Swedish Association of the Pharmaceutical Industry found that data auditing consumed 25% of study budgets, on average (Funning et al., 2009). We have identified key opportunities to maximize the research return on two-phase designs along what we call the audit-to-analysis pipeline: (i) design of Phase II, (ii) audit protocol to collect data in Phase II, and (iii) analysis of Phases I and II. In this dissertation, we propose novel methods to promote the statistical and practical efficiency of two-phase designs for data quality at each of these stages.

In practice, the size of a data audit is often resource-constrained, which can limit the numbers of patient records and variables that can be reviewed. Thus, selecting the most informative patients for validation is paramount. In Chapter 2, we use the asymptotic properties of the maximum likelihood estimator (MLE) to derive the optimal validation study design to obtain the most efficient log odds ratios under binary outcome and exposure misclassification. Since the optimal design is a function of unknown parameters, and thus not implementable in practice, we propose a multi-wave approximation to it, as well. The multi-wave optimal design breaks the audit into two "waves": the first wave can be used to estimate the necessary parameters to decide optimal allocation of subjects in the second wave. Since the variance of the MLE cannot be minimized directly (i.e., there is no closed-form solution), we propose a novel adaptive grid search routine to solve for the optimal design.

Source document verification (SDV), or data auditing, is a common method to assess data quality. SDV involves comparing data from the research database to

clinical source documents, e.g., patient charts, to identify and correct any discrepancies. These methods have long been standard in clinical trials (Weiss, 1998) and are catching on in observational research, as well (Chaulagai et al., 2005; Kimaro and Twaakyondo, 2005; Kiragga et al., 2011; Duda et al., 2012; Mphatswe et al., 2012; Giganti et al., 2019; Lotspeich et al., 2020). The most objective auditors might be external investigators, e.g., from the data coordinating center, but sending auditors to remote locations has its drawbacks, particularly in a multi-national cohort like the Caribbean, Central and South America Network for HIV epidemiology (CCASAnet). In Chapter 3, we propose a creative new protocol where site-level investigators in CCASAnet were trained to audit their own data. Eight clinical sites participated in these "self-audits," and three of them were additionally visited by external auditors for conventional "travel-audits." Using data from the doubly-audited sites (i.e., those who were both self- and travel-audited), we compare audit findings between the protocols, discuss lessons learned, and make recommendations about implementing self-audits in the future.

As biomedical research increasingly turns to secondary data sources like EHR, statistical methods are needed to obtain valid inference from error-prone data, ideally without sacrificing the high power of large datasets. In addition, relationships between error-prone and error-free variables add complexity to how we correct for them, especially when modeling the errors directly. Full-likelihood approaches have been proposed for logistic regression in two-phase studies (Tang et al., 2015), but they make many parametric assumptions and are limited to binary misclassified covariates. In Chapter 4, we propose a semiparametric likelihood approach to estimate odds ratios that uses all information from a two-phase study and accommodates a number of error mechanisms. Our approach handles error-prone covariates that can be categorical or continuous and leaves their distributions completely unspecified. In addition, the selection of the Phase II sample can depend on Phase I data in an arbitrary manner.

CHAPTER 2

OPTIMAL MULTI-WAVE VALIDATION FOR SECONDARY USE DATA WITH OUTCOME AND EXPOSURE MISCLASSIFICATION

2.1 Introduction

The ever-growing trove of patient information in observational databases, like electronic health records (EHR), provides unprecedented opportunities for biomedical researchers to investigate associations of scientific and clinical interest. However, these data are usually large and error-prone since they are "secondary use data," i.e., they were not primarily created for research purposes (Safran et al., 2007; Hersh et al., 2013; Kim et al., 2019). Ignoring the errors can yield biased results (Green, 2013; Chen et al., 2019; Giganti et al., 2019), and the interpretation, dissemination, or implementation of such results can be detrimental to the very patients whom the analysis sought to help.

To assess the quality of secondary use data, validation studies have been carried out wherein trained auditors compare clinical source documents (e.g., paper medical records) to database values and note any discrepancies between them (Duda et al., 2012). The Vanderbilt Comprehensive Care Clinic (VCCC) is an outpatient facility in Nashville, Tennessee that provides care for people living with HIV/AIDS (PLWH). Since investigators at the VCCC extract EHR data for research purposes, the VCCC validates all key study variables for all patients in the EHR. The VCCC data have demonstrated the importance of data validation, as estimates using the fully-validated data often differ substantially from those using the original unvalidated data extracted from the EHR (Oh et al., 2018; Giganti et al., 2020).

However, validating entire databases can be cost-prohibitive and often unattainable: in the VCCC, full-database validation of approximately 4000 patients costs over US$60,000 annually. A cost-effective alternative is to implement a two-phase design (White, 1982), or partial data audit, under which one collects the original error-prone data in Phase I and then uses this Phase I information to select a subset of records for validation/auditing in Phase II. This type of design greatly reduces the cost associated with data validation and has been implemented in cohorts using routinely collected data, like the Caribbean, Central, and South America network for HIV Epidemiology (CCASAnet) (McGowan et al., 2007).

CCASAnet is a large ($\sim$50,000 patients), multi-national HIV clinical research collaboration. Clinical sites in CCASAnet routinely collect important clinical variables,

and these site-level data are subsequently compiled into a collaborative CCASAnet database that is used for research. One interesting question for CCASAnet investigators is whether patients treated for tuberculosis (TB) are more likely to have better treatment outcomes if their TB diagnosis was bacteriologically confirmed. TB is difficult to diagnose and treat among PLWH, and some studies suggest that those treated for TB without a definitive diagnosis are more likely to subsequently die (Crabtree-Ramirez et al., 2019). Key study variables are available in or can be derived from the CCASAnet database, but both the outcome and exposure, successful treatment completion and bacteriological confirmation, respectively, can be misclassified in the database. For more than a decade, the CCASAnet Data Coordinating Center has performed partial data audits to ensure the integrity of its database (Duda et al., 2012; Giganti et al., 2019; Lotspeich et al., 2020), and plans are currently underway to validate these TB study variables on a subset of records in the near future. Site-stratified random sampling has been the most common selection mechanism for audits thus far, including a previous audit of the TB variables in 2009–2010. Now, we are interested in developing optimal designs that select subjects who are most informative about the association between bacteriologic confirmation and treatment completion.

Statistical methods have been proposed to combine Phase I and Phase II data from two-phase studies with binary outcome misclassification and covariate error. These methods can largely be grouped into likelihood- or design-based estimators. The former include the maximum likelihood estimator (MLE) (Tang et al., 2015) and semiparametric maximum likelihood estimator (SMLE) (Lotspeich et al., 2021), while the latter include the inverse probability weighted (IPW) estimator (Horvitz and Thompson, 1952), generalized raking/augmented IPW estimator (Deville et al., 1993; Robins et al., 1994; Lumley et al., 2011), and the mean score estimator (Reilly and Pepe, 1995). Likelihood-based estimators tend to be more efficient, while design-based estimators can be more robust.

Given the resource constraints imposed upon data audits, efficient designs that target the most informative patients are in high demand. Optimal designs have been derived for binary outcomes with error-prone covariates but not yet for binary outcome misclassification in addition to covariate error. Thus far, closed-form solutions exist for the optimal sampling proportions under covariate error for some design-based approaches, including the IPW and mean score estimators (Reilly and Pepe, 1995; McIsaac and Cook, 2014; Chen and Lumley, 2020). Optimal designs for likelihood-based estimators have also been considered, although the variance of these estimators

does not lend itself to a closed-form solution unless additional assumptions are made (Breslow and Cain, 1988; Holcroft and Spiegelman, 1999; McIsaac and Cook, 2014; Tao et al., 2020). While likelihood-based estimators can still gain efficiency under design-based optimal designs (McIsaac and Cook, 2014; Amorim et al., 2021), they will be most efficient under designs that are optimal for likelihood-based estimators.

Regardless of the estimator, optimal designs share common challenges; in particular, they require specification of unknown parameters. To overcome this, multi-wave designs have been proposed that estimate the unknown parameters with an internal pilot study and then use this information to approximate the optimal designs (McIsaac and Cook, 2015; Chen and Lumley, 2020; Han et al., 2020). Instead of selecting one Phase II subsample, multi-wave designs allow iterative selection of two or more waves of Phase II. This way, each wave gains insight from those that came before it. So far, multi-wave designs have only been used to adapt optimal designs for design-based estimators and under settings with covariate error alone. We focus on designing multi-wave validation studies to improve the statistical efficiency of likelihood-based estimators under the unaddressed setting of outcome and exposure misclassification.

Based on the asymptotic properties of the two-phase MLE for logistic regression, we derive the optimal validation study design to minimize the variance of the log odds ratio (OR) under differential outcome and exposure misclassification. In the absence of a closed-form solution, we devise an adaptive grid search algorithm. Since it is a function of unknown parameters, we introduce a two-wave approximation to the optimal design that can be implemented in practice. Through extensive simulations, the proposed optimal designs are compared to case-control or balanced case-control sampling for a variety of error settings. Notable gains in efficiency can be seen not only with the optimal design, but also with the two-wave approximation to it. Using the VCCC data, we compare the various designs by examining the efficiency of validating different subsets of the EHR data and comparing results from two-phase analyses to those from the full-cohort analysis using fully-validated data. Finally, we implement our approach to design the next round of CCASAnet audits.

2.2 Methods

2.2.1 Model and Data

Consider a binary outcome, Y, binary exposure, X, and covariates $\boldsymbol{Z}$ which are assumed to be related through the logistic regression model $P(Y = 1|X, \boldsymbol{Z}) = [1 +$

$\exp\{-(\beta_0 + \beta X + \boldsymbol{Z}\boldsymbol{\beta}_z)\}]^{-1}$. Instead of Y and X, error-prone measures denoted Y^* and X^*, respectively, are available, i.e., from an observational database; covariates $\boldsymbol{Z}$ are also available and error-free. Fortunately, n of the N subjects ($n < N$) will have their data validated through auditing. The validation indicator $V_i = 1$ if subject i ($i = 1, \ldots, N$) is audited and $V_i = 0$ otherwise. The joint probability of a complete observation is $P(V, Y^*, X^*, \boldsymbol{Z}, Y, X)$

$$= P(V|Y^*, X^*, \boldsymbol{Z})P(Y^*|X^*, \boldsymbol{Z}, Y, X)P(X^*|Y, X, \boldsymbol{Z})P(Y|X, \boldsymbol{Z})P(X|\boldsymbol{Z})P(\boldsymbol{Z}), \quad (2.1)$$

where $P(V|Y^*, X^*, \boldsymbol{Z})$ is the validation sampling probability; $P(Y|X, \boldsymbol{Z})$ is the logistic regression model of primary interest; $P(Y^*|X^*, \boldsymbol{Z}, Y, X)$ and $P(X^*|\boldsymbol{Z}, Y, X)$ are the outcome and exposure misclassification mechanisms, respectively; $P(X|\boldsymbol{Z})$ is the conditional probability of X given $\boldsymbol{Z}$; and $P(\boldsymbol{Z})$ is the marginal density of $\boldsymbol{Z}$. Sampling (i.e., V) is assumed to depend only on Phase I variables (Y^*, X^*, $\boldsymbol{Z}$), so (Y, X) are missing at random (MAR) for unaudited subjects (Little and Rubin, 2002). Equation (2.1) captures the most complex differential misclassification in both the outcome and exposure, but addresses other common settings as special cases. For classical scenarios of outcome or exposure misclassification alone, set $X^* = X$ or $Y^* = Y$, respectively. Nondifferential misclassification can be assumed if $P(Y^*|X^*, Y, X, \boldsymbol{Z}) = P(Y^*|Y, \boldsymbol{Z})$ or $P(X^*|Y, X, \boldsymbol{Z}) = P(X^*|X, \boldsymbol{Z})$ (Keogh et al., 2020).

All observations, $(V_i, Y_i^*, X_i^*, \boldsymbol{Z}_i, Y_i, X_i)$ ($i = 1, \ldots, N$), are assumed to be i.i.d. following equation (2.1). The necessary unknowns in equation (2.1) – specifically, $P(Y_i^*|X_i^*, Y_i, X_i, \boldsymbol{Z}_i)$, $P(X_i^*|Y_i, X_i, \boldsymbol{Z}_i)$, and $P(X_i|\boldsymbol{Z}_i)$ – are assumed to follow additional logistic regression models. All model parameters are denoted together by $\boldsymbol{\theta}$; since we focus on estimating β, all other nuisance parameters are denoted by $\boldsymbol{\eta}$ and $\boldsymbol{\theta} = (\beta, \boldsymbol{\eta}^T)^T$. Given that (Y_i, X_i) are incompletely observed, the observed-data log-likelihood for $\boldsymbol{\theta}$ is $l_N(\boldsymbol{\theta}) =$

$$\begin{aligned}
&\sum_{i=1}^{N} V_i \log\left\{P(Y_i^*|X_i^*, \boldsymbol{Z}_i, Y_i, X_i)P(X_i^*|\boldsymbol{Z}_i, Y_i, X_i)P(Y_i|X_i, \boldsymbol{Z}_i)P(X_i|\boldsymbol{Z}_i)\right\} \\
&+ \sum_{i=1}^{N}(1 - V_i) \log\left\{\sum_{y=0}^{1}\sum_{x=0}^{1} P(Y_i^*|X_i^*, \boldsymbol{Z}_i, y, x)P(X_i^*|\boldsymbol{Z}_i, y, x)P(y|x, \boldsymbol{Z}_i)P(x|\boldsymbol{Z}_i)\right\}.
\end{aligned} \quad (2.2)$$

The distribution of V can be omitted because the Phase II variables are MAR. The fully-parametric MLE (Tang et al., 2015) can be obtained by maximizing equation (2.2). In deriving the optimal design, we wish to obtain the most efficient estimator

of β, the conditional log OR for X on Y.

2.2.2 Optimal Design

Under standard MLE theory, we have $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightsquigarrow \boldsymbol{N}_d(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1})$, where $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\boldsymbol{\eta}}^T)^T$ are the MLE, $\boldsymbol{\theta}$ denotes the true parameter values, $\rightsquigarrow$ represents convergence in distribution, and $\boldsymbol{N}_d(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta})^{-1})$ is a multivariate normal distribution centered at $\mathbf{0}$ with variance equal to the inverse of the Fisher information. The Fisher information is defined as $\mathcal{I}(\boldsymbol{\theta})$

$$= E\left\{S_i(\boldsymbol{\theta})S_i(\boldsymbol{\theta})^T\right\} = \begin{bmatrix} E\left\{S_i(\beta)^2\right\} & E\left\{S_i(\beta)S_i(\boldsymbol{\eta})^T\right\} \\ E\left\{S_i(\beta)S_i(\boldsymbol{\eta})\right\} & E\left\{S_i(\boldsymbol{\eta})S_i(\boldsymbol{\eta})^T\right\} \end{bmatrix} = \begin{bmatrix} \mathcal{I}(\beta,\beta) & \mathcal{I}(\beta,\boldsymbol{\eta})^T \\ \mathcal{I}(\beta,\boldsymbol{\eta}) & \mathcal{I}(\boldsymbol{\eta},\boldsymbol{\eta}) \end{bmatrix},$$

where $S_i(\boldsymbol{\theta})^T = (S_i(\beta), S_i(\boldsymbol{\eta})^T)^T$ $(i = 1, \ldots, N)$ is the score vector for a single observation based on expression (2.2). We wish to minimize $Var(\hat{\beta})$ with the optimal design, which can be expressed as

$$Var(\hat{\beta}) = N^{-1}\left\{\mathcal{I}(\boldsymbol{\theta})^{-1}\right\}_{[1,1]} = N^{-1}\left\{\mathcal{I}(\beta,\beta) - \mathcal{I}(\beta,\boldsymbol{\eta})^T\mathcal{I}(\boldsymbol{\eta},\boldsymbol{\eta})^{-1}\mathcal{I}(\beta,\boldsymbol{\eta})\right\}^{-1}, \quad (2.3)$$

as long as $\mathcal{I}(\boldsymbol{\theta})$ is invertible and the models are correctly specified.

The elements of $\mathcal{I}(\boldsymbol{\theta})$ are expectations taken with respect to the complete data, which allows us to express them as functions of the sampling probabilities, $\pi_{y^*x^*z} \equiv P(V = 1|Y^* = y^*, X^* = x^*, Z = z)$, and model parameters, $\boldsymbol{\theta}$. To demonstrate, consider the element $\mathcal{I}(\theta_j, \theta_{j'})$ $(\theta_j, \theta_{j'} \in \boldsymbol{\theta})$

$$\begin{aligned} &= \sum_{y^*=0}^{1}\sum_{x^*=0}^{1}\sum_{z=1}^{q}\pi_{y^*x^*z}\sum_{y=0}^{1}\sum_{x=0}^{1}S^v(\theta_j; y^*, x^*, z, y, x)S^v(\theta_{j'}; y^*, x^*, z, y, x)P(y^*, x^*, z, y, x) \\ &\quad + \sum_{y^*=0}^{1}\sum_{x^*=0}^{1}\sum_{z=1}^{q}(1 - \pi_{y^*x^*z})S^{\bar{v}}(\theta_j; y^*, x^*, z)S^{\bar{v}}(\theta_{j'}; y^*, x^*, z)\sum_{y=0}^{1}\sum_{x=0}^{1}P(y^*, x^*, z, y, x), \end{aligned} \quad (2.4)$$

where $S^v(\cdot)$ and $S^{\bar{v}}(\cdot)$ are the score functions of validated and unvalidated subjects, respectively (see Section 2.7.1 for definitions). For notational simplicity, our derivations are based on designing validation studies with Y^*, X^*, and a q-level categorical Z, which may include all observed combinations of multiple categorical covariates or discretized continuous Z. Note that if we discretize continuous Z to define strata for sampling, we would still retain its continuous value for inference.

Since $\boldsymbol{\theta}$ is fixed, we see from equation (2.4) that the efficiency of the MLE can only

be improved through how we define the sampling probabilities. Thus, the optimal design will be the one that chooses $\{\pi_{y^*x^*z}\}$ to minimize the asymptotic variance of the MLE through the elements defined in expression (2.4). The size of the audit, n, is assumed to be constrained by budget, time, or other practicalities. This is expressed as

$$n = \left(\sum_{y^*=0}^{1} \sum_{x^*=0}^{1} \sum_{z=1}^{q} \pi_{y^*x^*z} N_{y^*x^*z} \right), \tag{2.5}$$

where $N_{y^*x^*z}$ is the observed size of the Phase I stratum with ($Y_i^* = y^*$, $X_i^* = x^*$, $Z_i = z$). Constrained optimization of $Var(\hat{\beta})$, e.g., with Lagrange multipliers for the audit size constraint, does not yield a closed-form solution. Therefore, we devise a novel grid search algorithm to find the optimal values of $\{\pi_{y^*x^*z}\}$.

2.2.3 Adaptive Grid Search

The challenge at hand is one of combinatorics: of all the possible designs that satisfy the audit size constraint and are supported by the available Phase I data (i.e., the stratum sizes $N_{y^*x^*z}$), which minimizes $Var(\hat{\beta})$? To answer this, we have developed an adaptive grid search algorithm, where a series of grids are constructed at iteratively tighter scales and over more focused grid spaces, to locate the optimal design. "Grid" refers to the collection of all possible audit designs. In iteration t, the grid $\boldsymbol{G}^{(t)}$ can be represented by a matrix with columns for each stratum size such that each row is a potential design (Figure 2.1). The adaptive nature of our algorithm is necessitated by the computational strain of this grid, whose dimension increases with the Phase I–II sample sizes and the number of sampling strata.

The minimum stratum size, m, is introduced for stability of the MLE; constraints like this are needed to avoid degenerate optimal designs (Breslow and Cain, 1988). Let K denote the number of strata used in sampling. For the initial grid search, we consider audits made up of stratum sizes between the minimum and maximum allocations, m and $(n - Km)$, respectively. Stratum sizes are incremented by $s^{(1)}$ subjects between designs; given the large scope of this search area, we choose $s^{(1)}$ to be the largest grid scale. For all successive grids, the search space focuses around the previous iteration's lowest-variance design, denoted $\{n_{y^*x^*z}^{(t-1)}\}$. The previous step size, $s^{(t-1)}$, determines the window around $\{n_{y^*x^*z}^{(t-1)}\}$ to be searched, with stratum sizes between $n_{y^*x^*z}^{(t-1)} - s^{(t-1)}$ and $n_{y^*x^*z}^{(t-1)} + s^{(t-1)}$ considered at a smaller scale of $s^{(t)}$ subjects ($s^{(t)} < s^{(t-1)}$). These steps are detailed below.

(a) Matrix representation

$\boldsymbol{G}^{(1)}$

Design	n_{01}	n_{10}	n_{11}
1	10	10	10
⋮	⋮	⋮	⋮
2532	100	85	190
2533	**115**	**85**	**190**
2543	10	100	190
⋮	⋮	⋮	⋮
2925	10	10	370

$\boldsymbol{G}^{(2)}$

Design	n_{01}	n_{10}	n_{11}
1	130	70	175
⋮	⋮	⋮	⋮
85	110	85	190
86	**115**	**85**	**190**
87	100	90	190
⋮	⋮	⋮	⋮
134	100	85	205

$\boldsymbol{G}^{(3)}$

Design	n_{01}	n_{10}	n_{11}
1	120	80	185
⋮	⋮	⋮	⋮
354	113	84	191
255	**114**	**84**	**191**
356	115	84	191
⋮	⋮	⋮	⋮
491	110	85	195

(b) Graphical representation

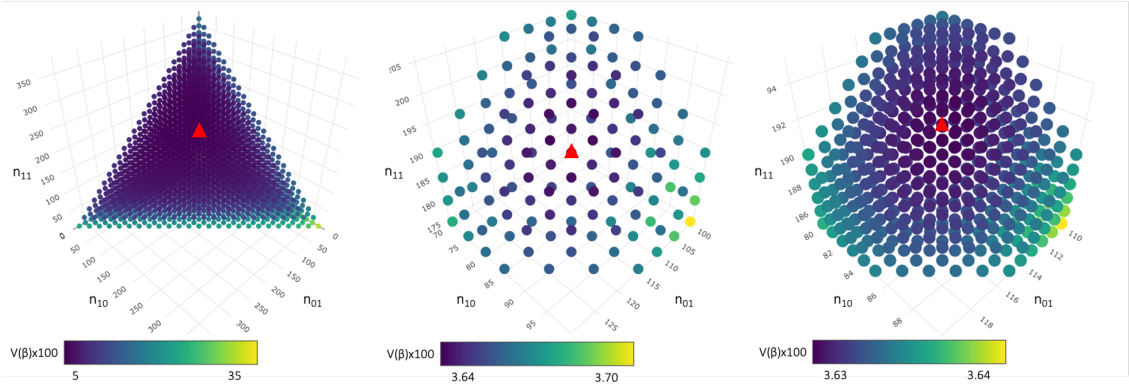


Figure 2.1: Matrix (a) and graphical (b) representations of a three-step adaptive grid search with audit size $n = 400$; minimum stratum size $m = 10$; and grid scales of $s^{(1)} = 15$, $s^{(2)} = 5$, and $s^{(1)} = 1$ subject(s). In (a), the bold row indicates the design achieving the lowest $Var(\hat{\beta})$; in (b) the triangle does. n_{00} can be omitted because it is determined by the audit size constraint.

1. Construct the grid $\boldsymbol{G}^{(t)}$ of possible audits from stratum sample sizes $n_{y^*x^*z}$ within the search window, varying by increments of $s^{(t)}$ subjects between designs.

   1.a. If $t = 1$, $\boldsymbol{G}^{(t)}$ is comprised of all combinations of stratum sample sizes between $[m, (n - Km)]$ in increments of $s^{(t)}$ that satisfy equation (2.5). If $N_{y^*x^*z} < m$ or $N_{y^*x^*z} < (n - Km)$, stratum sizes are adjusted accordingly.

   1.b. If $t > 1$, $\boldsymbol{G}^{(t)}$ is made up of all combinations of stratum sizes between $[n^{(t-1)}_{y^*x^*z} - s^{(t-1)}, n^{(t-1)}_{y^*x^*z} + s^{(t-1)}]$ in increments of $s^{(t)}$ that satisfy equation (2.5). If $n^{(t-1)}_{y^*x^*z} - s_{(t-1)} < m$ or $n^{(t-1)}_{y^*x^*z} + s_{(t-1)} > N_{y^*x^*z}$, strata are restricted accordingly.

2. Based on $\boldsymbol{\theta}$ and $\{\pi_{y^*x^*z} = n_{y^*x^*z}/N_{y^*x^*z}\}$, evaluate equation (2.3) for each row of $\boldsymbol{G}^{(t)}$ to estimate $Var(\hat{\beta})$ under each design.

3. Let $\{n^{(t)}_{y^*x^*z}\}$ denote the design that achieved the minimum values for $Var(\hat{\beta})$.

4. Repeat steps 1–3, reducing the scale of the grid in each iteration until $s^{(t)} = 1$.

The lowest-variance design from the final iteration is the optimal design, the optMLE. In rare situations, the grid search may be stuck in some local area of designs with extremely imbalanced strata sizes, rendering singular or nearly-singular information matrices despite the minimum stratum size requirement. In this situation, we need to retune the grids in the previous iteration or restart the algorithm using a different grid.

2.2.4 Two-Wave Approximate Optimal Design

Clearly, the optimal design derived in the Section 2.2.2 relies on the model parameters $\boldsymbol{\theta}$, which are unknown. Thus, application of the optMLE design in practice requires reliable estimates of these parameters. If available, historical data from a previous audit could be used to estimate these parameters. Otherwise, we propose a two-wave design. Whereas traditional two-phase studies require all of the information up front (at Phase I), multi-wave designs allow sampling to adapt as information accumulates.

We separate the validation study into two waves, labeled Phase II(a) and Phase II(b), respectively, and denote their corresponding sample sizes as $n^{(a)}$ and $n^{(b)}$ ($n^{(a)} + n^{(b)} = n$). Fully-adaptive designs have been considered elsewhere wherein, after an initial wave of Phase II, the design is re-approximated with each new person sampled. However, two waves with a 50/50 split between them were seen to be sufficient (McIsaac and Cook, 2015). Following this discussion, we choose $n^{(a)} = n/2$ subjects in Phase II(a); selection of these subjects will be through balanced case-control sampling of the Phase I data if no prior information is available. The unknown parameters can then be estimated following collection of validated data on these subjects. Therefore, the optimal design can be approximated to determine the allocation of the remaining subjects in Phase II(b). This is our two-wave approximate optimal design, the optMLE-2.

2.3 Simulations

2.3.1 Validation Study Designs

In the simulations that follow, we compare the performance of five audit designs in two-phase analyses under differential outcome and exposure misclassification. Since optimal designs have not yet been proposed for this setting, the proposed designs are compared to existing case-control and balanced case-control designs based on the error-prone ("unvalidated") Phase I data. While the total size of the validation study is the same, allocation of subjects between the Phase I strata differs between designs.

Simple random sampling (SRS): All subjects in Phase I have equal probability of inclusion in Phase II.

Unvalidated case-control sampling (CC*) Subjects are stratified on Y^* and separate random samples of size $n/2$ are drawn from each stratum (Tosteston and Ware, 1990).

Unvalidated balanced case-control sampling (BCC*) Subjects are jointly stratified on (Y^*, X^*) and separate random samples of size $n/4$ subjects are drawn from each stratum (Breslow and Cain, 1988; Tosteston and Ware, 1990).

Optimal design (optMLE) Subjects are jointly stratified on (Y^*, X^*), and stratum sizes are chosen following Section 2.2.2. The optMLE design is included as a "gold standard" design since it requires knowing the parameters $\boldsymbol{\theta}$.

Two-wave approximate optimal design (optMLE-2) Subjects are jointly stratified on (Y^*, X^*). In the first wave, $n/2$ subjects are selected using BCC*, and in the second wave the remaining subjects are chosen following the design in Section 2.2.2.

All simulations include differential outcome and exposure misclassification, and Section 2.3.3 extends to include an additional error-free covariate.

Designs are compared based on two precision measures: the relative efficiency (RE), defined as the ratio of empirical variances of parameter estimates, and the relative interquartile range (RI), defined as the ratio of the width of the empirical interquartile range (McIsaac and Cook, 2015). The optimal design based on true parameter values and observed stratum sizes was treated as the reference standard. Optimal designs based on (a) true parameter values and expected stratum sizes or (b) full cohort parameter estimates and observed stratum sizes were also considered, but results were similar (Table 2.5). RE and RI values > 1 indicate better precision than the optMLE design while values < 1 indicate worse.

2.3.2 Outcome and Exposure Misclassification

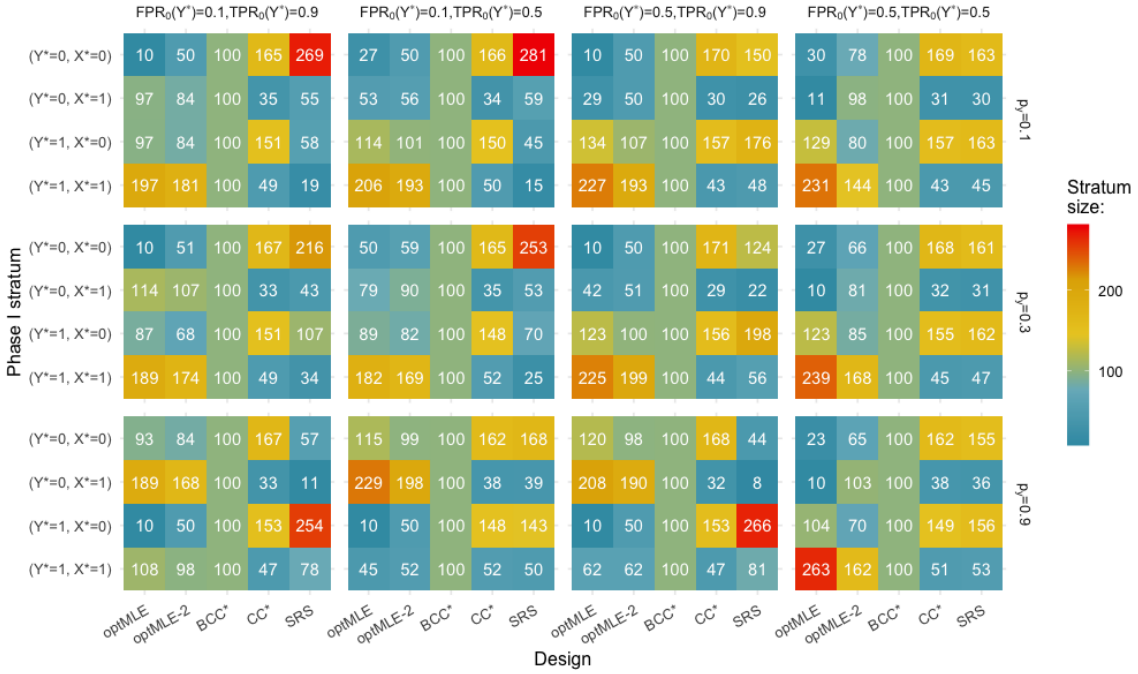
Data were generated for a Phase I sample of $N = 10{,}000$ subjects according to equation (2.1). True X and Y were generated from Bernoulli distributions with $p_x = P(X = 1)$ and $P(Y = 1|X) = [1 + \exp\{-(\beta_0 + 0.3X)\}]^{-1}$. The approximate outcome prevalence $p_y = P(Y = 1|X = 0)$ was used to define $\beta_0 = \log\{p_y/(1 - p_y)\}$. Error-prone Y^* and X^* were generated from Bernoulli distributions with $P(X^* = 1|Y, X) = [1 + \exp\{-(\gamma_0 + 0.45Y + \gamma_1 X)\}]^{-1}$ and $P(Y^* = 1|X^*, Y, X) = [1 + \exp\{-(\alpha_0 + 0.275X^* + \alpha_1 Y + 0.275X)\}]^{-1}$, where (γ_0, γ_1) and (α_0, α_1) control the strength of the relationship between error-prone and error-free values. We define the "baseline" false positive and true positive rates for X^*, denoted $FPR_0(X^*)$ and $TPR_0(X^*)$, respectively, as the false positive and true positive rates of X^* when $Y = 0$. Similarly, $FPR_0(Y^*)$ and $TPR_0(Y^*)$ are the false positive and true positive rates for Y^* when $X = X^* = 0$. With these definitions, we have $\alpha_0 = -\log\left\{\frac{1-FPR_0(Y^*)}{FPR_0(Y^*)}\right\}$, $\alpha_1 = -\log\left\{\frac{1-TPR_0(Y^*)}{TPR_0(Y^*)}\right\} - \alpha_0$, $\gamma_0 = -\log\left\{\frac{1-FPR_0(X^*)}{FPR_0(X^*)}\right\}$, and $\gamma_1 = -\log\left\{\frac{1-TPR_0(X^*)}{TPR_0(X^*)}\right\} - \gamma_0$. When FPR_0 and TPR_0 are both 0.5, the error-prone values are not informative about error-free ones. Using the designs in Section 2.3.1, $n = 400$ subjects were selected in Phase II. Minimum stratum sizes of $m = 10$–50 were considered for the optMLE design (Figure 2.6); all yielded stable estimates, so $m = 10$ was used for all simulations herein.

2.3.2.1 Varied outcome misclassification rates

We fixed exposure misclassification rates at $FPR_0(X^*) = 0.1$ and $TPR_0(X^*) = 0.9$ and varied outcome misclassification rates between combinations of $FPR_0(Y^*) = 0.1, 0.5$ and $TPR_0(Y^*) = 0.9, 0.5$. We also varied $p_y = 0.1, 0.3, 0.9$ for fixed $p_x = 0.1$. Designs are illustrated in Figure 2.2(a). The composition of the optMLE design depended on the frequencies of Phase I strata and misclassification rates. It generally favored oversampling from the smaller strata; since $p_x = 0.1$, optMLE design targeted records with $X^* = 1$ and those with $Y^* = 1$ or $Y^* = 0$ when $p_y <$ or > 0.5, respectively. The oversampling of less-frequent Y^* strata was typically heightened in higher-error settings. For example, for $p_y = 0.1$, the optMLE design selected more subjects with $(Y^* = 1, X^* = 1)$ as $FPR_0(Y^*)$ increased and $TPR_0(Y^*)$ decreased. When Y^* was more error-prone than X^*, the optMLE design redistributed more of the audit to the less-frequent strata of Y^* with less emphasis on the less-frequent value of X^*. We note that the initial BCC* sample of $n^{(a)} = 200$ in Phase II(a) kept the optMLE-2 design from being as extreme as the optMLE, but overall the optimal designs were similar. With $FPR_0(Y^*) = TPR_0(Y^*) = 0.5$, the optimal designs

were relatively unchanged by p_y; the optMLE and optMLE-2 designs were also less similar, which we attribute to added uncertainty from estimating parameters from weakly informative Phase I variables (Figure 2.7(a)).

(a) Exposure misclassification rates were fixed: $FPR_0(X^*) = 0.1$, $TPR_0(X^*) = 0.9$.



(b) Outcome misclassification rates were fixed: $FPR_0(Y^*) = 0.1$, $TPR_0(Y^*) = 0.9$.

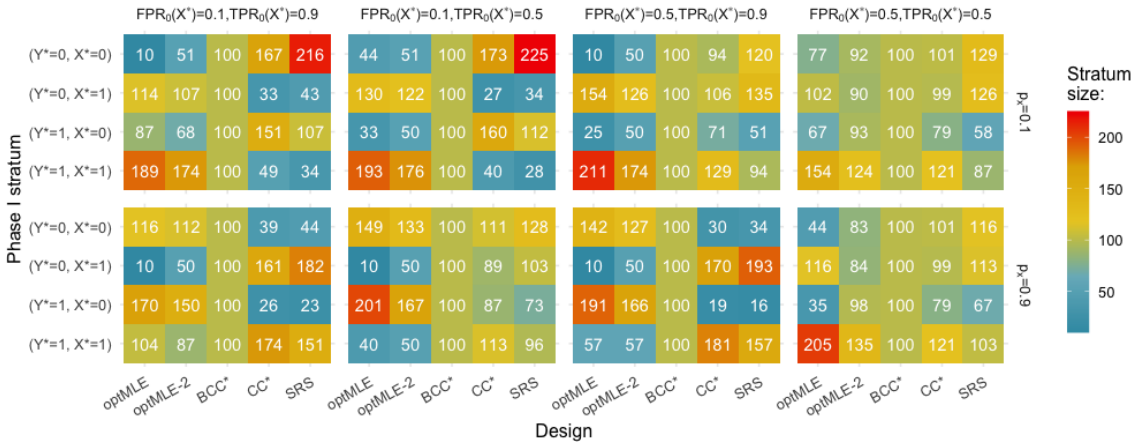


Figure 2.2: Average Phase II stratum sizes $n_{y^*x^*}$ under outcome and exposure misclassification.

The MLE was essentially unbiased, though SRS or CC* designs encountered some extreme replicates (Table 2.6(a)). The empirical standard error (SE), RE, and RI for the MLE under these designs are included in Table 2.1(a). The optMLE-2 design lost little efficiency to the optMLE design with RE > 0.9 and RI > 0.95 in most settings.

We note that the RE and RI for the optMLE-2 design could be > 1 since the optMLE design is asymptotically optimal but not necessarily optimal with finite samples. In most settings the optMLE-2 design exhibited sizeable gains over the BCC*, CC*, and SRS designs, with gains as high as 43%, 74%, and 83%, respectively. The optMLE-2 design saw the largest gains when the error-prone variables were informative (i.e., $FPR_0(Y^*) \neq 0.5$ and/or $TPR_0(Y^*) \neq 0.5$) or Y^* was less balanced (driven by choices of p_y farther from 0.5). The grid search successfully located the optMLE and optMLE-2 designs in all and $\geq 93\%$ replicates per setting, respectively. The grid search failed to locate the optMLE-2 design in a few replicates because it was stuck in some local area of designs that rendered singular information matrices. Ideally, we should retune the grids or restart the algorithm using different grids for these replicates. However, the specific solution needs to be tailored for each problematic replicate separately and thus is time-consuming to implement in large-scale simulations. For ease of implementation, we discarded the few problematic replicates. In practice, we recommend retuning the grids or restarting the algorithm.

2.3.2.2 Varied exposure misclassification error rates

Outcome misclassification rates were fixed at $FPR_0(Y^*) = 0.1$ and $TPR_0(Y^*) = 0.9$, while exposure rates were varied in $FPR_0(X^*) = 0.1, 0.5$ and $TPR_0(X^*) = 0.9, 0.5$. We varied $p_x = 0.1, 0.9$ for fixed $p_y = 0.3$. Interpretations of the optimal designs were very similar to Section 2.3.2.1 but with an emphasis on X^* rather than Y^* (Figure 2.2(b)). Briefly, the optMLE and optMLE-2 designs targeted less-frequent values of Y^* and X^* but oversampled from less-frequent X^* with greater intensity under higher-error settings. In the highest error setting, the optimal designs favored $X^* = 1$ strata for either p_x and the optMLE-2 design was more balanced (Figure 2.7(b)). Tables 2.1(b) and Table 2.6(b) include simulation results for the MLE under these settings. The MLE was always unbiased. The optMLE-2 design remained close in efficiency to the optMLE design and continued to demonstrate notable gains over the BCC*, CC*, or SRS designs, with as much as 30%, 63%, or 69% higher efficiency, respectively.

2.3.3 Outcome and Exposure Misclassification with an Additional Error-Free Covariate

Following equation (2.1), data were generated for a Phase I sample of $N = 10{,}000$ subjects. Error-free binary covariate Z was generated from a Bernoulli distribution

Table 2.1: Simulation results under outcome and exposure misclassification

a) Varied Outcome Misclassification Rates/Prevalence

	Errors in Y^*		optMLE-2			BCC*			CC*			SRS		
p_y	FPR_0	TPR_0	SE	RE	RI	SE	RE	RI	SE	RE	RI	SE	RE	RI
0.1	0.1	0.9	0.214	1.028	1.015	0.254	0.728	0.855	0.347	0.391	0.640	0.516	0.176	0.459
		0.5	0.241	0.908	0.960	0.286	0.643	0.815	0.362	0.403	0.679	0.512	0.201	0.508
	0.5	0.9	0.321	0.935	1.017	0.409	0.578	0.763	0.560	0.308	0.570	0.563	0.305	0.552
		0.5	0.361	1.067	1.008	0.377	0.982	1.004	0.512	0.531	0.767	0.543	0.472	0.700
0.3	0.1	0.9	0.190	1.009	0.983	0.223	0.734	0.855	0.297	0.413	0.683	0.333	0.329	0.569
		0.5	0.219	1.003	1.048	0.226	0.941	1.087	0.317	0.480	0.723	0.344	0.406	0.658
	0.5	0.9	0.241	0.924	0.879	0.274	0.717	0.814	0.386	0.360	0.588	0.357	0.421	0.626
		0.5	0.248	0.918	0.957	0.240	0.982	1.035	0.369	0.416	0.664	0.369	0.415	0.675
0.9	0.1	0.9	0.249	0.927	0.963	0.277	0.749	0.874	0.430	0.310	0.551	0.592	0.164	0.370
		0.5	0.381	0.887	1.040	0.505	0.505	0.774	0.600	0.357	0.663	0.596	0.362	0.660
	0.5	0.9	0.279	0.916	0.966	0.342	0.608	0.826	0.543	0.242	0.517	0.589	0.205	0.443
		0.5	0.491	0.926	0.950	0.515	0.842	0.985	0.620	0.582	0.779	0.608	0.605	0.757

b) Varied Exposure Misclassification Rates/Prevalence

	Errors in X^*		optMLE-2			BCC*			CC*			SRS		
p_x	FPR_0	TPR_0	SE	RE	RI	SE	RE	RI	SE	RE	RI	SE	RE	RI
0.1	0.1	0.9	0.190	1.009	0.983	0.223	0.734	0.855	0.297	0.413	0.683	0.333	0.329	0.569
		0.5	0.218	0.998	0.986	0.247	0.781	0.860	0.336	0.420	0.600	0.338	0.414	0.637
	0.5	0.9	0.295	0.977	1.015	0.351	0.691	0.866	0.342	0.730	0.885	0.351	0.693	0.866
		0.5	0.343	1.020	1.028	0.342	1.028	0.993	0.348	0.993	1.026	0.357	0.942	0.997
0.9	0.1	0.9	0.189	0.851	0.940	0.201	0.750	0.879	0.310	0.316	0.584	0.339	0.265	0.520
		0.5	0.290	0.960	0.910	0.343	0.685	0.811	0.345	0.678	0.771	0.381	0.555	0.750
	0.5	0.9	0.221	0.977	0.920	0.264	0.681	0.838	0.337	0.418	0.600	0.366	0.355	0.576
		0.5	0.364	1.008	0.984	0.366	0.996	0.975	0.360	1.029	0.983	0.387	0.890	0.941

Note: Misclassification rates for X^* and Y^* were fixed at $FPR_0 = 0.1$ and $TPR_0 = 0.9$ in a) and b), respectively. SE is the empirical standard error of the MLE. RE and RI are the empirical relative efficiency and relative interquartile range of the design to the optMLE design, respectively. When $p_y \neq 0.3$, select error settings encountered replicates where the SRS, CC*, or BCC* estimates could be > 5 in magnitude; this happened in $< 1\%$ for any individual setting when $p_y = 0.1$ and $< 5\%$ when $p_y = 0.9$ so these replicates were excluded. All other entries are based on 1000 replicates.

with $P(Z = 1) = p_z = 0.25, 0.5$. True X and Y were generated from Bernoulli distributions with $P(X = 1|Z) = [1+\exp\{-(-2.2+0.5Z)\}]^{-1}$ and $P(Y = 1|X, Z) = [1+\exp\{-(-0.85+0.3X+\beta_z Z)\}]^{-1}$, for $\beta_z = -0.25, 0, 0.25$. Baseline misclassification rates were fixed at $FPR_0 = 0.25$ and $TPR_0 = 0.75$ such that X^* and Y^* were generated from Bernoulli distributions with $P(X^* = 1|Y, X, Z) = [1 + \exp\{-(-1.1 + 0.45Y + 2.2X + \lambda Z)\}]^{-1}$ and $P(Y^* = 1|X^*, Y, X, Z) = [1 + \exp\{-(-1.1 + 0.275X^* + 2.2Y + 0.275X + \lambda Z)\}]^{-1}$, where $\lambda = -1, 0, 1$. In Phase II, $n = 400$ subjects were selected by extensions of Section 2.3.1 to sample on (Y^*, X^*, Z). The optMLE design sampled ≥ 10 subjects from each stratum.

Typical Phase II stratum sizes for the designs when $p_z = 0.5$ are depicted in Figure 2.3; designs were similar when $p_z = 0.25$ (Figure 2.8). The optimal designs favored subjects with $Z = 1$, sampling minimally from the $Z = 0$ strata regardless

of the values for (Y^*, X^*). This was partly because $\text{Var}(X|Z = 1)$ was larger than $\text{Var}(X|Z = 0)$, such that the true value of X was harder to "guess" when $Z = 1$, and validating X among subjects with $Z = 1$ was more "rewarding" than validating X among subjects with $Z = 0$. Within the $Z = 1$ strata, the optimal designs preferred subjects with $(Y^* = 1, X^* = 1)$, in alignment with Section 2.3.2.

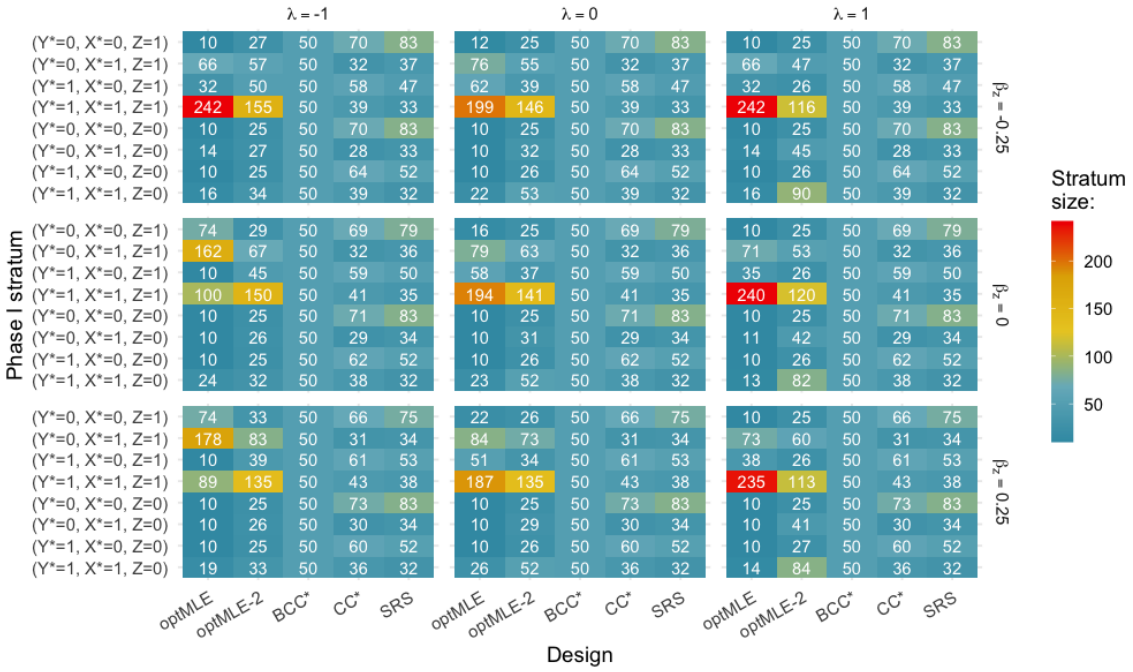


Figure 2.3: Average Phase II stratum stratum sizes $n_{y^*x^*z}$ under outcome and exposure misclassification when additional error-free covariate information was included in sampling. Error-free covariate Z had prevalence $p_z = 0.5$.

Simulation results for the MLE are included in Tables 2.2 and 2.7. The MLE was always unbiased. The optMLE-2 design was competitively efficient to the optMLE design and surpassed the BCC*, CC*, or SRS designs, with efficiency gains as high as 43%, 56%, or 59%, respectively. Gains were more pronounced in settings with $\lambda \leq 0$, such that $Z = 1$ decreased the probability that $Y^* = 1$, or Z was less common.

2.3.4 Special Cases of Misclassification

Special cases with outcome or exposure misclassification alone are illustrated with additional simulations in Sections 2.7.2.1 and 2.7.2.2, respectively. The optimal designs targeted the less-frequent value of the error-prone variable, with little regard for the value of the error-free variable (Figures 2.4 and 2.5). In both settings, the optMLE-2 design approximated the optMLE design well and continued to offer sizable

Table 2.2: Simulation results under outcome and exposure misclassification with available error-free covariate information

			optMLE-2			BCC*			CC*			SRS		
p_z	λ	β_Z	SE	RE	RI	SE	RE	RI	SE	RE	RI	SE	RE	RI
0.25	−1	−0.25	0.225	1.329	1.246	0.296	0.767	0.835	0.321	0.653	0.833	0.352	0.543	0.770
		0.00	0.227	1.097	1.024	0.288	0.683	0.792	0.325	0.535	0.734	0.332	0.515	0.755
		0.25	0.221	1.036	0.991	0.287	0.613	0.764	0.333	0.454	0.681	0.337	0.444	0.682
	0	−0.25	0.249	0.975	0.960	0.296	0.690	0.790	0.321	0.587	0.788	0.352	0.489	0.729
		0.00	0.245	1.021	0.965	0.288	0.735	0.816	0.325	0.577	0.756	0.332	0.554	0.778
		0.25	0.242	0.879	0.925	0.287	0.625	0.787	0.333	0.463	0.701	0.337	0.453	0.702
	1	−0.25	0.267	0.943	1.000	0.296	0.767	0.835	0.321	0.653	0.833	0.352	0.543	0.770
		0.00	0.275	0.840	0.940	0.288	0.766	0.857	0.325	0.601	0.794	0.332	0.578	0.816
		0.25	0.270	0.932	1.065	0.287	0.823	0.966	0.333	0.610	0.860	0.337	0.596	0.861
0.50	−1	−0.25	0.229	1.295	1.241	0.304	0.735	0.971	0.310	0.707	0.913	0.338	0.596	0.832
		0.00	0.228	1.088	1.093	0.264	0.811	0.902	0.312	0.579	0.764	0.317	0.559	0.764
		0.25	0.225	0.994	0.972	0.287	0.612	0.776	0.308	0.531	0.685	0.304	0.545	0.715
	0	−0.25	0.250	0.927	1.052	0.304	0.628	0.869	0.310	0.603	0.817	0.338	0.509	0.745
		0.00	0.239	0.924	0.959	0.264	0.762	0.849	0.312	0.545	0.719	0.317	0.526	0.719
		0.25	0.251	0.836	1.010	0.287	0.639	0.839	0.308	0.555	0.740	0.304	0.569	0.773
	1	−0.25	0.277	0.890	1.029	0.304	0.735	0.971	0.310	0.707	0.913	0.338	0.596	0.832
		0.00	0.267	0.991	0.994	0.264	1.014	1.012	0.312	0.725	0.858	0.317	0.700	0.857
		0.25	0.261	0.916	0.939	0.287	0.760	0.867	0.308	0.660	0.765	0.304	0.678	0.799

Note: SE is the empirical standard error of the MLE. RE and RI are the empirical relative efficiency and relative interquartile range of the design to the optMLE design, respectively. Each entry is based on 1000 replicates.

efficiency gains over the BCC*, CC*, and SRS designs.

2.4 Comparing Partial- to Full-Audit Results in the VCCC

The VCCC EHR contains routinely collected patient data including demographics, antiretroviral therapy (ART), labs such as viral load or CD4 count, and clinical events. Since the VCCC data have been fully validated, pre-/post-validation datasets are available. These datasets can be used to compare two-phase designs and analyses that only validate a subset of the records to the gold standard analysis using fully-validated data. We used these data to assess the relative odds of having an AIDS-defining event (ADE) within one year of ART initiation between patients who were/were not ART naive at enrollment, controlling for CD4 (square root transformed) at ART initiation. There were $N = 2012$ records extracted from the EHR, with their unvalidated data suggesting that 73% were ART naive at enrollment and 8% experienced an ADE within one year. There was 6% misclassification in ADE, with 63% false positive rate (FPR) and only 1% false negative rate (FNR); ART naive status at enrollment had 11% misclassification with FPR = 13% and FNR = 3%. Only 19 subjects (1%) had both outcome and exposure misclassification. CD4 count was error-free.

For our two-phase analyses, subjects were first divided into four strata based on

Phase I ADE and ART naive status with stratum sizes $(N_{00}, N_{01}, N_{10}, N_{11}) = (504, 1350, 42, 116)$. Though a full audit was completed, only an artificial subsample of $n = 200$ subjects (10%) were assumed to have validated information available. We used the optMLE-2 design to choose subjects for data validation to maximize the efficiency of the MLE of the ART naive coefficient. The $n^{(a)} = 100$ subjects for Phase II(a) were selected via BCC*. Results using BCC*, CC*, and SRS designs are also included. The process was repeated 1000 times. The SRS, CC*, and BCC* audits chose, on average, Phase II strata of sizes $(n_{00}, n_{01}, n_{10}, n_{11}) = (56, 134, 4, 12)$; (27, 73, 26, 74); and (50, 50, 50, 50), respectively. The grid search successfully located the optMLE-2 design in 95% of replications, with average Phase II stratum sizes $(n_{00}, n_{01}, n_{10}, n_{11}) = (25, 39, 42, 95)$.

The partial-audit estimates were all reasonably close to the fully-validated gold standard and led to the same clinical interpretations: after controlling for $\sqrt{CD4}$ at ART initiation, ART naive status at enrollment was not associated with changes in the odds of ADE within one year of ART initiation (Table 2.3). However, the proposed optMLE-2 design gained 15% efficiency over CC* or BCC* in estimating the relationship between ADE and ART status at enrollment. The SE using SRS was dramatically larger, with the optMLE-2 design over 9× more efficient than SRS. There were also gains in estimating the intercept but not the log OR for $\sqrt{CD4}$. The latter is to be expected since the optMLE-2 design was specifically tuned to minimize the SE for the ART status coefficient. As in simulation, a small number of replicates ($\sim 7\%$) using the SRS led to extreme values and were excluded; see the distributions of estimates in Figure **??**.

We also considered designs that further stratified on error-free CD4 count. Since CD4 count is continuous, we dichotomized it at the median, 238 cells/mm^3. Patients were divided into eight strata, jointly defined on their Phase I ADE, ART status, and CD4 category, of sizes $(N_{000}, N_{010}, N_{100}, N_{110}, N_{001}, N_{011}, N_{101}, N_{111}) = (171, 701, 34, 93, 333, 649, 8, 23)$. The SRS and CC* results were unchanged since these designs do not use covariate information. The BCC* selected $(n_{000}, n_{010}, n_{100}, n_{110}, n_{001}, n_{011}, n_{101}, n_{111}) = (28, 28, 28, 29, 28, 28, 8, 23)$ subjects in Phase II. The grid search successfully located the optMLE-2 design 95% of the time, selecting $(n_{000}, n_{010}, n_{100}, n_{110}, n_{001}, n_{011}, n_{101}, n_{111}) = (14, 32, 33, 70, 13, 13, 8, 16)$ in Phase II; it seemed to favor patients with below-average CD4 counts. In Table 2.3, we see that further stratification on CD4 count offered little improvement to design efficiency.

Table 2.3: log OR and standard errors from the analysis of the VCCC dataset

	(Intercept)		ART Status		$\sqrt{CD4}$	
Design	**log OR**	**SE**	**log OR**	**SE**	**log OR**	**SE**
Full cohort analysis						
Gold standard	−1.184	0.294	0.032	0.260	−0.180	0.022
Naive	−0.043	0.234	−0.308	0.200	−0.148	0.015
Two-phase analysis						
	Sampling on ADE and ART Status					
SRS	−1.044	0.821	−0.163	1.173	−0.189	0.058
CC*	−1.536	0.439	0.092	0.395	−0.148	0.034
BCC*	−1.336	0.419	0.001	0.394	−0.168	0.035
optMLE-2	−1.542	0.394	0.118	0.368	−0.151	0.035
	Sampling on ADE, ART Status, and CD4 Count					
BCC*	−1.402	0.421	0.115	0.406	−0.163	0.034
optMLE-2	−1.495	0.392	0.105	0.362	−0.150	0.034

Note: log OR and SE are, respectively, the empirical means of the log odds ratio and the empirical standard error estimates of the MLE. Each two-phase analysis was repeated 1000 times; full cohort analyses were run once. The Naive analysis used Phase I data only. The SRS and CC* designs are the same whether CD4 Count was included in sampling or not.

2.5 Prospective Audit Planning in CCASAnet

Researchers are interested in assessing the association between bacteriological confirmation of TB and successful treatment outcomes among PLWH who are treated for TB. We are in the process of designing a multi-site audit of $n = 500$ patients to validate key variables and better estimate this association in the CCASAnet cohort. The outcome of interest (Y) is successful completion of TB treatment within one year of diagnosis; among patients who did not complete treatment, this captures unfavorable outcomes of death, TB recurrence, or loss to follow-up (with each of these outcomes also of interest in secondary analyses). Bacterial confirmation (X) is defined as any positive diagnostic test result, e.g., culture, smear, or PCR. The CCASAnet database contains error-prone values (Y^*, X^*) of these variables.

The Phase I sample comes from the current CCASAnet research database and includes all patients initiating TB treatment between January 1, 2010 and December 31, 2018. There were $N = 3478$ TB cases across sites in five countries (anonymously

labeled Countries A–E) during this period. Patients were stratified on (Y^*, X^*) within Countries A–E to create 20 strata of sizes $(N_{00}, N_{01}, N_{10}, N_{11}) = (704, 246, 1015, 415)$; (239, 139, 336, 218); (3, 7, 5, 17); (6, 9, 15, 14); and (12, 16, 36, 26), respectively.

To implement the optMLE-2 design as in Sections 2.3–2.4, $n^{(a)} = 250$ patients would be chosen in Phase II(a) using BCC* from the 20 $(Y^*, X^*,$ Country) strata. Site-level designs (Y^*, X^*) would be $(n_{00}^{(a)}, n_{01}^{(a)}, n_{10}^{(a)}, n_{11}^{(a)}) = (12, 12, 13, 13)$ in Countries A, B, and E. Countries C and D were smaller, so we would sample all subjects from them such that $(n_{00}^{(a)}, n_{01}^{(a)}, n_{10}^{(a)}, n_{11}^{(a)}) = (3, 7, 5, 17)$ and $(6, 9, 15, 14)$, respectively. We note that in this case, $n^{(a)} < 250$, but Phase II(a) strata are large enough for stable estimates so the extra subjects can be deferred to Phase II(b). After completing this preliminary audit, the parameter estimates can be used to optimize Phase II(b). However, prior information is available to design Phase II(a), so a naive BCC* design is not necessary.

Between 2009–2010, on-site chart reviews were conducted by external auditors in the five CCASAnet sites. A total of 595 TB cases were chosen for validation via site-stratified SRS. Based on original data, 70% of cases completed treatment within one year and 68% had bacteriological confirmation of TB. Due to time constraints, review of these records was incomplete; validated TB treatment and diagnosis were available for 40 subjects. We observed 13% and 20% misclassification in Y^* (FPR= 7%, FNR= 23%) and X^* (FPR= 39%, FNR= 5%), respectively. No subjects had both their outcome and exposure misclassified. We demonstrate two ways to use these historic audits to design an optimal validation study for the next round of CCASAnet audits: (i) to derive the optMLE design to allocate all $n = 500$ subjects in a single Phase II subsample or (ii) to inform the first wave of $n^{(a)} = 250$ for the optMLE-2 design. Strategy (i) puts more trust in the historic audits, while strategy (ii) allows the design to adjust accordingly if Phase II(a) parameters differ from historic ones.

Given the small size of the historic audit, it was not possible to obtain country-level estimates of all parameters needed to derive the optimal design. Instead, we created country groupings (CoG), based on site-specific audit results (Table 2.8). CoG was defined as a three-level categorical variable: CoG = 0 for countries with errors in Y^* or X^* (Countries A–B), CoG = 1 for countries with errors in both Y^* and X^* (Countries C–D), and CoG = 2 for countries without errors (Country E). These groupings were used to obtain the MLE for the historic data. Since audits will be conducted at the site level, we applied these parameters to the 20 Phase I strata from the current data by assuming the same coefficients for countries in a given grouping (Table 2.8).

First, we derived the optMLE design for $n = 500$ based on the historic parameters, setting the minimum stratum size to be $m = 10$. The optimal design was composed of the following (Y^*, X^*) strata from Countries A–E, respectively: $(n_{00}, n_{01}, n_{10}, n_{11})$ = (10, 15, 10, 21); (20, 80, 11, 168); (3, 7 , 5, 17); (6, 9, 15, 14); and (12, 16, 35, 26). All, or nearly all, available subjects were taken from Countries C–E. In Countries A and B, subjects with $X^* = 1$ were preferred, particularly with $Y^* = 1$. We note that the number of records in Country B is smaller, so it was oversampled more than Country A.

Then, we derived the optMLE design for just the first $n^{(a)} = 250$ subjects as a more-informed first wave for the optMLE-2 design. Once again, we assumed that $m = 10$. Site-level (Y^*, X^*) strata of $(n_{00}^{(a)}, n_{01}^{(a)}, n_{10}^{(a)}, n_{11}^{(a)})$ = (10, 10, 10, 10); (10, 10, 10, 45); (3, 7, 5, 17); (6, 9, 10, 13); and (12, 16, 11, 26) from Countries A–E, respectively, were recommended in Phase II(a). This time, audit resources were more focused on sampling from the smallest countries (C–E). Country A was sampled minimally, as was Country B except for the $(Y^* = 1, X^* = 1)$ stratum. Data on these subjects can subsequently be used to re-estimate the model parameters and derive the optimal allocation of the remaining subjects in Phase II(b). Alternatively, the Phase II(a) and historic data could be pooled to re-estimate the parameters; since the historic audits were much smaller, Phase II(a) would likely still dominate the pooled estimates. However, if Phase II(a) were smaller, it might be beneficial to combine the data for more stable estimates.

Ultimately, the choice between these validation designs is determined by logistics and our confidence in the historic data. For the CCASAnet audits, we plan to use the optMLE-2 design informed by the prior audits. This decision is based on our lack of confidence in estimates derived from the prior audits and the results of others who have seen that incorporating prior, even somewhat biased, information can improve multi-wave sampling designs (Chen and Lumley, 2020). Also, multiple waves of validation is feasible in the CCASAnet cohort if performed by in-country investigators (Lotspeich et al., 2020).

2.6 Discussion

Validation studies are integral to many statistical methods to correct for errors in secondary use data, but they are resource-intensive undertakings. This can limit the numbers of records and variables that are reviewed. Thus, selecting the most informative records is key to maximizing the benefits of data validation. We in-

troduced a new optimal design, and a multi-wave approximation to it, which maximizes the efficiency of the MLE under differential outcome and exposure misclassification – a setting for which optimal designs have not yet been proposed. We provide a novel method to minimize the asymptotic variance of a full-likelihood estimator, which has no analytical solution; our adaptive grid search is provided in the *auditDesignR* software as an R package (available on GitHub at https://github.com/sarahlotspeich/auditDesignR) or a Shiny application (accessible at https://sarahlotspeich.shinyapps.io/auditDesignR/).

We focused on designs for likelihood-based estimators because they tend to be more efficient than design-based estimators and remain valid under a wide range of Phase II designs. To the latter point, data from design-based optimal designs can be analyzed with likelihood-based estimators, but the opposite is not necessarily true: likelihood-based optimal designs can be too extreme for analysis via design-based estimators. Earlier work has suggested to design validation studies that are optimal for design-based estimators because these designs can be used with all estimators and still result in improved efficiency for likelihood-based estimators over other traditional designs (McIsaac and Cook, 2014). However, in other work, we have seen substantial efficiency gains by using an optimal likelihood-based design over an optimal design-based design when using a likelihood-based analysis (Amorim et al., 2021).

While the MLE makes parametric assumptions, in earlier work we found that using logistic regression for outcome and exposure misclassification models can be fairly robust for two-phase analyses (Lotspeich et al., 2021). Note that optimal designs for the MLE are also likely optimal or near-optimal for other likelihood-based alternatives such as the SMLE (Lotspeich et al., 2021), which makes no assumptions about the exposure error mechanism. In an additional simulation, we saw that the efficiency gains of the SMLE based on the optMLE and optMLE-2 designs over the BCC*, CC*, SRS designs were essentially identical to those of the MLE (Table 2.10).

Future research could consider extending the proposed optimal designs to likelihood-based estimators for more general outcomes and exposures. Our designs can be implemented with continuous covariates or exposures but require categorization. How to best stratify continuous covariates for design purposes or obtain optimal designs that do not require stratifying continuous covariates could be investigated.

2.7 Appendix A

2.7.1 Derivations of $S^v(\cdot)$ and $S^{\bar{v}}(\cdot)$

We denote the log-likelihood contribution of the ith subject by $l^v(\boldsymbol{\theta}; Y_i^*, X_i^*, \boldsymbol{Z}_i, Y_i, X_i)$ or $l^{\bar{v}}(\boldsymbol{\theta}; Y_i^*, X_i^*, \boldsymbol{Z}_i)$ $(i = 1, \ldots, N)$, respectively, depending on whether his/her data have been validated or not. We denote the score vector for the ith subject by $S_i(\boldsymbol{\theta}) = (S_i(\beta), S_i(\boldsymbol{\eta})^{\mathrm{T}})^{\mathrm{T}}$, where

$$\begin{aligned} S_i(\theta_j) =& V_i \frac{\partial}{\partial \theta_j} l_i^v(\boldsymbol{\theta}; Y_i^*, X_i^*, \boldsymbol{Z}_i, Y_i, X_i) + (1 - V_i) \frac{\partial}{\partial \theta_j} l_i^{\bar{v}}(\boldsymbol{\theta}; Y_i^*, X_i^*, \boldsymbol{Z}_i) \\ \equiv& V_i S^v(\theta_j; Y_i^*, X_i^*, \boldsymbol{Z}_i, Y_i, X_i) + (1 - V_i) S^{\bar{v}}(\theta_j; Y_i^*, X_i^*, \boldsymbol{Z}_i), \ \forall\ \theta_j \in \boldsymbol{\theta}. \end{aligned}$$

We decompose $\boldsymbol{\eta}^{\mathrm{T}}$ into $(\boldsymbol{\eta}_{y^*}^{\mathrm{T}}, \boldsymbol{\eta}_{x^*}^{\mathrm{T}}, \boldsymbol{\eta}_y^{\mathrm{T}}, \boldsymbol{\eta}_x^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\eta}_{y^*}$, $\boldsymbol{\eta}_{x^*}$, $\boldsymbol{\eta}_y$, and $\boldsymbol{\eta}_x$ correspond to the nuisance parameters in models $P_{\boldsymbol{\eta}_{y^*}}(Y^*|X^*, \boldsymbol{Z}, Y, X)$, $P_{\boldsymbol{\eta}_{x^*}}(X^*|\boldsymbol{Z}, Y, X)$, $P_{\boldsymbol{\eta}_y}(Y|X, \boldsymbol{Z})$, and $P_{\boldsymbol{\eta}_x}(X|\boldsymbol{Z})$, respectively. Then, we have

$$\begin{aligned} & S^v(\theta_j; Y_i^*, X_i^*, \boldsymbol{Z}_i, Y_i, X_i) \\ =& \begin{cases} \left\{ \frac{\partial}{\partial \theta_j} P(Y_i^*|X_i^*, \boldsymbol{Z}_i, Y_i, X_i) \right\} \left\{ P(Y_i^*|X_i^*, \boldsymbol{Z}_i, Y_i, X_i) \right\}^{-1}, & \text{if } \theta_j \in \boldsymbol{\eta}_{y^*}, \\ \left\{ \frac{\partial}{\partial \theta_j} P(X_i^*|\boldsymbol{Z}_i, Y_i, X_i) \right\} \left\{ P(X_i^*|\boldsymbol{Z}_i, Y_i, X_i) \right\}^{-1}, & \text{if } \theta_j \in \boldsymbol{\eta}_{x^*}, \\ \left\{ \frac{\partial}{\partial \theta_j} P(Y_i|X_i, \boldsymbol{Z}_i) \right\} \left\{ P(Y_i|X_i, \boldsymbol{Z}_i) \right\}^{-1}, & \text{if } \theta_j \in (\beta, \boldsymbol{\eta}_y), \\ \left\{ \frac{\partial}{\partial \theta_j} P(X_i|\boldsymbol{Z}_i) \right\} \left\{ P(X_i|\boldsymbol{Z}_i) \right\}^{-1}, & \text{if } \theta_j \in \boldsymbol{\eta}_x, \end{cases} \\ & S^{\bar{v}}(\theta_j; Y_i^*, X_i^*, \boldsymbol{Z}_i) \\ =& \frac{\sum_{y=0}^{1} \sum_{x=0}^{1} \frac{\partial}{\partial \theta_j} \left\{ P(Y_i^*|X_i^*, \boldsymbol{Z}_i, y, x) P(X_i^*|\boldsymbol{Z}_i, y, x) P(y|x, \boldsymbol{Z}_i) P(x|\boldsymbol{Z}_i) \right\}}{\sum_{y=0}^{1} \sum_{x=0}^{1} P(Y_i^*|X_i^*, \boldsymbol{Z}_i, y, x) P(X_i^*|\boldsymbol{Z}_i, y, x) P(y|x, \boldsymbol{Z}_i) P(x|\boldsymbol{Z}_i)}. \end{aligned}$$

2.7.2 Additional Simulations

2.7.2.1 Outcome misclassification only

For the special scenario of outcome misclassification alone, $X^* = X$ such that equation (2.1) reduces to $P(Y^*, Y, X) = P(Y^*|Y, X)P(Y|X)P(X)$. We generated Y and X in the same way as in Section 2.3.2, with $p_y = 0.3$ and $p_x = 0.1$, for a sample of $N = 10{,}000$ subjects. We generated error-prone Y^* from a Bernoulli distribution with $P(Y^* = 1|Y, X) = [1 + \exp\{-(\alpha_0 + \alpha_1 Y + 0.28X)\}]^{-1}$, where α_0 and α_1 were defined in the same way as in Section 3.2 with $FPR_0(Y^*) \in \{0.1, 0.5\}$ and $TPR_0(Y^*) \in \{0.9, 0.5\}$. We set $n = 400$. Without exposure misclassification, the sampling strata for the BCC*, optMLE, and optMLE-2 designs were defined by (Y^*, X). The grid search located the optimal designs in $\geq 98\%$ of replicates.

Table 2.4: Simulation results under outcome or exposure misclassification only

(a) Outcome Misclassification Only

Errors in $Y^\star$		optMLE-2				BCC*				CC*				SRS			
FPR_0	TPR_0	Bias	SE	RE	RI	Bias	SE	RE	RI	Bias	SE	RE	RI	Bias	SE	RE	RI
0.1	0.9	0.002	0.125	0.843	0.840	−0.002	0.133	0.741	0.800	0.005	0.209	0.302	0.526	0.003	0.234	0.240	0.452
	0.5	0.018	0.183	0.829	0.952	−0.002	0.192	0.754	0.902	−0.009	0.293	0.325	0.563	0.009	0.315	0.281	0.548
0.5	0.9	0.000	0.199	0.859	0.898	−0.001	0.218	0.717	0.790	−0.013	0.376	0.241	0.477	0.006	0.345	0.287	0.519
	0.5	0.010	0.219	0.914	0.997	0.004	0.207	1.019	1.061	−0.009	0.373	0.314	0.623	0.001	0.348	0.362	0.657

(b) Exposure Misclassification Only

Errors in $X^\star$		optMLE-2				BCC*				CC*				SRS			
FPR_0	TPR_0	Bias	SE	RE	RI	Bias	SE	RE	RI	Bias	SE	RE	RI	Bias	SE	RE	RI
0.1	0.9	−0.005	0.158	0.861	0.917	0.009	0.180	0.660	0.838	0.005	0.282	0.270	0.536	0.003	0.285	0.264	0.540
	0.5	0.003	0.206	1.002	0.928	0.013	0.237	0.754	0.788	0.008	0.313	0.433	0.607	0.002	0.332	0.385	0.618
0.5	0.9	−0.011	0.298	0.852	0.902	0.008	0.361	0.580	0.736	−0.010	0.327	0.706	0.831	−0.014	0.336	0.667	0.795
	0.5	−0.030	0.345	0.941	0.957	0.017	0.343	0.949	0.959	0.006	0.348	0.923	0.964	−0.012	0.362	0.854	0.880

Note: Bias and SE are, respectively, the empirical bias and standard error of the MLE. Each entry is based on 1000 replicates.

Figure 2.4 shows the average Phase II stratum sizes selected under each of the designs. The optimal designs favored strata with the less-frequent value of Y^* (i.e., $Y^* = 1$) in all settings where it was informative (i.e., $FPR_0(Y^*) \neq 0.5$ or $TPR_0(Y^*) \neq 0.5$). In the highest error setting, the optimal designs appeared to be similar to the BCC* design. Simulation results for the MLE are included in Table 2.4(a). The optMLE-2 design did not lose much efficiency to the optMLE design and typically surpassed the efficiencies of the BCC*, CC*, and SRS designs, with gains as high as 17%, 72%, and 72%, respectively.

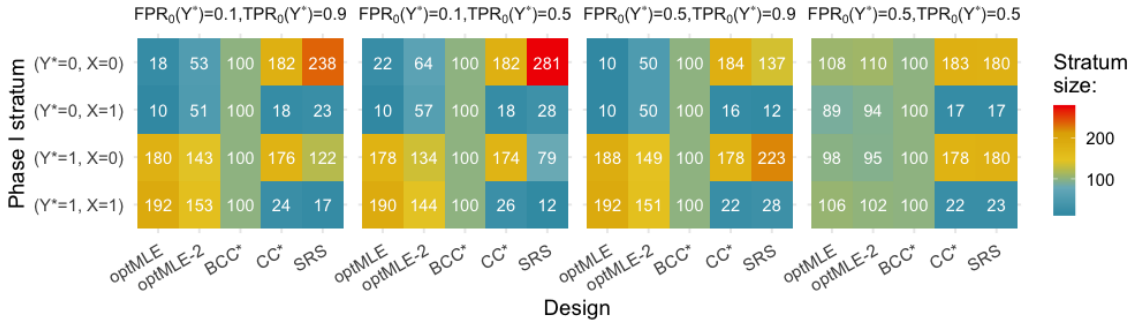


Figure 2.4: Average Phase II stratum sizes n_{y^*x} under outcome misclassification.

2.7.2.2 Exposure misclassification only

For the special scenario of exposure misclassification alone, $Y^* = Y$ such that equation (2.1) reduces to $P(X^*, Y, X) = P(X^*|Y, X)P(Y|X)P(X)$. We generated Y and X in the same way as in 2.7.2.1 for a Phase I sample of $N = 10{,}000$ subjects. We generated error-prone X^* from a Bernoulli distribution with $P(X^* = 1|Y, X) =$

$[1 + \exp\{-(\gamma_0 + 0.45Y + \gamma_1 X)\}]^{-1}$, where γ_0 and γ_1 were defined in the same way as in Section 2.3.2 with $FPR_0(X^*) \in \{0.1, 0.5\}$ and $TPR_0(X^*) \in \{0.9, 0.5\}$. We set $n = 400$. Without outcome misclassification, the sampling strata for the BCC*, optMLE, and optMLE-2 designs were defined by (Y, X^*). The grid search successfully located the optimal designs in all replicates.

Figure 2.5 shows the average Phase II stratum sizes selected under each of the designs. The optimal designs favored strata with the less-frequent value of X^* (i.e., $X^* = 1$) in all settings where it was informative (i.e., $FPR_0(X^*) \neq 0.5$ or $TPR_0(X^*) \neq 0.5$). In the highest error setting, the optimal designs appeared to be similar to the BCC* design. Together with 2.7.2.1, these results suggest that the optimal designs seemed to target the less-frequent value of the error-prone variable with very little regard for the error-free variable. Simulation results for the MLE are included in Web Table 2.4(b). The optMLE-2 design did not lose much efficiency to the optMLE design and typically surpassed the efficiencies of the BCC*, CC*, and SRS designs, with gains as high as 32%, 69% and 69%, respectively.

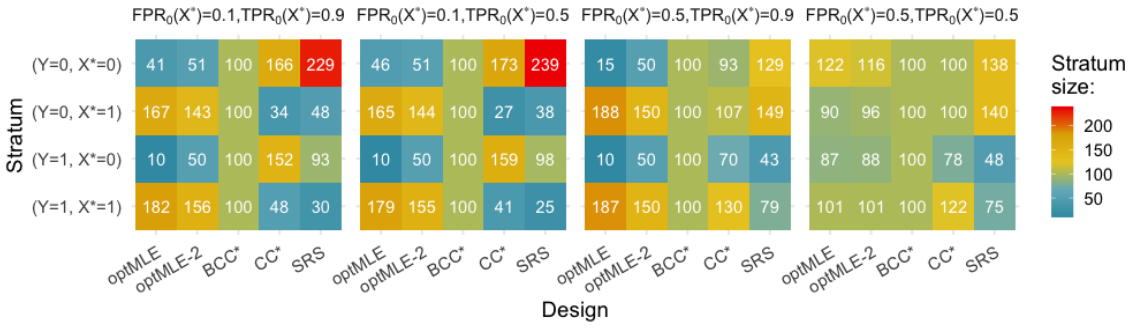
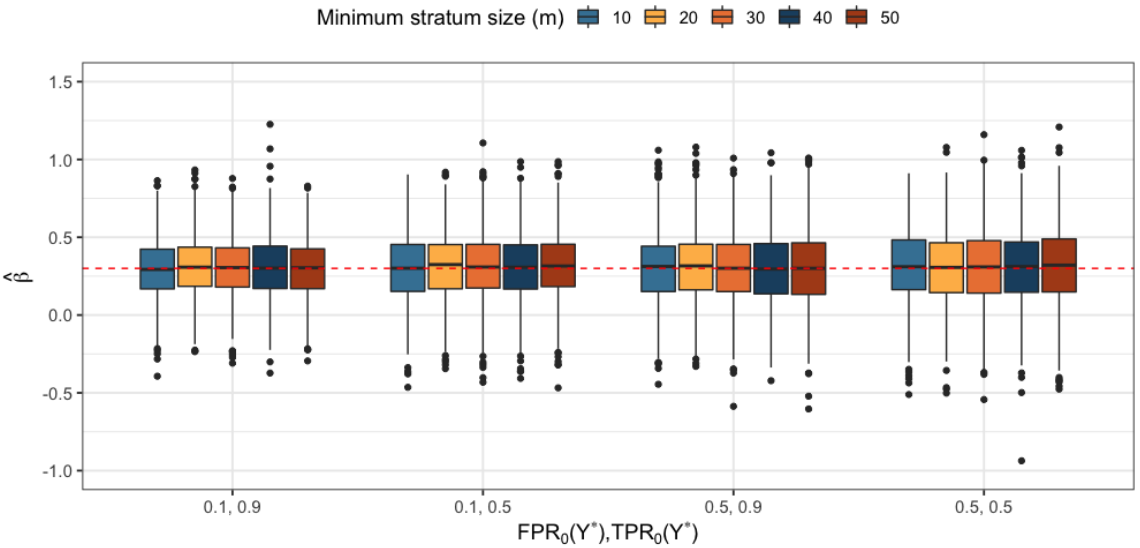


Figure 2.5: Average Phase II stratum sizes n_{yx^*} under exposure misclassification.

2.7.3 Additional Figures and Tables

(a) Exposure misclassification rates were fixed at $FPR_0(X^*) = 0.1$ and $TPR_0(X^*) = 0.9$.



(b) Outcome misclassification rates were fixed at $FPR_0(Y^*) = 0.1$ and $TPR_0(Y^*) = 0.9$.

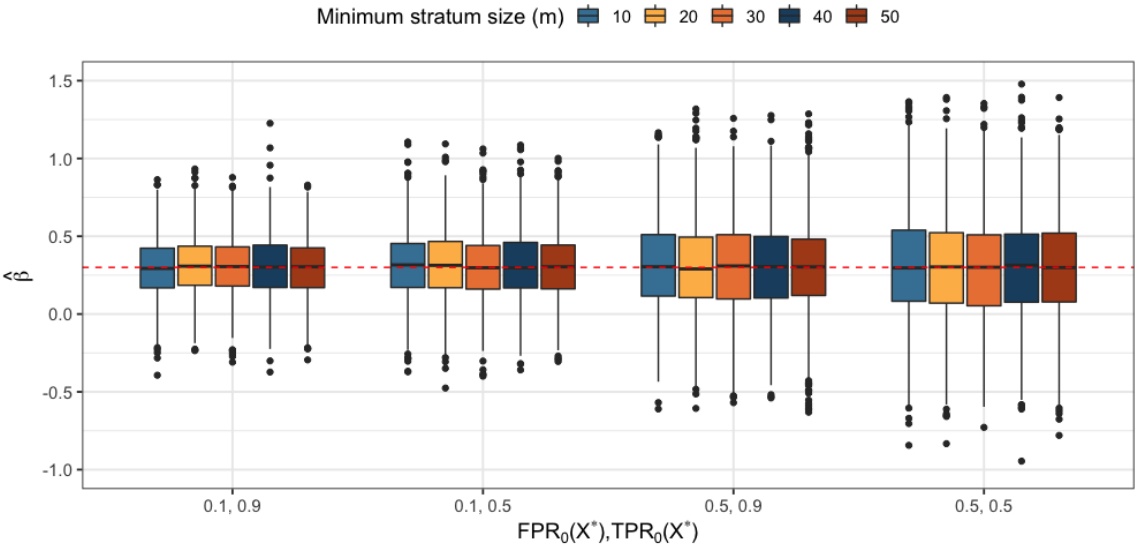
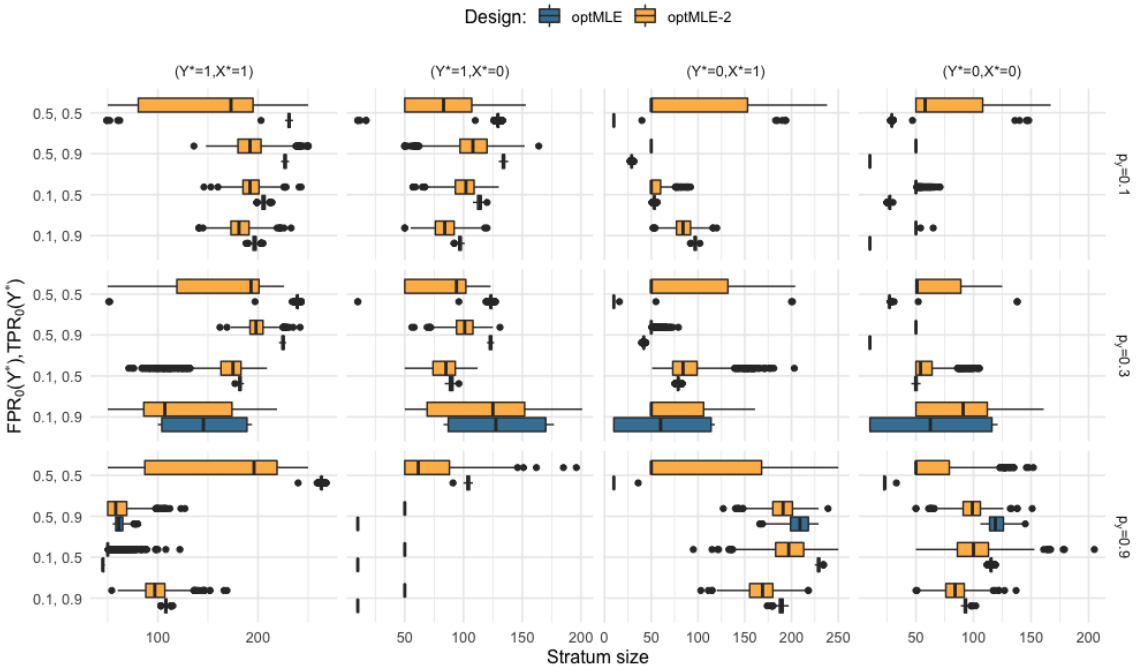


Figure 2.6: Distribution of $\hat{\beta}$ under the optMLE design with outcome and exposure misclassification. The dashed line denotes the true value $\beta = 0.3$.

(a) Exposure misclassification rates were fixed: $FPR_0(X^*) = 0.1$, $TPR_0(X^*) = 0.9$.



(b) Outcome misclassification rates were fixed: $FPR_0(Y^*) = 0.1$, $TPR_0(Y^*) = 0.9$.

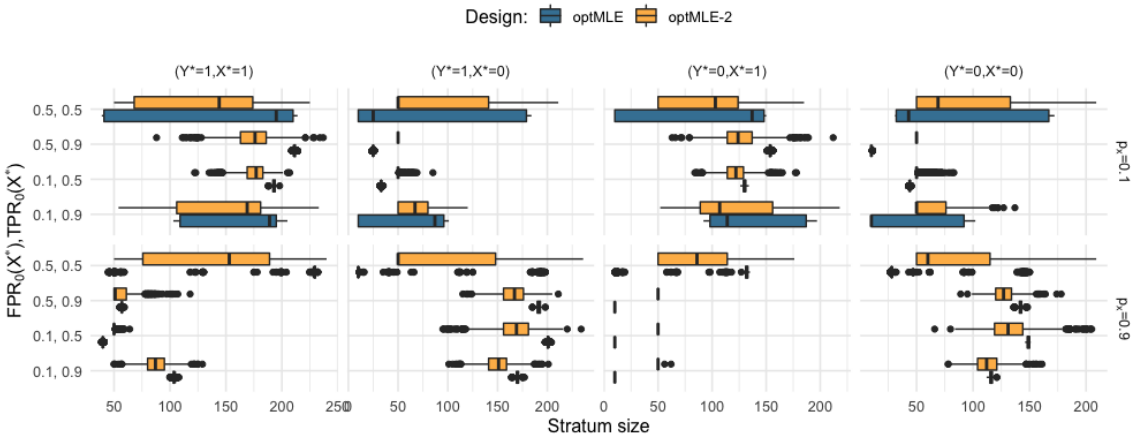


Figure 2.7: Distribution of Phase II stratum sizes $n_{y^*x^*}$ under outcome and exposure misclassification.

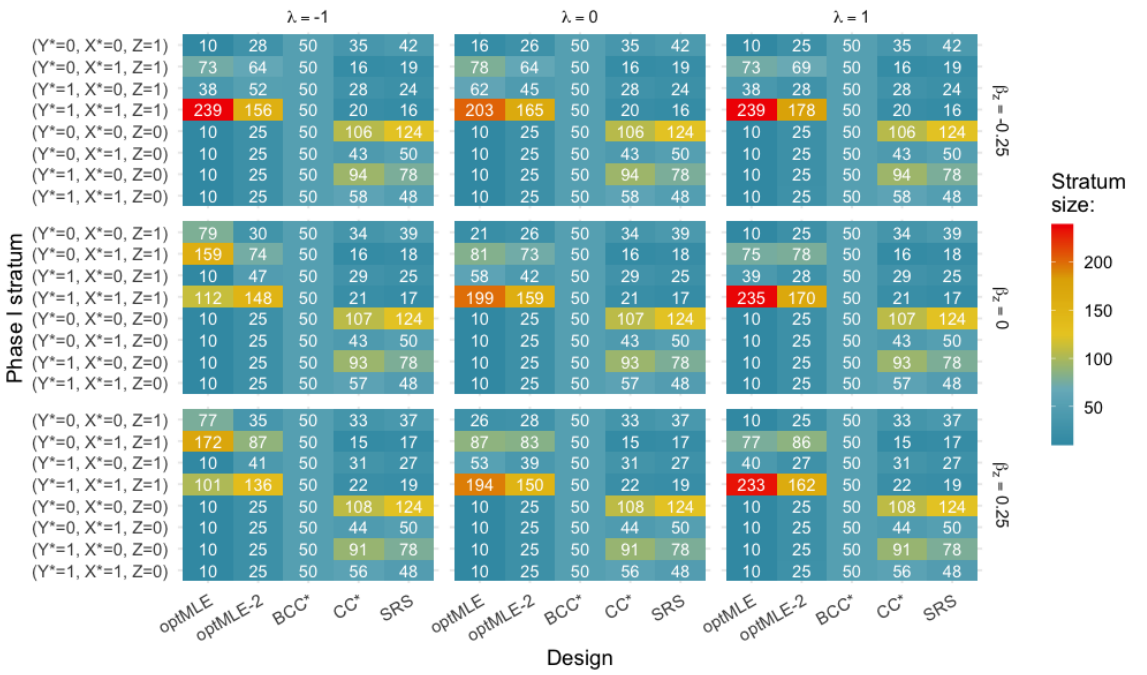


Figure 2.8: Average Phase II stratum sizes $n_{y^*x^*z}$ under outcome and exposure misclassification when an error-free binary covariate Z with 25% prevalence was used in sampling.

Table 2.5: Three versions of the optimal design under outcome and exposure misclassification

Errors in Y^*		**Errors in X^***		**optMLE**		**optMLE-EXP**				**optMLE-FC**			
FPR_0	TPR_0	FPR_0	TPR_0	**Bias**	**SE**	**Bias**	**SE**	**RE**	**RI**	**Bias**	**SE**	**RE**	**RI**
0.1	0.9	0.1	0.9	−0.006	0.191	−0.006	0.193	0.977	0.983	0.001	0.192	0.987	0.976
	0.5			0.003	0.220	0.014	0.225	0.956	1.019	0.007	0.215	1.039	1.012
0.5	0.9			0.004	0.232	0.008	0.221	1.094	0.940	0.007	0.222	1.093	1.024
	0.5			0.014	0.238	0.010	0.239	0.988	0.949	0.019	0.248	0.921	0.982
0.1	0.9	0.1	0.9	−0.006	0.191	−0.006	0.193	0.977	0.983	0.001	0.192	0.987	0.976
			0.5	0.015	0.218	0.013	0.213	1.048	1.002	0.026	0.216	1.013	0.967
		0.5	0.5	0.008	0.347	0.024	0.341	1.031	1.007	0.002	0.338	1.054	1.057
			0.9	0.014	0.292	0.025	0.287	1.033	1.032	0.013	0.285	1.051	0.995

Note: The optMLE design was based on the true parameters $\boldsymbol{\theta}$ and the observed stratum sizes $\{N_{y^*x^*}\}$ in each replicate. The optMLE-EXP design was based on the true parameters $\boldsymbol{\theta}$ and the expected stratum sizes $E(N_{y^*x^*}) = N \times P_{\beta,\boldsymbol{\eta}}(Y^* = y^*, X^* = x^*)$; this design was the same for each replicate. The optMLE-FC design was based on the full cohort parameter estimates $\hat{\boldsymbol{\theta}}^{FC}$ and the observed stratum sizes $\{N_{y^*x^*}\}$ in each replicate. Bias and SE are, respectively, the empirical standard error and bias of the MLE. Each entry is based on 1000 replicates.

Table 2.6: Additional simulation results under outcome and exposure misclassification

a) Varied Outcome Misclassification Rates

	Errors in Y^*		optMLE		optMLE-2		BCC*		CC*		SRS	
p_y	FPR_0	TPR_0	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
0.1	0.1	0.9	−0.012	0.217	−0.009	0.214	−0.012	0.254	−0.028	0.347	−0.058	0.516
		0.5	0.024	0.230	0.003	0.241	0.004	0.286	−0.012	0.362	−0.052	0.512
	0.5	0.9	−0.003	0.311	−0.009	0.321	−0.009	0.409	−0.065	0.560	−0.053	0.563
		0.5	0.011	0.373	−0.005	0.361	−0.019	0.377	−0.032	0.512	−0.081	0.543
0.3	0.1	0.9	−0.006	0.191	0.004	0.190	0.001	0.223	0.022	0.297	−0.004	0.333
		0.5	0.003	0.220	0.015	0.219	−0.002	0.226	0.020	0.317	0.004	0.344
	0.5	0.9	0.004	0.232	0.002	0.241	−0.001	0.274	−0.023	0.386	0.006	0.357
		0.5	0.014	0.238	0.005	0.248	−0.002	0.240	−0.003	0.369	−0.003	0.369
0.9	0.1	0.9	0.011	0.240	0.003	0.249	0.009	0.277	0.046	0.430	0.093	0.592
		0.5	0.030	0.359	0.000	0.381	0.064	0.505	0.043	0.600	0.082	0.596
	0.5	0.9	0.024	0.267	0.002	0.279	0.022	0.342	0.043	0.543	0.048	0.589
		0.5	0.019	0.473	0.068	0.491	0.071	0.515	0.070	0.620	0.052	0.608

b) Varied Exposure Misclassification Rates

	Errors in X^*		optMLE		optMLE-2		BCC*		CC*		SRS	
p_x	FPR_0	TPR_0	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
0.1	0.1	0.9	−0.006	0.191	0.004	0.190	0.001	0.223	0.022	0.297	−0.004	0.333
		0.5	0.015	0.218	0.013	0.218	0.003	0.247	0.010	0.336	−0.005	0.338
	0.5	0.9	0.014	0.292	0.006	0.295	−0.006	0.351	0.011	0.342	0.000	0.351
		0.5	0.008	0.347	0.008	0.343	0.001	0.342	0.007	0.348	0.014	0.357
0.9	0.1	0.9	−0.001	0.174	0.006	0.189	−0.003	0.201	0.012	0.310	0.000	0.339
		0.5	0.012	0.284	−0.016	0.290	0.013	0.343	0.007	0.345	0.003	0.381
	0.5	0.9	0.011	0.218	−0.006	0.221	0.008	0.264	−0.006	0.337	0.017	0.366
		0.5	0.021	0.365	0.018	0.364	0.023	0.366	0.006	0.360	0.036	0.387

Note: Misclassification rates for X^* and Y^* were fixed at $FPR_0 = 0.1$ and $TPR_0 = 0.9$ in a) and b), respectively. SE is the empirical standard error of the MLE. When $p_y \neq 0.3$, we excluded a few replicates with unstable estimates under the SRS, CC*, or BCC* design. All other entries are based on 1000 replicates.

Table 2.7: Additional simulation results under outcome and exposure misclassification with available error-free covariate information

			optMLE		optMLE-2		BCC*		CC*		SRS	
p_z	λ	β_Z	Bias	SE	Bias	SE	Bias	SE	Bias	SE	Bias	SE
0.25	−1	−0.25	−0.012	0.259	−0.011	0.225	−0.005	0.296	−0.016	0.321	−0.037	0.352
		0.00	0.000	0.238	0.001	0.227	−0.002	0.288	0.010	0.325	−0.010	0.332
		0.25	0.003	0.225	−0.014	0.221	0.023	0.287	−0.006	0.333	0.001	0.337
	0	−0.25	−0.009	0.246	−0.010	0.249	−0.005	0.296	−0.016	0.321	−0.037	0.352
		0.00	−0.004	0.247	−0.010	0.245	−0.002	0.288	0.010	0.325	−0.010	0.332
		0.25	0.000	0.227	−0.006	0.242	0.023	0.287	−0.006	0.333	0.001	0.337
	1	−0.25	−0.012	0.259	−0.016	0.267	−0.005	0.296	−0.016	0.321	−0.037	0.352
		0.00	−0.017	0.252	−0.003	0.275	−0.002	0.288	0.010	0.325	−0.010	0.332
		0.25	−0.007	0.260	0.001	0.270	0.023	0.287	−0.006	0.333	0.001	0.337
0.50	−1	−0.25	−0.014	0.261	−0.018	0.229	0.002	0.304	−0.007	0.310	−0.012	0.338
		0.00	−0.010	0.237	−0.028	0.228	−0.004	0.264	−0.001	0.312	−0.018	0.317
		0.25	−0.006	0.225	−0.018	0.225	0.013	0.287	−0.003	0.308	0.005	0.304
	0	−0.25	0.000	0.241	−0.002	0.250	0.002	0.304	−0.007	0.310	−0.012	0.338
		0.00	0.014	0.230	−0.003	0.239	−0.004	0.264	−0.001	0.312	−0.018	0.317
		0.25	0.009	0.229	−0.007	0.251	0.013	0.287	−0.003	0.308	0.005	0.304
	1	−0.25	−0.014	0.261	−0.028	0.277	0.002	0.304	−0.007	0.310	−0.012	0.338
		0.00	0.002	0.265	−0.005	0.267	−0.004	0.264	−0.001	0.312	−0.018	0.317
		0.25	0.008	0.250	0.008	0.261	0.013	0.287	−0.003	0.308	0.005	0.304

Note: Bias and SE are the empirical bias and standard error of the MLE, respectively. Each entry is based on 1000 replicates.

Table 2.8: Historic TB audits results in CCASAnet

		Misclassified (%)	
Country	**Audited**	**Treatment Completion (Y^*)**	**Bacterial Confirmation (X^*)**
Country Grouping = 0			
A	6	2 (33.3%)	0 (0.0%)
B	7	0 (0.0%)	2 (28.6%)
Country Grouping = 1			
C	6	1 (16.7%)	1 (16.7%)
D	10	2 (20.0%)	5 (50.0%)
Country Grouping = 2			
E	4	0 (0.0%)	0 (0.0%)

Note: No subjects had both outcome and exposure misclassification.

Table 2.9: Parameter estimates for TB analysis in CCASAnet using historic audits

Coeff	**log OR**
Analysis model (Y)	
Intercept	0.752
X	−0.415
CoG = 0 (Co = A–B)	Referent
CoG = 1 (Co = C–D)	0.601
CoG = 2 (Co = E)	0.211
Outcome misclassification mechanism (Y^*)	
Intercept	2.088
X^*	0.156
Y^*	4.644
X	2.485
CoG = 0 (Co = A–B)	Referent
CoG = 1 (Co = C–D)	−1.182
CoG = 2 (Co = E)	−0.956
Exposure misclassification mechanism (X^*)	
Intercept	−0.600
Y	−2.611
X	4.770
CoG = 0 (Co = A–B)	Referent
CoG = 1 (Co = C–D)	1.685
CoG = 2 (Co = E)	0.170
Exposure model (X)	
Intercept	−1.017
CoG = 0 (Co = A–B)	Referent
CoG = 1 (Co = C–D)	−0.160
CoG = 2 (Co = E)	−0.592

Note: The optimal designs for CCASAnet were based on parameters $\hat{\beta} = -0.415$ and $\hat{\boldsymbol{\eta}}^{\mathrm{T}} = (0.752,\ 0,\ 0.601,\ 0.601,\ 0.211,\ 2.088,\ 0.156,\ 4.644,\ 2.485,\ 0,\ -1.182,\ -1.182,\ -0.956,\ -0.6,\ -2.611,\ 4.77,\ 0,\ 1.685,\ 1.685,\ 0.17,\ -1.017,\ 0,\ -0.16,\ -0.16,\ -0.592)^{\mathrm{T}}$.

Table 2.10: Simulation results comparing the MLE and SMLE

	MLE				SMLE			
Design	**Bias**	**SE**	**RE**	**RI**	**Bias**	**SE**	**RE**	**RI**
optMLE	0.010	0.176	1.000	1.000	0.008	0.176	1.000	1.000
optMLE-2	0.003	0.181	0.942	0.933	0.001	0.181	0.946	0.934
BCC*	0.002	0.198	0.788	0.853	0.001	0.198	0.785	0.855
CC*	0.000	0.270	0.426	0.616	0.000	0.270	0.425	0.618
SRS	−0.017	0.305	0.333	0.541	−0.017	0.305	0.332	0.544

Note: The SMLE was proposed with X^* as a surrogate for X such that $(Y \perp X^*)|X$. Thus, X^* was generated from a Bernoulli distribution with $P(X^* = 1|Y, X) = [1 + \exp\{-(\gamma_0 + \gamma_1 X)\}]^{-1}$. All other variables were generated as in Section 2.3.2, with $p_y = 0.3$, $p_x = 0.1$, $FPR_0(Y^*) = 0.1$, $TPR_0(Y^*) = 0.9$, $FPR_0(X^*) = 0.1$, and $TPR_0(X^*) = 0.9$. Bias and SE are, respectively, the empirical standard error and bias of the estimators. Each entry is based on 1000 replicates.

CHAPTER 3

SELF-AUDITS AS ALTERNATIVES TO TRAVEL-AUDITS FOR IMPROVING DATA QUALITY IN THE CARIBBEAN, CENTRAL AND SOUTH AMERICA NETWORK FOR HIV EPIDEMIOLOGY

This chapter is adapted from "Self-audits as alternatives to travel-audits for improving data quality in the Caribbean, Central and South America network for HIV epidemiology" published in *Journal of Clinical and Translational Science* and has been reproduced with the permission of the publisher and my co-authors Mark J. Giganti, Marcelle Maia, Renalice Vieira, Daisy Maria Machado, Regina Célia Succi, Sayonara Ribeiro, Mario Sergio Pereira, Maria Fernanda Rodriguez, Gaetane Julmiste, Marco Tulio Luque, Yanink Caro-Vega, Fernando Mejia, Bryan E. Shepherd, Catherine C. McGowan and Stephany N. Duda.

3.1 Introduction

High quality data are essential for valid inference and decision making in observational HIV cohort research. Source document verification, or data auditing, is the standard for ensuring high quality data in clinical trials (Weiss, 1998) and has also been used to assess data quality in observational studies (Chaulagai et al., 2005; Kimaro and Twaakyondo, 2005; Kiragga et al., 2011; Mphatswe et al., 2012; Duda et al., 2012; Giganti et al., 2019). The data audit process involves a group of external data auditors visiting the research site, comparing records in the research database to clinical source documents, and reporting any discrepancies. In addition to assessing the accuracy and completeness of existing data, audits can help educate local staff on good data management practices, highlight areas for improvement in data collection methods, and provide a deterrent against data fraud. Statistical methods have been developed that incorporate audit information into analyses, potentially providing more accurate estimates based on error-prone data (Shepherd and Yu, 2011).

Despite its benefits, source document verification of entire databases, or even of a subset of records and variables, is a resource-intensive exercise that often exceeds the available capacity and budget of research studies. We have developed methodologies and tools to simplify the audit preparation and feedback process (Duda et al., 2011, 2012), but the practice remains relatively uncommon among observational HIV cohorts. Although the most objective audits are conducted by external auditors, sending

trained auditors to distant sites within multi-national networks (on-site monitoring or travel-auditing) requires extensive travel funds and dedicated personnel effort (De, 2011). External auditors require additional time to familiarize themselves with local source documents and procedures, while sites may spend unplanned time obtaining patient charts for review and hosting the visitors. Language differences can further complicate the audit process.

To address the logistical and financial challenges of these travel-audits, the present work investigates the efficacy of audits executed by local sites themselves, referred to as "self-audits." These self-audits explore a creative way to continue maintaining high data quality standards, while markedly lowering the costs of performing the audits. Several novel internal checks are built into our self-audits in an attempt to strengthen their validity: (1) records to be audited are randomly selected by the data coordinating center rather than the sites themselves and (2) prior to performing their self-audits, local personnel are notified that several sites will be randomly selected to have their self-audits verified by travel auditors, i.e., that external auditors will travel to some of the sites and validate the same records as self-auditors. We describe our experience conducting a self-audit process within a multi-national HIV cohort and compare the findings from self- and travel-audits at those sites randomly selected to receive both.

3.2 Methods

3.2.1 Cohort

The Caribbean, Central, and South America network for HIV epidemiology (CCASAnet) is a consortium of HIV care and treatment clinics in seven countries in Latin America. CCASAnet clinics pool their routine clinical care data to conduct collaborative research to better understand the HIV epidemic in the region (McGowan et al., 2007). In early 2017, investigators from eight sites (six adult and two pediatric) participated in a new self-audit process to review their data. Participating sites included Instituto Nacional de Infectologia Evandro Chagas in Rio de Janeiro, Brazil; Universidade Federal de Sao Paulo, Brazil; Universidade Federal de Minas Gerais, Brazil; Fundación Arriarán in Santiago, Chile; Le Groupe Haïtien d'Etude du Sarcome de Kaposi et des Infections Opportunistes in Port-au-Prince, Haiti; Instituto Hondureño de Seguridad Social and Hospital Escuela Universitario in Tegucigalpa, Honduras; El Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán in Mexico City, Mexico; and Instituto de Medicina Tropical Alexander von Humboldt

in Lima, Peru.

The CCASAnet data structure roughly follows data exchange protocols outlined by the HIV Cohorts Data Exchange Protocol (HICDEP) and the International epidemiology Databases to Evaluate AIDS (IeDEA) (*IeDEA - Data Exchange Standard (DES)*, 2017; *HICDEP 1.110.*, 2017). Variables captured in the following patient forms were audited: basic demographic information typically captured at enrollment (tblBAS), information on visits and vital status (tblVIS and tblLTFU), CD4+ cell count measurements (tblLAB_CD4), HIV viral load measurements (tblLAB_RNA), antiretroviral therapy regimens and dates (tblART), and clinical endpoints (tblCEP). An abbreviated data dictionary for these tables is provided in Section 3.5 (Table 3.3). Of 28 audited variables, 13 were captured at each patient's first appointment only (e.g., sex, birthdate), and 15 could be collected multiple times per patient during subsequent clinic visits (e.g., weight, height, CD4+ cell count). For clarity, we refer to individual occurrences or measurements of these data values as "data entries." Clinical source documents were in the form of parallel paper-based forms at each of the study sites.

3.2.2 Study Design

For each site, the CCASAnet Data Coordinating Center at Vanderbilt (VDCC) selected 40 patient records to be audited. Of these, 20 records were randomly selected among patients enrolling within the previous year to assess the quality of recent data capture, and 20 records were randomly selected among all patients. For the six adult sites, an additional 10 records were chosen from those audited in a previous study (Giganti et al., 2019). Institutional review board approval was obtained from each site and the VDCC.

Prior to the audit, each site selected representatives to attend a two-hour online training session that explained procedures for conducting a data audit and documenting findings. Following completion of the training session, sites were given two weeks to complete the self-audit. Upon return of self-audit results, sites were compensated $2000 US for their efforts. In June 2017, two VDCC investigators performed on-site audits at three randomly selected HIV clinics (one adult and two pediatric) using published audit procedures (Duda et al., 2011, 2012) to audit the same patient records that were chosen for the self-audit. We refer to this as the travel-audit. Travel-auditors spent two and a half days on average performing audits at each site. The self- and travel-audits for this study were completed between May and July 2017.

The VDCC developed a REDCap web-based data capture interface, so site and travel auditors could view audit records and enter audit findings and data corrections in a structured, electronic format (Harris et al., 2009). These REDCap forms were based on audit templates developed in prior audit work (Duda et al., 2011, 2012). The most recently submitted study data for the randomly selected audit patients were imported into corresponding fields in the REDCap database. Each site could only see its own audit records.

For each data entry, both sets of auditors compared the value in REDCap, representing data the site had submitted to CCASAnet for research studies, to the site's source documents, including paper patient charts, laboratory summaries, and electronic data systems. Within the REDCap interface, auditors were asked to categorize their findings as one of the following: "Value matches the chart (correct)," "Value doesn't match the chart," or "Can't find this value in the chart." For data entries in error, auditors were prompted to provide corrected values. If they identified a new data entry in the source documents that was not in the REDCap data but should have been, they entered it into a blank supplemental data field with the label "Found value in the chart (was not in the study data)." For our analysis, findings were collapsed into "correct" (matches the chart) and "incorrect" (does not match the chart, could not be found in the chart, or was found in the chart but not present in the study data). Following completion of the audits, audit findings and corrections were extracted from REDCap for analysis.

3.2.3 Analysis

R Statistical Software (R Core Team, 2019) was used for all analyses. Analyses focused on data entries that were reviewed by both sets of auditors, referred to as the "doubly-audited sample." All patient records were reviewed by self-auditors but due to time constraints, many records were not reviewed by travel-auditors. Characteristics of doubly-audited (included) and self-audited only (excluded) patient records were compared using logistic regression models, controlling for site. Because our sample included both pediatric and adult sites, we excluded two variables, receiving PMTCT as an infant and birth mode, which are specific to pediatric patients. Entries for clinical AIDS diagnosis prior to first visit, date of prior clinical AIDS diagnosis, and WHO stage were also excluded because they were incompletely audited. (The related variables, prior AIDS diagnosis and date, were included.) All other variables reviewed during self- or travel-audits were included in analyses.

To directly compare findings, we defined an entry in "discordance" to mean that self- and travel-audit findings did not agree. Rates of audit discordance were calculated by variable and by site. Variables collected only once (e.g., sex, birthdate) were calculated as the proportion of patients' entries in discordance. Repeatedly collected entries (e.g., visit date, height, and labs) were calculated as both the average per-person proportion of entries in discordance and the total percentage of entries in discordance. The Kappa statistic was computed to estimate overall agreement between self- and travel-audit entries.

Further descriptive analyses addressed differences in how self- and travel-auditors recorded fixes for incorrect entries, focusing on entries agreed to be incorrect by both sets of auditors and excluding those that could not be found in the patient chart. With these, we inspected different fixes submitted for the same incorrect entries (called "mismatched corrections"). Mismatched corrections were reviewed by two investigators to identify possible causes for the mismatch, which we categorized as audit protocol issues, date approximations, genuine differences, near-equivalent coding, or typographical errors (typos). Audit protocol issues included entries that one team declared incorrect while the other team labeled "could not verify/no source document" (a matter of interpretation or thoroughness of chart searching) or some incomplete data entries that self-auditors corrected in a way that created duplicate data, while travel-auditors labeled them "could not find in the chart" in order to avoid duplicate data rows. For example, an instance of ce_d (start date of clinical endpoint) was corrected by self-auditors to the same date as the previous ce_d (creating a duplicate of this clinical endpoint), while travel-auditors reported that they could not find the original ce_d value. Other correction mismatches occurred because of the combination of dates and date approximation variables. One audit team might record a corrected date "exact to the day" whereas another recorded a correction that was only "exact to the month." Although one correction was more precise than the other, both were technically correct.

3.3 Results

A total of 39,269 entries in 130 patient records were selected for self- and travel-audit across the three sites. Figure 3.1 (left panel) summarizes audit results for these entries. Travel-auditors faced time constraints and therefore audited fewer records: 29,965 of the selected entries (76%) were self- but not travel-audited. Patients whose records were not at all travel-audited (n=65, 50%) were not materially different

from those who were at least partially travel-audited, since the order of auditing was essentially random (Section 3.5, Table 3.4). Among patients that were both self- and travel-audited (n=65), some were not fully audited. There were 52 patients (80%) whose original records contained entries that were self- but not travel-audited, and 41 patients (63%) with entries that were travel- but not self-audited. While these percentages appear somewhat similar, we note that only 298 entries (less than 1% of the original sample) were travel- but not self-audited whereas 29,965 entries (76%) were self- but not travel-audited.

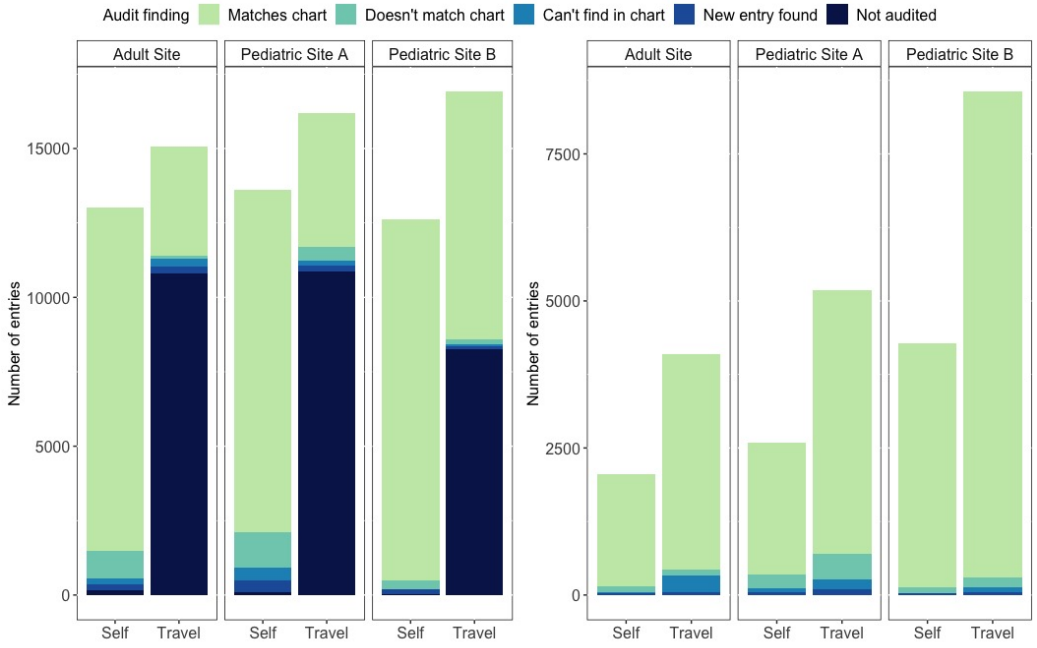


Figure 3.1: Comparison of audit findings between self- and travel-auditors at the three sites (left) and among only doubly-audited entries (right).

3.3.1 Overall Data Quality

Across 65 patient records, 8919 data entries capturing 28 clinical and demographic variables were both self- and travel-audited. Figure 3.1 (right panel) shows the number and distribution of errors by audit site for entries that were both self- and travel-audited. Across all variables, records, and sites, self- and travel-auditors reported similar proportions of entries to be correct in the doubly-audited sample (93% versus 92%). Self-auditors reported slightly more values not matching the charts (5.0% versus 3.8%), while travel-auditors indicated that more values could not be found in charts (3.0% versus 1.0%). Auditors reported the same number of values that were found in the patient chart but not originally in the database (1.1%).

Figure 3.2: Percentage of audit findings by variable and audit type. Variable definitions are in Section 3.5.

Figure 3.2 shows the error rates for each variable from both travel- and self-audits among the doubly-audited sample. Independently, self- and travel-auditors reported similar rates of incorrect values for all variables in the ART, viral load, and CD4 data tables (all rates were less than 5% different). From the baseline and visit tables, both teams reported small error rates for sex, death (yes/no), height, weight, and visit date (all less <5%). However, audit results for date of death substantially differed, as self-auditors reported 0% (0 of 12) incorrect but travel-auditors reported 33% (4 of 12) incorrect. Similar rates of incorrect entries for clinical diagnosis dates (ce_d) were reported by both audit teams, but travel-auditors reported that 16% (30 of 185) of disease codes (ce_id) were incorrect whereas self-auditors reported only 2% (3 of 185) incorrect. Across all variables except AIDS diagnosis date (aids_d), travel-auditors reported a larger number of entries that could not be found in the chart. Details are shown in Table 3.1.

Table 3.1: Self- and travel-audit discordance by variable in the doubly-audited sample (n=8919 entries)

	Audited patient records	Audited entries	Discordant entries	Average discordance per record	Overall discordance
tblBAS[a]	**434**	**434**	**38**	**9%**	**9%**
male_y	65	65	0	0%	0%
birth_d	65	65	1	2%	2%
hiv_diagnosis_d	62	62	14	23%	23%
mode	63	63	5	8%	8%
recart_y	62	62	3	5%	5%
aids_y	36	36	4	11%	11%
aids_d	18	18	5	28%	28%
baseline_d	63	63	6	10%	10%
tblLTFU[a]	**168**	**168**	**18**	**11%**	**11%**
drop_y	52	52	1	2%	2%
drop_d	22	22	4	18%	18%
drop_rs	19	19	8	42%	42%
death_y	63	63	1	2%	2%
death_d	12	12	4	33%	33%
tblVIS[b]	**213**	**4248**	**206**	**11%**	**5%**
visit_d	57	1216	43	6%	4%
height	57	1075	36	5%	3%
weight	57	1061	52	12%	5%
cdcstage	42	896	75	26%	8%
tblLAB_CD4[b]	**164**	**1933**	**46**	**4%**	**2%**
cd4_d	55	652	19	6%	3%
cd4_v	55	649	8	2%	1%
cd4_per	54	632	19	3%	3%
tblLAB_RNA[b]	**96**	**1248**	**54**	**6%**	**4%**
rna_d	48	624	23	8%	4%
rna_v	48	624	31	5%	5%
tblART[b]	**193**	**514**	**69**	**13%**	**13%**
art_id	58	155	9	4%	6%
art_sd	58	153	25	20%	16%
art_ed	39	110	19	17%	17%
art_rs	38	96	16	15%	17%
tblCEP[b]	**100**	**374**	**79**	**20%**	**21%**
ce_d	50	189	48	23%	25%
ce_id	50	185	31	18%	17%

Note: [a]Variables from tblBAS and tblLTFU were collected once per record. [b]Variables from all other tables were repeatedly collected.

For quantitative variables height, weight, lab values, and dates, the median discrepancies between the original entries and the self- or travel-audit corrections (with

the interquartile range [IQR]) are included in Table 3.2. Lab dates (cd4_d and rna_d) were corrected in at least 20 entries by either set of auditors, with median corrections of 16–17 days for CD4 labs and 0–2 days for RNA labs. For ART regimens, self-auditors corrected both the start and end dates by about 1 month on average, while travel-auditors supplied slightly larger fixes to start dates than to end dates (median differences of 28 and 17 days, respectively). Many of the remaining variables were corrected on only a few entries (e.g., birth_d, aids_d, baseline_d, cd4_v, cd4_per), so median discrepancies may appear more extreme. Corrections made to height and weight were minor.

Table 3.2: Magnitude of discrepancies between original entries in quantitative variables found to not match the charts and corrections submitted by self- or travel-auditors.

	Self-audit		Travel-audit	
	Corrected entries	Median difference (IQR)	Corrected entries	Median difference (IQR)
tblBAS				
birth_d	3	6 (4, 34) days	2	34 (20, 47) days
hiv_diagnosis_d	14	6 (−377, 110) days	17	14 (−14, 155) days
aids_d	2	116 (43, 190) days	4	10 (−4, 81) days
baseline_d	6	−8 (−17, 2) days	2	817 (416, 1217) days
tblLTFU				
drop_d	3	0 (0, 33) days	1	−309 (NA, NA) days
death_d	0	NA (NA, NA) days	4	7 (−8, 19) days
tblVIS				
visit_d	17	−6 (−31, 29) days	7	−10 (−228, −5) days
height	12	−0.4 (−1.4, 2.9) cm	13	−0.5 (−1.0, 0.5) cm
weight	26	−0.1 (−0.1, −0.1) kg	27	−0.1 (−0.1, −0.1) kg
tblLAB_CD4				
cd4_d	26	17 (1, 38) days	20	16 (1, 40) days
cd4_v	4	−53 (−325, 441) cells/mm^3	5	−100 (−600, −5) cells/mm^3
cd4_per	8	−0.3 (−15.1, 4.4)%	5	−0.6 (−4.0, 1.0)%
tblLAB_RNA				
rna_d	26	2 (−4, 26) days	31	0 (−8, 27) days
rna_v	18	350 (30, 350) copies/mL	7	−101 (−498, 366) copies/mL
tblART				
art_sd	28	32 (0, 154) days	25	28 (0, 47) days
art_ed	23	−26 (−88, 149) days	19	17 (−17, 154) days
tblCEP				
ce_d	36	56 (−2, 147) days	11	30 (21, 227) days

Note: Entries from the doubly-audited sample (n = 8919) that received audit findings of "doesn't match chart" from self-auditors and/or from travel-auditors are included in the left and right halves of the table, respectively.

3.3.2 Comparing Audit Findings

Across all patient entries, 8409 (94%, Kappa = 0.59) received the same assessment from self- and travel-auditors (7988 correct and 421 incorrect). Of the 510 discordant entries, 44% were marked correct by travel-auditors but incorrect by self-auditors while 56% were marked correct by self-auditors but incorrect by travel-auditors. Pa-

tient sex was the only variable with no discordance, but other variables pertaining to CD4 and RNA labs, routine visit variables (visit_d, height, and weight), and baseline variables (recart_y, drop_y, death_y, and birth_d) exhibited less than 5% discordance based on more than 50 entries each.

Variables drop_rs (42%) and death_d (33%) from tblLTFU had the largest individual proportions discordant, but these estimates were based on only 19 and 12 entries, respectively, while many of the more concordant variables were based on >600 entries. Of the 8 discordant drop_rs entries, 6 were originally entered as "Other," which self-auditors considered to match the chart while travel-auditors corrected 3 to "LTFU/not known to be dead" and could not find the other 3. Disagreement in the 4 death_d entries came from self-auditors finding them "correct" while travel-auditors proposed corrections that were 0, 30, 34, and 14 days from the original entries. There was also 18% discordance in ce_id, with more than half of the discordant ce_id entries unable to be found by travel-auditors but assessed to be correct by self-auditors. Between-audit discordance rates are reported for all variables in Table 3.1.

Discordance rates for certain variables (e.g., CDC stage, reason for dropout, dropout date, and death date) differed between the three sites Figure 3.4). While these variables saw more variability in site-specific discordance rates, we note that rates for reason for dropout, dropout date, and death date were based on a small number of audited entries (≤ 10). We attribute the disparities between discordance rates for CDC stage to differences in the number of audited entries at the three sites: 32 at the adult site versus 292 and 572 at Pediatric Sites A and B, respectively.

3.3.3 Comparing Audit Fixes

Of 421 entries marked incorrect by both teams of auditors, 52 (12%) were not corrected by either because the original values could not be found in the patient charts; these were appropriately left uncorrected. Of the remaining 369 erroneous entries, 304 (82%) were corrected by both auditors (called the "doubly-corrected sample"). Most of the singly-corrected entries (n=60, 92%) were unable to be found by one team but were found by the other, which resulted in the blank corrections. All entries agreed to be incorrect are included in Figure 3.3, colored by the categorized comparison of the self- and travel-auditor corrections.

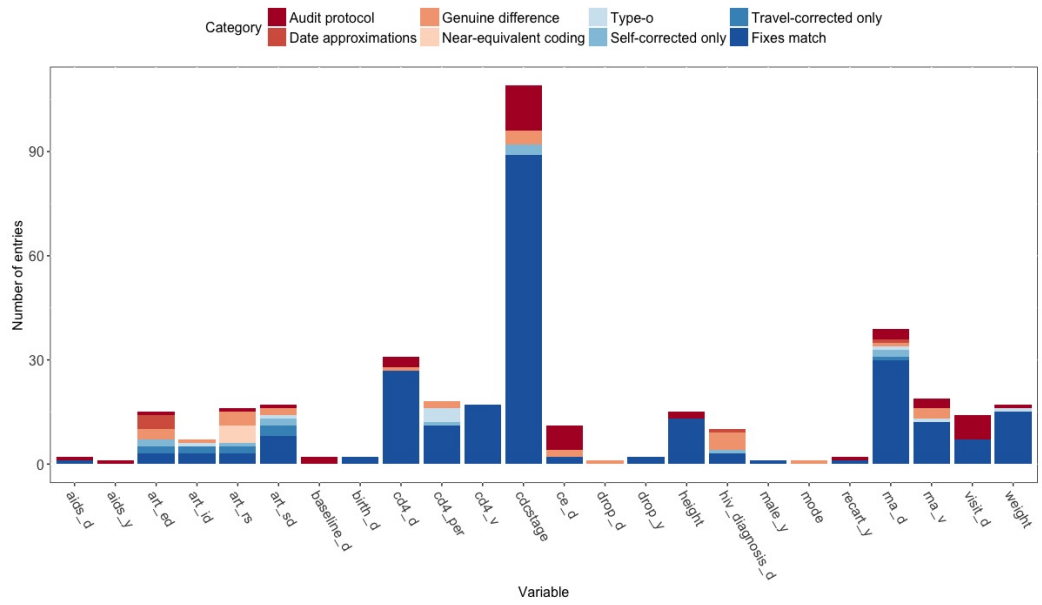


Figure 3.3: A comparison of corrections made to 421 entries assessed as incorrect by both self- and travel-auditors. This plot includes entries that neither set of auditors could find (which were appropriately left uncorrected and counted as "Fixes match"), as well as singly- and doubly-corrected entries.

Of 304 doubly-corrected records, 250 (82%) received the same corrections from self- and travel-auditors. While proportions of entries receiving mismatched corrections varied by variable, there were seven variables that received all matching fixes from self- and travel-auditors: CD4 value (n=17), height (n=13), birthdate (n=2), drop from cohort (n=2), prior ART (n=1), date of prior AIDS diagnosis (n=1), and sex (n=1). On the other hand, the mismatched corrections (n=54) occurred across 15 variables, with the largest numbers in "reason for changing ART regimen" (n=9), end date of ART regimen (n=8), viral load value (n=6), CD4% (n=6), and date of HIV diagnosis (n=6).

The largest number of mismatches were identified to be genuine differences between self- and travel-corrections (n=29, 54%). These mismatches were found primarily across variables CDC stage, reason for switching ART regimen, and date of HIV diagnosis variables (4–5 entries each). After this, there were fewer than 10 entries found to be attributable to typographical errors (n=9), differing interpretations of audit protocol (n=6), date approximations (n=6), or near-equivalent coding (n=4). Most typographical errors were found in lab values, where self-auditors entered commas as decimal separators (frequently used throughout South America) instead of periods as decimal point separators (e.g., "11,99" instead of 11.99, with software saving as "1199" when alerts were overridden). Near-equivalent coding applied mostly to

the "reason for switching ART regimen" variable (art_rs), where selection of codes depended on auditor interpretation (e.g., "availability of more effective treatment" vs. "simplified treatment available").

3.4 Discussion

In this study we describe a novel approach where sites perform self-audits supervised by a central data coordinating center. We compared audit findings between self- and travel-auditors on a sample of 8919 entries across 65 patients capturing 28 HIV variables at three HIV clinics in Latin American. Overall error rates were similar between self- and travel-auditors, with 94% of entries receiving the same assessment from self- and travel-auditors, and the majority (72%) of incorrect variables received the same corrections from both groups. Despite some discordance and mismatched corrections between auditors, our findings suggest that data audits carried out by local investigators can provide a viable alternative to travel-audits to investigate data quality.

Monitoring data quality in a multi-site research network can be logistically challenging, costly, and time-consuming (De, 2011). The time and costs have been well-documented in the clinical trials space, where extensive source document verification is often required for government regulatory agency approval of new drugs and devices (Houston et al., 2018). Cost-savings measures in clinical trials have prompted the roll-out of "remote auditing," where trial auditors log in to the electronic medical records of distant hospitals to review and verify patient information (De, 2011; Duda et al., 2011). Such solutions are not feasible in many global heath settings, however, where electronic systems are not designed for secure remote login capacity, large geographic distances produce high internet latency, patient charts are on paper, or internet is not available.

The CCASAnet cohort faces particular challenges because the sites are dispersed in seven countries throughout North and South America. Despite the demonstrated importance of data audits (Weiss, 1998; Kimaro and Twaakyondo, 2005; Chaulagai et al., 2005; Kiragga et al., 2011; Mphatswe et al., 2012; Duda et al., 2012; Giganti et al., 2019), constraints in both time and funding have limited the extent to which data audits have been performed in this network. The self-audits described in this manuscript allowed us to collect extensive data on all eight sites using fewer resources. The self-audit involved a two-hour online training session and compensation of $2,000 US to local investigators. Local sites were given two weeks to complete the self-audit,

compared to the approximately two and a half days travel-auditors spent at each site. The number of entries audited by self-auditors was strikingly higher than our previous efforts with added benefits of lower cost and roughly equivalent resulting quality.

Although results were largely similar between self- and travel-auditors, the between-auditor discrepancies that we observed illustrate some challenges with determining correct values even with an audit. Many audit decisions are not straightforward, and neither the self- nor travel-audits should be considered a gold standard. Self- and travel-auditors contributed insights into the reasons for mismatched corrections, and, when shown the results, recommended establishing a better protocol for definitions of certain variables for data auditing and collection. Both sets of auditors felt there could be ambiguity in interpretations of specific variables (e.g., can CDC stage go from C to B, or once C is it always maintained as C?), which could contribute to finding discordance and correction mismatches. Audit entries were labeled as "correct" or "incorrect," whereas there is often some ambiguity in the true value; for example, an auditor's inability to find an original value in the source document does not necessarily imply that the original data were incorrect.

Our study has several limitations. Travel-auditors were unable to completely audit all records and the subset of doubly-audited records may not be fully representative of all records selected by the data coordinating center to be audited. This analysis excluded three patient fields of potential interest (clinical AIDS diagnosis prior to first visit (yes/no), date of prior clinical AIDS diagnosis, and WHO stage) because they were incompletely audited, which poses limitations in extending these findings to these specific variables. At some sites, self-auditors may have been involved in the original data entry, which could potentially lead to underreporting of errors. Sites were aware that their records could be externally audited, which we believe strengthened the quality of their self-audits and lessened concern of such underreporting. In addition, sites were given a small amount of compensation for completing their self-audits. Self-audits without the possibility of external auditing or without compensation may perform differently. Only three sites were doubly-audited, and sites that were not selected for a travel-audit may have had substantially different levels of concordance between self- and travel-audits.

Finally, our travel- and self-audit results may not be applicable to other settings. Settings with paperless data capture may require different source data verification techniques, such as detailed consideration of all text and data fields in an electronic health record, some of which may not have made it into the research database. We recognize that all our sites had prior experience with audits, were active research

contributors, and engaged in data quality activities. If sites are unfamiliar with data quality concepts or when scientific misconduct or fraud is a potential concern, self-audits are unlikely to be an effective solution despite cost-savings.

Our study has several strengths. This study builds upon previous data quality initiatives in a diverse, multi-national HIV cohort. Initial self- and travel-audit findings gauged the overall integrity of data at the three sites, while the comparison of doubly-audited entries investigated the efficacy of the proposed self-audits to replace travel-audits in the future. The analysis built on a large dataset, capturing many facets of the patient record, which allowed us to draw conclusions not just about the general quality of data but also inspect the integrity of specific variables and forms. Additionally, anti-fraud precautions were incorporated into the self-audit methodology: 1) patient IDs were randomly selected by the VDCC (not the sites) and 2) a random subset of the eight sites were chosen for a travel-audit, as well.

With similar overall error rates, our findings suggest that self-audits are an effective approach for assessing data quality and should be considered in networks performing analyses using pooled HIV observational data. However, discrepancies observed between corrections illustrate challenges in determining correct values even with audits. For multi-site collaborations, we recommend first conducting baseline travel-audits. After the team is familiar with sites' data quality and the audit process, we recommend regular audits, which may include self-auditing in a manner similar to that described here.

3.5 Appendix B

Table 3.3: Data dictionary for CCASAnet variables

tblBAS: Contains basic patient information typically collected at enrollment	
record_id	Study ID
male_y	Sex
birth_d	Birth date
hiv_diagnosis_d	HIV diagnosis date
mode	Mode of transmission
recart_y	Prior ART?
aids_y	AIDS diagnosis prior to first visit?
aids_d	Date of prior AIDS diagnosis
aids_cl_y	Clinical AIDS diagnosis prior to first visit?
aids_cl_d	Date of clinical AIDS diagnosis prior to first visit
baseline_d	First visit date at CCASAnet clinic
pmtct	Received PMTCT as an infant? (pediatric sites only)
birth_mode	Birth mode (pediatric sites only)
tblLTFU: Contains death and drop-out information collected during follow-up	
drop_y	Has patient been dropped from cohort?
drop_d	Date of drop
drop_rs	Reason for dropping
death_y	Did the patient die?
death_d	Date of death
tblVIS: Contains visit-related information	
visit_d	Visit date
height	Height (in cm)
weight	Weight (in kg)
cdcstage	CDC stage
whostage	WHO stage
tblLAB_CD4: Contains CD4 lab measurements	
cd4_d	CD4 date
cd4_v	CD4 value
cd4_per	CD4 percent
tblLAB_RNA: Contains viral load lab measurements	
rna_d	Viral load date
rna_v	Viral load value
tblART: Contains antiretroviral therapy (ART) information	
art_id	ART regimen
art_sd	Start date of ART regimen
art_ed	End date of ART regimen
art_rs	Reason for stopping ART regimen
tblCEP: Contains clinical events including serious non-AIDS conditions	
ce_d	Date of disease diagnosis
ce_id	Disease code

Note: The CCASAnet data structure roughly follows data exchange protocols outlined by the HIV Cohorts Data Exchange Protocol (HICDEP) and the International epidemiology Databases to Evaluate AIDS (IeDEA).

Table 3.4: Comparing baseline and follow-up characteristics of patients who were only self-audited and who were self- and travel-audited

Variable	Self- and travel-audited (n = 65)	Self-audited only (n = 65)	P-Value
Age at first clinic visit mean (sd)	15.3 years (17.9)	18.5 years (18.9)	0.55
Age at death mean (sd)	18.0 years (21.2)	31.3 years (24.5)	0.36
Number of form entries mean (sd)			
ART regimen	2.7 (2.5)	3.6 (3.0)	0.06
CD4 labs	17.4 (16.3)	21.4 (19.2)	0.09
Clinical endpoints	5.9 (9.2)	6.6 (8.0)	0.62
Viral load labs	17.8 (16.3)	20.6 (18.1)	0.2
Visits	43.4 (44.8)	57.4 (58.1)	0.15
Sex (% male)	n = 41 (63%)	n = 37 (57%)	0.52
Dead (% yes)	n = 12 (18%)	n = 14 (22%)	0.51
AIDS (% yes)	n = 20 (31%)	n = 17 (26%)	0.12
Prior ART (% yes)	n = 10 (15%)	n = 9 (14%)	0.83
Drop from cohort (% yes)	n = 22 (34%)	n = 16 (25%)	0.44
Reason for dropping[a]			
LTFU/ not known to be dead	n = 8 (36%)	n = 5 (31%)	Ref
Other	n = 8 (36%)	n = 7 (44%)	0.77
Transfer to another center	n = 6 (27%)	n = 4 (25%)	0.77
Mode of transmission			
Heterosexual Contact	n = 7 (11%)	n = 13 (20%)	Ref
Homosexual/bisexual contact	n = 9 (14%)	n = 9 (14%)	0.28
Perinatal	n = 39 (60%)	n = 31 (48%)	0.45
Other	n = 2 (3%)	n = 2 (3%)	0.86
Unknown	n = 8 (12%)	n = 10 (15%)	0.49

Note: P-values are from the Wald tests for the coefficient effects in logistic regression models for whether the patient was included in the analysis, controlling for site. [a]Proportions of reason for dropping are limited to subjects who were dropped.

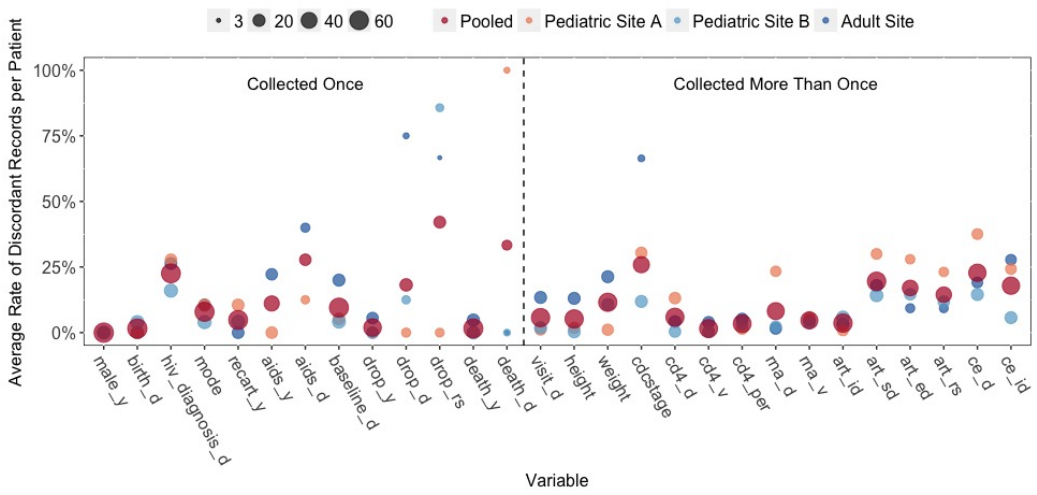


Figure 3.4: Average proportion of discordant entries per patient by variable by site (sized by number of audited patient records). For variables that were collected once, this is the proportion of patients whose entries were discordant. For variables that were collected more than once, this is the average of the per-patient proportions of entries that were discordant.

CHAPTER 4

EFFICIENT ODDS RATIO ESTIMATION UNDER TWO-PHASE SAMPLING USING ERROR-PRONE DATA FROM A MULTI-NATIONAL HIV RESEARCH COHORT

4.1 Introduction

Electronic health records (EHR) and other observational databases routinely collect a wide array of information on large numbers of patients. While initially created to support clinical care, financial billing, and insurance claims (Nordo et al., 2019), these databases are increasingly being used for clinical investigations that can influence disease prevention and policy-making. In particular, there has been an uptake in observational studies of HIV/AIDS, where patients engage in regular, frequent care, generating large amounts of routine clinical data (Zaniewski et al., 2018).

Observational databases can be error-prone since data collection is often secondary to clinical care, and the error mechanisms can be quite complicated. For example, a binary outcome of interest may be differentially misclassified; that is, the sensitivity and specificity may depend on other variables such as study site. Errors in continuous variables can be even more so: they can be additive or multiplicative, symmetric or skewed, and centered around zero or systematically biased. In addition, errors can be correlated across multiple error-prone variables. Naive logistic regression analyses could yield biased odds ratio (OR) and standard error estimates in the presence of outcome misclassification and/or covariate errors (Barron, 1977; Copeland et al., 1977; Quake et al., 1980; Neuhaus, 1999).

To ensure data integrity, clinical trials have long relied on source document verification and data auditing. Observational studies have begun to advocate for these procedures as well (Duda et al., 2012; Giganti et al., 2019; Lotspeich et al., 2020). In a typical data audit, external, trained auditors visit the site, compare existing records to clinical source documents, and report discrepancies. Complete data re-entry would be ideal, but could be too resource-intensive, especially for large databases like EHR. A cost-effective alternative is the partial audit or two-phase design, wherein error-prone variables are observed for all subjects from the research database during Phase I, and then this information is used to select a validation subsample in Phase II. Thus, the available data consist of validated records for subjects chosen in Phase II and unvalidated records for everyone else. Two-phase sampling greatly reduces the burden of data validation and thus has been used in several large-scale observational studies,

including the Caribbean, Central, and South America Network for HIV Epidemiology (CCASAnet).

CCASAnet is a multi-site research network that uses routinely collected HIV patient care data to address questions about the characteristics of the HIV epidemic and to improve the quality and consistency of clinical research activities across Latin America. Individual-level data from CCASAnet clinical sites are sent to the Data Coordinating Center at Vanderbilt University (VDCC) in Nashville, Tennessee, where they are compiled into a research database (McGowan et al., 2007). Participating sites are regularly audited by the VDCC to ensure data quality; we focus on one round of audits from 2013–2014 (Giganti et al., 2019).

We are interested in estimating the associations between risk factors CD4 count and AIDS, both measured at the time of antiretroviral therapy (ART) initiation, and the odds of subsequently developing an AIDS-defining event (ADE) within two years after initiating ART. Error-prone data were available for 5109 patients in the collaborative CCASAnet database (Phase I sample), and audit data were available for a site-stratified simple random sample of 117 patients (Phase II sample). The audits revealed many data errors. Giganti et al. (2019) noted a higher prevalence of ADE in the audit database than in the original research database. In addition, from previous audits we have learned that date variables tend to be particularly prone to errors. Since the outcome (ADE within two years of ART initiation) and primary predictors (CD4 and AIDS diagnosis at ART initiation) were derived based on the date of ART initiation, errors in the date of ART initiation could induce dependent errors in these key study variables. More details on the definitions of these variables can be found in Section 4.4. Here we propose a method to combine the audit data with the pre-audit data to obtain unbiased and efficient OR estimators.

Statistical methods have been developed to obtain valid inference from error-prone data under a two-phase design. The majority of those methods address classical measurement error in covariates only (Carroll et al., 2006) or binary outcome misclassification alone (Magder and Hughes, 1997; Neuhaus, 1999). Fewer methods have been developed to handle outcome misclassification and covariate measurement error simultaneously. When sensitivity and specificity or false positive and false negative rates (FPR and FNR, respectively) of a binary outcome and covariate are known or can be reliably estimated, the matrix method (Barron, 1977) or inverse matrix method (Marshall, 1990) can be extended to correct naive OR estimates based on misclassified data (Tang et al., 2013). Fully-parametric maximum likelihood estimators (MLE) have been proposed for outcome and/or binary covariate misclassification

(Tang et al., 2015), but they do not accommodate error-prone continuous covariates. In addition, they require explicit specification of the distributions of measurement errors, which may render bias when models are misspecified (Robins et al., 1994). Design-based estimators, such as the Horvitz–Thompson (HT) (Horvitz and Thompson, 1952) and generalized raking (Deville et al., 1993) estimators, can also be used (Lumley et al., 2011). While these estimators do not require specification of the error models, they tend to be less efficient than model-based estimators, especially when the Phase II sampling probabilities are extremely unequal.

In this manuscript, we propose a general framework to accommodate a misclassified binary outcome and error-prone categorical or continuous covariates under a two-phase design. Our methods handle settings where the error rates depend on other error-prone and error-free covariates. We model the covariate error distribution nonparametrically with B-spline sieves (Grenander, 1981) to accommodate continuous covariates subject to arbitrary measurement error patterns. We develop a computationally efficient and numerically stable EM algorithm to maximize the resulting semiparametric likelihood. Our estimators are shown to be consistent, asymptotically normal, and asymptotically efficient.

4.2 Methods

Consider a binary outcome, Y, and vector of continuous or categorical covariates, $\boldsymbol{X}$, which are assumed to be related through the logistic regression model $P_{\boldsymbol{\theta}}(Y = 1|\boldsymbol{X}) = [1 + \exp\{-(\alpha + \boldsymbol{X}\boldsymbol{\beta})\}]^{-1}$, where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$. Error-prone measures of the outcome and covariates recorded in the database are denoted by Y^* and $\boldsymbol{X}^*$, respectively. Throughout the paper, we use "error-prone" to describe both categorical variables subject to misclassification and continuous variables subject to measurement error. A complete, validated observation $(Y^*, \boldsymbol{X}^*, Y, \boldsymbol{X})$ is assumed to be generated from

$$P(Y^*, \boldsymbol{X}^*, Y, \boldsymbol{X}) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})P(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})P(\boldsymbol{X}|\boldsymbol{X}^*)P(\boldsymbol{X}^*), \tag{4.1}$$

where $P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})$ is the logistic regression model of primary interest, $P(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})$ is the conditional probability of Y^* given $(\boldsymbol{X}^*, Y, \boldsymbol{X})$, $P(\boldsymbol{X}|\boldsymbol{X}^*)$ is the conditional density of $\boldsymbol{X}$ given $\boldsymbol{X}^*$, and $P(\boldsymbol{X}^*)$ is the marginal density of $\boldsymbol{X}^*$. Conditional independence of Y and $\boldsymbol{X}^*$ given $\boldsymbol{X}$ is assumed, such that $P(Y|\boldsymbol{X}, \boldsymbol{X}^*) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})$. No additional assumptions are made, and expression (4.1) allows for differential outcome misclassification and correlated covariate errors. This general setup covers the clas-

sical scenarios with either (1) outcome misclassification only, letting $\boldsymbol{X}^* = \boldsymbol{X}$ and $P(Y^*, Y, \boldsymbol{X}) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})P(Y^*|Y, \boldsymbol{X})P(\boldsymbol{X})$, or (2) covariate measurement error only, setting $Y^* = Y$ and $P(\boldsymbol{X}^*, Y, \boldsymbol{X}) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})P(\boldsymbol{X}|\boldsymbol{X}^*)P(\boldsymbol{X}^*)$. Error-free covariates can also be included by decomposing $\boldsymbol{X} = (\boldsymbol{X}_a^T, \boldsymbol{X}_b^T)^T$ and $\boldsymbol{X}^* = (\boldsymbol{X}_a^T, \boldsymbol{X}_b^{*T})^T$, where subscripts a and b denote error-free and error-prone covariates, respectively. Thus, complete observations are assumed to be generated from

$$P(Y^*, \boldsymbol{X}^*, Y, \boldsymbol{X}) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X}_a, \boldsymbol{X}_b)P(Y^*|\boldsymbol{X}_b^*, Y, \boldsymbol{X}_a, \boldsymbol{X}_b)P(\boldsymbol{X}_b|\boldsymbol{X}_b^*, \boldsymbol{X}_a)P(\boldsymbol{X}_b^*, \boldsymbol{X}_a).$$

With complete data on all N subjects, estimates can be obtained by maximizing the usual likelihood $\prod_{i=1}^N P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i)$. In a two-phase study, however, the likelihood contributions of validated and unvalidated subjects are different. While validated subjects contribute complete data to the likelihood via equation (4.1), unvalidated subjects can only contribute incomplete data to the likelihood by marginalizing out the unobserved Phase II variables. Let V denote the validation indicator, where $V_i = 1$ if subject i was selected for Phase II and 0 otherwise. The observed-data log-likelihood can be expressed as

$$\begin{aligned}
&\sum_{i=1}^N V_i \Big\{ \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i) + \log P(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i) + \log P(\boldsymbol{X}_i|\boldsymbol{X}_i^*) \Big\} \\
&+ \sum_{i=1}^N (1 - V_i) \log \left\{ \sum_{y=0}^1 \int_{\boldsymbol{x}} P_{\boldsymbol{\theta}}(y|\boldsymbol{x}) P(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}) P(\boldsymbol{x}|\boldsymbol{X}_i^*) d\boldsymbol{x} \right\}. \qquad (4.2)
\end{aligned}$$

Because $\boldsymbol{X}^*$ is fully observed at Phase I, $P(\boldsymbol{X}^*)$ can be ignored in the inference of $\boldsymbol{\theta}$. Further, two-phase designs assume that the selection of the Phase II sample depends on Phase I variables Y^* and $\boldsymbol{X}^*$ only. Thus, the Phase II variables are missing at random and the distribution of V can be omitted from expression (4.2).

Our primary interest lies in the inference of $\boldsymbol{\theta}$, the true conditional log OR of $\boldsymbol{X}$ on Y. The unknown error mechanisms $P(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})$ and $P(\boldsymbol{X}|\boldsymbol{X}^*)$ are nuisance parameters, rarely of interest on their own. Tang et al. (2015) proposed a fully-parametric MLE for misclassified Y^* with a single binary error-prone covariate X_b and error-free covariates $\boldsymbol{X}_a$. They instead factored the joint density as $P(Y^*, X_b^*, Y, X_b, \boldsymbol{X}_a) = P_{\boldsymbol{\theta}}(Y|X_b, \boldsymbol{X}_a)P(Y^*|X_b^*, Y, X_b, \boldsymbol{X}_a)P(X_b^*|Y, X_b, \boldsymbol{X}_a)P(X_b|\boldsymbol{X}_a)$ $P(\boldsymbol{X}_a)$, modeled the first four terms with four logistic regressions, and ignored the last term because $\boldsymbol{X}_a$ was fully observed at Phase I. This approach explicitly specifies all error mechanisms, and thus could perform poorly if not done so correctly. In particular, the MLE will be biased if errors are incorrectly assumed to be independent

of other variables. Usually, little is known about how errors were introduced to data, so proper specification of error models can be challenging, especially with continuous error-prone covariates.

As we extend to settings with continuous covariate error, we aim to develop a more robust estimator for $\boldsymbol{\theta}$ by requiring fewer assumptions on the error mechanisms. Since we already assume that $Y|\boldsymbol{X}$ follows a logistic model, it is reasonable to assume a logistic model for the outcome error mechanism $P_{\boldsymbol{\gamma}}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})$ in a similar manner, where $\boldsymbol{\gamma}$ denotes the model parameters. The algorithm derived herein treats this model generally and does not dictate the form of the linear predictor. No assumption is made about the distribution of $\boldsymbol{X}|\boldsymbol{X}^*$, i.e., $P(\boldsymbol{X}|\boldsymbol{X}^*)$ is estimated nonparametrically. Let m denote the number of distinct values of $\boldsymbol{X}$ in the validation sample, and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ denote these values. For each $\boldsymbol{X}^* = \boldsymbol{x}^*$, we use discrete probability functions to estimate $P(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{X}^* = \boldsymbol{x}^*)$ $(k = 1, \ldots, m)$. This works when $\boldsymbol{X}^*$ is categorical. However, with continuous components of $\boldsymbol{X}^*$, few validated subjects will have $\boldsymbol{X}^* = \boldsymbol{x}^*$ for each distinct $\boldsymbol{x}^*$. In this situation, this nonparametric estimator for $P(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{X}^* = \boldsymbol{x}^*)$ is not directly applicable and smoothing techniques are required.

We extend Tao et al. (2017) and use the method of sieves (Grenander, 1981) to handle error-prone continuous covariates. Specifically, B-splines (Schumaker, 1981) are used to approximate the covariate error mechanism. We note that B-splines have been used to promote robustness in measurement error settings elsewhere (Staudenmayer et al., 2008; Sarkar et al., 2014). The error-prone covariates are assumed to have bounded support. Without loss of generality, each component X^* in $\boldsymbol{X}^*$ is standardized to have support on the interval [0,1]. Let q and b_N denote the order and number of interior knots for the B-splines basis, respectively, and let $\{t_{-q+1}, \ldots, t_{q+b_N}\}$ denote the knots. The b_N interior knots are assumed to be evenly spaced across the range of $\boldsymbol{X}^*$; this can be revised in practice to best suit the covariate data. Then, the interval [0,1] can be partitioned around the knots as $\Delta \equiv \{t_{-q+1} = \ldots\ t_{-1} = 0 = t_0 < t_1 < \cdots < t_{b_N+1} = 1 = \cdots = t_{q+b_N}\}$. For one covariate X^* in $\boldsymbol{X}^*$, the q order B-spline basis associated with the partition Δ is denoted by $\{N_l^q(X^*)\}_{l=-q+1}^{b_N}$, where $N_l^q(X^*)$ corresponds to the lth B-spline basis function of order q and is defined according to the recursive formula $N_l^q(X^*) = \frac{X^* - t_l}{t_{l+q-1} - t_l} N_l^{q-1}(X^*) + \frac{t_{l+q} - X^*}{t_{l+q} - t_{l+1}} N_{l+1}^{q-1}(X^*)$ $(l = -q+1, \ldots, b_N)$. The first order B-spline basis function is defined as the histogram function $N_l^1(X^*) = I(t_l \leq X^* < t_{l+1})$ with $I(\cdot)$ being the indicator function. The multivariate B-spline basis function on the full set of d covariates $\boldsymbol{X}^* \equiv (X_1^*, \ldots, X_d^*)^{\mathrm{T}}$ is defined as $B_j^q(\boldsymbol{X}^*) = \prod_{s=1}^d N_{l_s}^q(X_s^*)$ $(s = 1, \ldots, d;\ l_s = -q+1, \ldots, b_N;\ j = 1, \ldots, s_N)$, where $s_N = (b_N + q)^d$ represents the total number of multivariate B-spline basis func-

tions. In essence, the scalar index j replaces the multivariate index $(l_1, \ldots, l_d)$ to simplify notation.

We approximate $\log P(\boldsymbol{X}_i|\boldsymbol{X}_i^*)$ and $P(\boldsymbol{x}_k|\boldsymbol{X}_i^*)$ in expression (4.2) with

$$\sum_{k=1}^{m} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \sum_{j=1}^{s_N} \log p_{kj} B_j^q(\boldsymbol{X}_i^*) \text{ and } \sum_{k=1}^{m} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \sum_{j=1}^{s_N} p_{kj} B_j^q(\boldsymbol{X}_i^*), \tag{4.3}$$

respectively, where p_{kj} is the coefficient of $B_j^q(\boldsymbol{X}^*)$ at $\boldsymbol{x}_k$ $(k = 1, \ldots, m;\ j = 1, \ldots, s_N)$. Thus, using the approximations in expression (4.3) the observed-data log-likelihood (expression (4.2)) can be rewritten as $l_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \{p_{kj}\})$

$$\begin{aligned} &= \sum_{i=1}^{N} V_i \left\{ \log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i) + \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i) + \sum_{k=1}^{m} \sum_{j=1}^{s_N} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \log p_{kj} B_j^q(\boldsymbol{X}_i^*) \right\} \\ &\quad + \sum_{i=1}^{N} (1 - V_i) \log \left\{ \sum_{y=0}^{1} \sum_{k=1}^{m} P_{\boldsymbol{\theta}}(y|\boldsymbol{x}_k) P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k) \sum_{j=1}^{s_N} p_{kj} B_j^q(\boldsymbol{X}_i^*) \right\}, \end{aligned} \tag{4.4}$$

which is to be maximized with respect to $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and $\{p_{kj}\}$ under constraints

$$\sum_{k=1}^{m} p_{kj} = 1, \text{ and } p_{kj} \geq 0, \ (j = 1, \ldots, s_N; k = 1, \ldots, m). \tag{4.5}$$

The maximization of the right-hand side of expression (4.4) is carried out through an EM algorithm (Dempster et al., 1977).

Remark. *The selection of b_N and q need to satisfy condition (C4) in Section 4.6.1. Specifically, b_N should increase at a much slower rate than the Phase I and Phase II sample sizes, and q should increase with the dimension of the error-prone continuous covariates but is usually restricted to be less than or equal to four (corresponding to cubic splines). In practice, b_N and q can be chosen in a data-adaptive manner such as cross-validation. For any fixed b_N and q, one evaluates expression (4.4) in the validation fold using estimates obtained from the training folds. The optimal b_N and q are those that maximize the average cross-validation likelihood. For the purposes of this paper, and to save computation time, we choose q and b_N such that the model is stable within a reasonable range of b_N values.*

4.2.1 EM Algorithm

Direct maximization of the observed-data log-likelihood is difficult due to the intractable term of the likelihood contribution of unvalidated subjects. Following Tao

et al. (2017), we devise an EM algorithm to maximize expression (4.4). Specifically, a latent variable $Z \in \{1/s_N, 2/s_N \ldots, 1\}$ that satisfies the constraints $P(Z = j/s_N|\boldsymbol{X}^*) = B_j^q(\boldsymbol{X}^*)$, $P(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{X}^*, Z = j/s_N) = P(\boldsymbol{X} = \boldsymbol{x}_k|Z = j/s_N) = p_{kj}$, $P(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X}, Z) = P_{\boldsymbol{\gamma}}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})$, and $P(Y|\boldsymbol{X}, Z) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})$ is introduced for unvalidated subjects such that the corresponding observed-data log-likelihood can be interpreted as

$$\begin{aligned}
&\sum_{i=1}^{N}(1-V_i)\log\left\{\sum_{y=0}^{1}\sum_{k=1}^{m}P_{\boldsymbol{\theta}}(y|\boldsymbol{x}_k)P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k)\sum_{j=1}^{s_N}p_{kj}B_j^q(\boldsymbol{X}_i^*)\right\} \\
=&\sum_{i=1}^{N}(1-V_i)\log\left\{\sum_{y=0}^{1}\sum_{k=1}^{m}\sum_{j=1}^{s_N}P(Y_i^*, \boldsymbol{X}_i^*, y, \boldsymbol{x}_k, Z = j/s_N)\right\}, \qquad (4.6)
\end{aligned}$$

i.e., assuming complete data consist of $(Y^*, \boldsymbol{X}^*, Y, \boldsymbol{X}, Z)$ but with $(Y, \boldsymbol{X}, Z)$ missing. Thus, the complete-data log-likelihood is

$$\begin{aligned}
&\sum_{i=1}^{N}V_i\left\{\log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i) + \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i) + \log P(\boldsymbol{X}_i|\boldsymbol{X}_i^*)\right\}+ \\
&\sum_{i=1}^{N}(1-V_i)\left\{\log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i) + \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i) + \log P(\boldsymbol{X}_i|\boldsymbol{X}_i^*, Z_i) + \log P(Z_i|\boldsymbol{X}_i^*)\right\} \\
=&\sum_{i=1}^{N}V_i\left\{\log P_{\boldsymbol{\theta}}(Y_i|\boldsymbol{X}_i) + \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i) + \sum_{k=1}^{m}I(\boldsymbol{X}_i = \boldsymbol{x}_k)\sum_{j=1}^{s_n}\log p_{kj}B_j^q(\boldsymbol{X}_i^*)\right\}+ \\
&\sum_{i=1}^{N}(1-V_i)\left[\sum_{y=0}^{1}\sum_{k=1}^{m}I(Y_i = y, \boldsymbol{X}_i = \boldsymbol{x}_k)\left\{\log P_{\boldsymbol{\theta}}(y|\boldsymbol{x}_k) + \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k)\right\}\right. \\
&\qquad\left.+\sum_{k=1}^{m}\sum_{j=1}^{s_N}I(\boldsymbol{X}_i = \boldsymbol{x}_k, Z_i = j/s_N)\log p_{kj} + \sum_{j=1}^{s_N}I(Z_i = j/s_n)\log B_j^q(\boldsymbol{X}_i^*)\right]. \qquad (4.7)
\end{aligned}$$

In the E-step, conditional expectations of $I(Y_i = y, \boldsymbol{X}_i = \boldsymbol{x}_k)$ and $I(\boldsymbol{X}_i = \boldsymbol{x}_k, Z_i = j/s_n)$ are calculated given Phase I data. Specifically, in the $(t+1)$th iteration we calculate

$$\begin{aligned}
\hat{\psi}_{kyji}^{(t+1)} \equiv& E\left\{I(Y_i = y, \boldsymbol{X}_i = \boldsymbol{x}_k, Z_i = j/s_N)|Y_i^*, \boldsymbol{X}_i^*\right\} \\
=&\frac{P_{\hat{\boldsymbol{\theta}}^{(t)}}(y|\boldsymbol{x}_k)P_{\hat{\boldsymbol{\gamma}}^{(t)}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k)\hat{p}_{kj}^{(t)}B_j^q(\boldsymbol{X}_i^*)}{\sum_{y'=0}^{1}\sum_{k'=1}^{m}P_{\hat{\boldsymbol{\theta}}^{(t)}}(y'|\boldsymbol{x}_{k'})P_{\hat{\boldsymbol{\gamma}}^{(t)}}(Y_i^*|\boldsymbol{X}_i^*, y', \boldsymbol{x}_{k'})\sum_{j'=1}^{s_N}\hat{p}_{k'j'}^{(t)}B_{j'}^q(\boldsymbol{X}_i^*)}
\end{aligned}$$

for each unvalidated subject. Then, estimates of $E\left\{I(Y_i = y, \boldsymbol{X}_i = \boldsymbol{x}_k)|Y_i^*, \boldsymbol{X}_i^*\right\}$ and

$E\{I(\boldsymbol{X}_i = \boldsymbol{x}_k, Z_i = j/s_N)|Y_i^*, \boldsymbol{X}_i^*\}$ are obtained as

$$\hat{w}_{kyi}^{(t+1)} = \begin{cases} I(Y_i = y, \boldsymbol{X}_i = \boldsymbol{x}_k) & \text{if } V_i = 1, \\ \sum_{j=1}^{s_N} \hat{\psi}_{kyji}^{(t+1)} & \text{if } V_i = 0, \end{cases} \tag{4.8}$$

and

$$\hat{u}_{kji}^{(t+1)} = \sum_{y=0}^{1} \hat{\psi}_{kyji}^{(t+1)} \text{ if } V_i = 0, \tag{4.9}$$

respectively. Substituting equations (4.8) and (4.9) into the right-hand side of equation (4.7) yields the following objective function for the M-step:

$$\sum_{i=1}^{N} \Bigg[\sum_{y=0}^{1} \sum_{k=1}^{m} \hat{w}_{kyi}^{(t+1)} \{\log P_{\boldsymbol{\theta}}(y|\boldsymbol{x}_k) + \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k)\} + \sum_{k=1}^{m} \sum_{j=1}^{s_N} \log p_{kj} \left\{ V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*) + (1 - V_i) \hat{u}_{kji}^{(t+1)} \right\} \Bigg]. \tag{4.10}$$

In the M-step, $\hat{\boldsymbol{\theta}}^{(t+1)}$ and $\hat{\boldsymbol{\gamma}}^{(t+1)}$ are updated by maximizing

$$\sum_{i=1}^{N} \sum_{y=0}^{1} \sum_{k=1}^{m} \hat{w}_{kyi}^{(t+1)} \log P_{\boldsymbol{\theta}}(y|\boldsymbol{x}_k) \text{ and } \sum_{i=1}^{N} \sum_{y=0}^{1} \sum_{k=1}^{m} \hat{w}_{kyi}^{(t+1)} \log P_{\boldsymbol{\gamma}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k),$$

respectively, both of which are log-likelihoods of weighted logistic regression models. The nuisance parameters $\hat{p}_{kj}^{(t+1)}$ are updated by maximizing

$$\sum_{i=1}^{N} \sum_{k=1}^{m} \sum_{j=1}^{s_N} \log p_{kj} \left\{ V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*) + (1 - V_i) \hat{u}_{kji}^{(t+1)} \right\},$$

such that

$$\hat{p}_{kj}^{(t+1)} = \frac{\sum_{i=1}^{N} \left\{ V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*) + (1 - V_i) \hat{u}_{kji}^{(t+1)} \right\}}{\sum_{k'=1}^{m} \sum_{i=1}^{N} \left\{ V_i I(\boldsymbol{X}_i = \boldsymbol{x}_{k'}) B_j^q(\boldsymbol{X}_i^*) + (1 - V_i) \hat{u}_{k'ji}^{(t+1)} \right\}}$$

$(k = 1, \ldots, m; j = 1, \ldots, s_N)$. Notice that the constraints of $\{p_{kj}\}$ from expression (4.5) are satisfied in each iteration. Starting with initial values $\hat{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$, $\hat{\boldsymbol{\gamma}}^{(0)} = \mathbf{0}$, and $\hat{p}_{kj}^{(0)} = \sum_{i=1}^{n} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*) / \sum_{i=1}^{n} V_i B_j^q(\boldsymbol{X}_i^*)$, the sieve maximum likelihood estimators (SMLE) $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\gamma}}$, and $\{\hat{p}_{kj}\}$ are obtained by iterating between the E- and M-steps until convergence.

4.2.2 Asymptotic Properties

Let the true values of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and F (the joint cumuulative distribution function of $\boldsymbol{X}$ and $\boldsymbol{X}^*$) be denoted $\boldsymbol{\theta}_0$, $\boldsymbol{\gamma}_0$, and F_0, respectively. We impose the following regularity conditions:

(C1) The set of covariates $(\boldsymbol{X}, \boldsymbol{X}^*)$ has bounded support.

(C2) If there exist two sets of parameters $(\boldsymbol{\theta}_1, \boldsymbol{\gamma}_1, F_1)$ and $(\boldsymbol{\theta}_2, \boldsymbol{\gamma}_2, F_2)$ such that

$$P_{\boldsymbol{\theta}_1}(Y|\boldsymbol{X})P_{\boldsymbol{\gamma}_1}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})F_1(\boldsymbol{X}, \boldsymbol{X}^*) = P_{\boldsymbol{\theta}_2}(Y|\boldsymbol{X})P_{\boldsymbol{\gamma}_2}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})F_2(\boldsymbol{X}, \boldsymbol{X}^*),$$

where $(Y, \boldsymbol{X}, Y^*, \boldsymbol{X}^*) \in \mathcal{C} \equiv \{(y, \boldsymbol{x}, y^*, \boldsymbol{x}^*) : P(V = 1|y^*, \boldsymbol{x}^*) > 0\}$, then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$, and $F_1 = F_2$. Further, if there exist a vector of constants $\boldsymbol{c}$ such that

$$\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{P_{\boldsymbol{\theta}_0}(y_1|\boldsymbol{x})}{P_{\boldsymbol{\theta}_0}(y_2|\boldsymbol{x})} + \frac{\partial}{\partial \boldsymbol{\gamma}} \log \frac{P_{\boldsymbol{\gamma}_0}(y_1^*|\boldsymbol{x}^*, y_1, \boldsymbol{x})}{P_{\boldsymbol{\gamma}_0}(y_2^*|\boldsymbol{x}^*, y_2, \boldsymbol{x})} \right\}^T \boldsymbol{c} = 0$$

for any $(y_i^*, \boldsymbol{x}^*, y_i, \boldsymbol{x}) \in \mathcal{C}$, $i = 1, 2$, then $\boldsymbol{c} = \boldsymbol{0}$.

(C3) The distribution function F_0 is positive in its support and q-times differentiable with respect to a suitable measure.

(C4) As $N \to \infty$, $s_N \to \infty$, and $N^{1/2} s_N^{-q/d} \to 0$.

(C5) The function $E(V|\boldsymbol{X}, \boldsymbol{X}^*)$ is q-times continuously differentiable with respect to $\boldsymbol{X}$ and $\boldsymbol{X}^*$.

We state the asymptotic results for the proposed SMLE in two theorems, with proof provided in Section 4.6.1.

Theorem 4.2.1. *Under conditions (C1)-(C5),*

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|| + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| + \underset{\boldsymbol{x}, \boldsymbol{x}^*}{sup} |\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*) - F_0(\boldsymbol{x}, \boldsymbol{x}^*)| \to 0$$

almost surely.

Theorem 4.2.2. *Under conditions (C1)-(C5), $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance attains the semiparametric efficiency bound.*

In short, the proposed estimators possess desirable statistical properties: they are consistent, asymptotically normal, and asymptotically efficient.

4.2.3 Variance Estimation

Variance estimator of $\hat{\boldsymbol{\theta}}$ is obtained by inverting the observed profile information via the profile likelihood method of Murphy and Van der Vaart (2000). The profile likelihood is defined by $pl(\boldsymbol{\theta}) = \sup_{\boldsymbol{\gamma},\{p_{kj}\}} l_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \{p_{kj}\})$. In practice, $pl(\boldsymbol{\theta})$ is obtained by holding $\boldsymbol{\theta}$ fixed in the EM algorithm and calculating the observed-data log-likelihood at convergence. The (k,l)th element of the observed profile information $I\left(\hat{\boldsymbol{\theta}}\right)$ is calculated as

$$-\frac{1}{h_N^2}\left[pl\left(\hat{\boldsymbol{\theta}} + \boldsymbol{e}_k h_N + \boldsymbol{e}_l h_N\right) - pl\left(\hat{\boldsymbol{\theta}} + \boldsymbol{e}_k h_N\right) - pl\left(\hat{\boldsymbol{\theta}} + \boldsymbol{e}_l h_N\right) + pl\left(\hat{\boldsymbol{\theta}}\right)\right]$$

where h_N is a constant of order $N^{-1/2}$ and $\boldsymbol{e}_k$ is the kth canonical vector. The variance $Var\left(\hat{\boldsymbol{\theta}}\right)$ is estimated by $\left\{I\left(\hat{\boldsymbol{\theta}}\right)\right\}^{-1}$.

4.3 Simulation Studies

Our simulation studies begin with the setting of an error-prone outcome and binary covariate (Section 4.3.1). The SMLE, fully-parametric MLE (Tang et al., 2015), complete-case analysis, and design-based generalized raking and HT estimators are compared. Next, binary outcome misclassification and a continuous error-prone covariate are considered (Section 4.3.2), illustrating the performance of the SMLE under settings where the MLE does not apply. Method performance is assessed on bias, confidence interval (CI) coverage, and efficiency. Error-free covariates are included. All settings focus on estimation of β: the conditional log OR for X_b on Y. The bootstrap method of Koehler et al. (2009) was used to estimate the Monte Carlo simulation errors for bias and coverage.

4.3.1 Error-Prone Binary Covariate

True binary covariate X_b, error-free covariate X_a, and true outcome Y were generated from Bernoulli distributions with $P(X_b = 1) = 0.5$, $P(X_a = 1) = 0.25$, and $P(Y = 1|X_b, X_a) = [1 + \exp\{-(-0.65 - 0.2X_b - 0.1X_a)\}]^{-1}$. Error-prone X_b^* and Y^* were generated from Bernoulli distributions with $P(X_b^* = 1|X_b, Y, X_a) = [1 + \exp\{-(-1.1 + 2.2X_b + 0.5X_a)\}]^{-1}$ and $P(Y^* = 1|X_b^*, Y, X_b, X_a) = [1 + \exp\{-(-2.2 - 0.2X_b^* + 5.14Y - 0.2X_b - 0.1X_a)\}]^{-1}$. These settings followed from Tang et al. (2015), except that conditional independence between X_b^* and Y given (X_b, X_a) was assumed here. There was approximately 32% and 35% prevalence of Y and Y^*, respectively,

and 8% and 25% misclassification in Y^* (FPR = 8%; FNR = 6%) and X_b^* (FPR = 28%; FNR = 23%), respectively. From $N = 1000$ and 2000 Phase I subjects, proportions of $p_v = 0.1$, 0.25, or 0.5 were selected for validation through simple random sampling (SRS) or 1:1 case-control sampling based on Y^* (naive case-control). The likelihood for the fully-parametric MLE (Tang et al., 2015) was correctly specified for the data generation scheme, and the *nlm* function in R (R Core Team, 2019) was used to obtain the estimates. For generalized raking, we calibrated sampling weights to the naive, error-prone influence functions following Chen and Lumley (2020) and used the *survey* package in R (Lumley, 2019).

Results for these settings are presented in Table 4.1. All methods were essentially unbiased. The average standard error estimates (SEE) for the SMLE were reasonably close to the empirical standard error (SE), and improved with increasing N. As expected, for a fixed N, increasing p_v decreased both bias and standard error. Results were similar with $N = 5000$ (Table 4.6). In comparison, a naive analysis using only the error-prone Phase I data yielded an average bias of 14%. All estimators had smaller standard errors under naive case-control than SRS. The SMLE was as efficient as the MLE, suggesting that the added robustness of the SMLE came at little cost. The complete-case, HT, and raking estimators, by comparison, lost as much as 31%, 33%, and 23% efficiency to the SMLE, respectively. Our EM algorithm was stable, with convergence rates $\geq 99\%$ for audit sizes $n > 100$. Additional simulations, reported in Section 4.6.2.2, compared the SMLE with the MLE under model misspecification. In general, these simulations suggest that with binary X_b/X_b^*, both estimators behaved similarly and were fairly robust to misspecification.

4.3.2 Error-Prone Continuous Covariate

4.3.2.1 Varying covariate error variance

Continuous covariate X_b was generated from a standard normal distribution. Error-free covariate X_a and outcome Y were generated from Bernoulli distributions with $P(X_a = 1) = 0.25$ and $P(Y = 1|X_b, X_a) = [1 + \exp\{-(-1 + X_b - 0.5X_a)\}]^{-1}$. Error-prone covariate X_b^* was constructed as $X_b^* = X_b + U$, where U was a normal random variable with mean zero and variance σ_U^2. We considered σ_U^2 values of 0.1, 0.25, 0.5, and 1, corresponding to correlations of 0.95, 0.9, 0.82, and 0.71, respectively, between X_b and X_b^*. The error-prone outcome Y^* was generated from a Bernoulli distribution with $P(Y^* = 1|X_b^*, Y, X_b, X_a) = [1 + \exp\{-(-2.2 + X_b^* + 5.14Y + X_b - 0.5X_a)\}]^{-1}$. This simulation setup yielded approximately 28% and 37% prevalence of

Table 4.1: Simulation results for outcome misclassification and a binary error-prone covariate for increasing Phase I sample size N and audit proportion p_v

		SMLE				MLE			Complete-Case			HT			Raking		
N	p_v	Bias	SE	SEE	CP	Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE	RE
								SRS									
1000	0.10	−0.003	0.381	0.369	0.950	0.010	0.376	1.027	−0.012	0.452	0.711	−0.012	0.452	0.711	−0.008	0.419	0.825
	0.25	0.007	0.247	0.242	0.950	0.007	0.246	1.008	−0.008	0.283	0.732	−0.008	0.283	0.732	−0.003	0.271	0.831
	0.50	0.001	0.183	0.181	0.938	0.001	0.183	1.000	0.000	0.193	0.902	0.000	0.193	0.902	−0.001	0.187	0.954
2000	0.10	0.002	0.271	0.258	0.946	0.004	0.270	1.007	−0.008	0.310	0.762	−0.008	0.310	0.762	−0.006	0.285	0.905
	0.25	0.001	0.177	0.171	0.941	0.001	0.177	1.000	−0.005	0.195	0.820	−0.005	0.195	0.820	−0.008	0.178	0.984
	0.50	0.000	0.128	0.128	0.954	0.000	0.128	1.000	−0.007	0.133	0.924	−0.007	0.133	0.924	−0.006	0.128	1.001
								1:1 case-control sampling based on Y^*									
1000	0.10	0.018	0.366	0.343	0.936	0.020	0.363	1.017	−0.015	0.423	0.750	−0.026	0.430	0.724	−0.022	0.393	0.869
	0.25	−0.011	0.239	0.228	0.943	−0.011	0.238	1.008	0.001	0.263	0.824	−0.009	0.267	0.799	−0.010	0.253	0.895
	0.50	−0.001	0.169	0.171	0.954	−0.001	0.169	1.000	0.016	0.183	0.856	0.009	0.186	0.827	0.007	0.181	0.874
2000	0.10	−0.008	0.242	0.240	0.952	−0.009	0.241	1.008	0.002	0.291	0.690	−0.008	0.295	0.671	−0.013	0.276	0.770
	0.25	0.004	0.163	0.161	0.949	0.004	0.162	1.012	0.003	0.183	0.800	−0.005	0.184	0.784	−0.002	0.169	0.930
	0.50	0.005	0.118	0.121	0.957	0.005	0.118	1.000	0.002	0.129	0.835	−0.006	0.131	0.811	−0.006	0.125	0.894

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the average of the standard error estimator; CP is the coverage probability of the 95% confidence interval. RE is the relative efficiency of the estimator to the SMLE. Each entry is based on 1000 replicates. Convergence rates for the SMLE with an audit size of $n = 100$ were 89% and 94%, respectively, under SRS and 1:1 case-control sampling based on Y^*. This was due to complete or quasi-complete separation of the outcome error model $P(Y^*|X_b^*, Y, X_b, C)$ in these settings. The SMLE had greater than 99% convergence rates in other settings. The Monte Carlo simulation error for the bias and CP of the SMLE did not exceed 0.013 and 0.8%, respectively.

Y and Y^*, respectively, regardless of the choice of σ_U^2, and 12%–13% misclassification in Y^* (FPR =14%–15%; FNR = 6%–7%). We used a cubic B-spline basis ($q = 4$) and varied b_N from 16 to 28 to assess its effects on model fitting, maintaining a 3:1 ratio of knots allocated to subjects with $X_a = 0$:$X_a = 1$. This ratio allocated the knots proportionally to the available data, distributing 25% of the knots to the 25% of subjects with $X_a = 1$. When $N = 1000$, the results were very similar for $b_N \geq 20$, i.e., the maximum difference in the coverage probability of the 95% CI was less than 0.5%. Consequently, separate cubic B-splines with 15 and 5 interior knots were used for subjects with $X_a = 0$ and $X_a = 1$, respectively; when $N = 2000$, 18 and 6 interior knots were used.

Simulation results using SRS to select Phase II are shown in Table 4.2. The proposed SMLE continued to be unbiased, with accurate SEE and reasonable coverage probabilities. The EM algorithm remained stable, converging in $\geq 96\%$ of replicates. The complete-case and HT estimators are equivalent under SRS. The efficiency gain of the SMLE, which used all available information on all subjects, over the complete-case analyses, which used information on audited subjects only, was higher for smaller values of σ_U^2. This makes sense because X_b^* was more informative about X_b when σ_U^2 was smaller. Thus, more information could be gained by including the Phase I data. In some settings, the complete-case was as much as 41% less efficient than the SMLE. The SMLE was generally, although not always, more efficient than the raking estimator. For a fixed σ_U^2, the relative efficiency (RE) of the SMLE to the complete-case or raking estimators decreased as p_v increased. This also makes sense

Table 4.2: Simulation results for outcome misclassification and a continuous covariate with varied additive measurement error variance when the Phase II design is simple random sampling

			SMLE				Complete-Case/HT			Raking		
σ_U^2	N	p_v	Bias	SE	SEE	CP	Bias	SE	RE	Bias	SE	RE
0.10	1000	0.10	0.005	0.250	0.237	0.943	0.069	0.323	0.599	0.038	0.271	0.849
		0.25	−0.005	0.157	0.160	0.953	0.019	0.183	0.736	0.016	0.166	0.898
		0.50	0.005	0.121	0.121	0.940	0.009	0.132	0.840	0.005	0.116	1.085
	2000	0.10	−0.012	0.166	0.164	0.953	0.031	0.216	0.592	0.020	0.181	0.843
		0.25	−0.011	0.112	0.112	0.949	0.010	0.132	0.721	0.005	0.119	0.887
		0.50	0.002	0.083	0.085	0.956	0.002	0.087	0.905	0.004	0.085	0.949
0.25	0.10	1000	0.003	0.267	0.251	0.950	0.070	0.322	0.688	0.045	0.285	0.878
		0.25	−0.004	0.166	0.166	0.959	0.019	0.183	0.823	0.019	0.173	0.925
		0.50	0.007	0.125	0.123	0.944	0.009	0.132	0.897	0.006	0.120	1.091
	2000	0.10	−0.015	0.179	0.173	0.943	0.031	0.216	0.686	0.025	0.187	0.915
		0.25	−0.011	0.117	0.116	0.944	0.010	0.132	0.784	0.005	0.122	0.918
		0.50	0.004	0.084	0.086	0.956	0.002	0.087	0.936	0.004	0.086	0.958
0.50	1000	0.10	0.030	0.292	0.273	0.948	0.067	0.318	0.843	0.050	0.298	0.959
		0.25	0.001	0.171	0.173	0.957	0.019	0.183	0.873	0.020	0.179	0.910
		0.50	0.008	0.128	0.126	0.941	0.009	0.132	0.940	0.005	0.122	1.103
	2000	0.10	0.001	0.196	0.187	0.941	0.031	0.216	0.824	0.029	0.194	1.021
		0.25	−0.006	0.123	0.121	0.949	0.010	0.132	0.862	0.006	0.127	0.931
		0.50	0.003	0.085	0.088	0.961	0.002	0.087	0.965	0.004	0.089	0.922
1.00	1000	0.10	0.060	0.318	0.292	0.951	0.070	0.322	0.975	0.053	0.310	1.052
		0.25	0.010	0.177	0.180	0.964	0.019	0.183	0.936	0.022	0.183	0.931
		0.50	0.008	0.129	0.128	0.948	0.009	0.132	0.955	0.006	0.124	1.083
	2000	0.10	0.026	0.212	0.201	0.940	0.031	0.216	0.960	0.032	0.202	1.097
		0.25	0.001	0.126	0.125	0.953	0.010	0.132	0.908	0.005	0.129	0.951
		0.50	0.002	0.086	0.089	0.957	0.002	0.087	0.988	0.003	0.091	0.903

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the average of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the relative efficiency of the estimator to the SMLE. Under SRS the HT and complete-case estimators are equivalent. Each entry is based on 1000 replicates. The SMLE had greater than 96% convergence rates in all settings. The Monte Carlo simulation errors for bias and CP did not exceed 0.01 and 0.8%, respectively.

because, as audited data became available on more subjects, less information could be extracted from the unvalidated subjects. The naive analysis was most biased in settings where σ_U^2 was smaller and improved as σ_U^2 increased. Specifically, the naive estimator yielded an average of 65%, 57%, 36%, and 3% bias when $\sigma_U^2 = 0.1$, 0.25, 0.5, and 1, respectively. This counterintuitive phenomenon was due to the way we generated X_b^* and Y^*. In additional simulations (Section 4.6.2.4), we found that the bias of the naive estimator could increase as σ_U^2 increased or reverse direction in various settings.

Simulation results using naive case-control to select Phase II subjects are included in Table 4.3. The SMLE continued to perform well under this sampling scheme, with smaller standard errors than under SRS. The complete-case estimators were

Table 4.3: Simulation results for outcome misclassification and a continuous covariate with varied additive measurement error variance when the Phase II design is 1:1 case-control sampling based on Y^*

			SMLE				Complete-Case		HT			Raking		
σ_U^2	N	p_v	Bias	SE	SEE	CP	Bias	SE	Bias	SE	RE	Bias	SE	RE
0.10	1000	0.10	−0.049	0.234	0.222	0.932	−0.059	0.286	0.046	0.298	0.617	0.043	0.248	0.891
		0.25	−0.028	0.148	0.151	0.952	−0.077	0.160	0.028	0.170	0.758	0.004	0.156	0.898
		0.50	−0.008	0.118	0.115	0.941	−0.091	0.115	0.009	0.124	0.906	0.006	0.118	1.004
	2000	0.10	−0.046	0.159	0.155	0.930	−0.073	0.192	0.031	0.207	0.590	0.024	0.172	0.853
		0.25	−0.026	0.103	0.106	0.945	−0.090	0.111	0.010	0.119	0.749	0.015	0.111	0.863
		0.50	−0.006	0.079	0.080	0.946	−0.096	0.078	0.005	0.084	0.884	0.002	0.081	0.947
0.25	1000	0.10	−0.042	0.233	0.237	0.950	−0.047	0.274	0.054	0.294	0.650	0.040	0.267	0.764
		0.25	−0.021	0.150	0.157	0.958	−0.081	0.161	0.021	0.170	0.853	0.005	0.162	0.856
		0.50	0.000	0.118	0.117	0.949	−0.085	0.115	0.013	0.124	0.890	0.009	0.115	1.057
	2000	0.10	−0.038	0.172	0.165	0.930	−0.069	0.186	0.028	0.200	0.681	0.026	0.176	0.960
		0.25	−0.029	0.109	0.110	0.933	−0.094	0.112	0.002	0.120	0.840	0.011	0.113	0.924
		0.50	−0.004	0.080	0.082	0.948	−0.093	0.079	0.005	0.086	0.909	0.001	0.081	0.968
0.50	1000	0.10	−0.004	0.270	0.256	0.940	−0.038	0.270	0.060	0.292	0.855	0.037	0.272	0.987
		0.25	−0.006	0.160	0.165	0.958	−0.068	0.162	0.027	0.172	0.865	0.013	0.157	1.037
		0.50	0.005	0.122	0.119	0.946	−0.075	0.118	0.019	0.127	0.923	0.012	0.125	0.954
	2000	0.10	−0.019	0.178	0.177	0.938	−0.068	0.183	0.028	0.193	0.851	0.020	0.190	0.880
		0.25	−0.017	0.114	0.114	0.954	−0.084	0.115	0.007	0.123	0.859	0.013	0.117	0.946
		0.50	−0.005	0.080	0.083	0.962	−0.089	0.078	0.004	0.084	0.907	0.010	0.084	0.902
1.00	1000	0.10	0.013	0.288	0.270	0.941	−0.034	0.285	0.051	0.300	0.922	0.044	0.287	1.007
		0.25	0.001	0.172	0.169	0.942	−0.063	0.168	0.021	0.179	0.923	0.012	0.180	0.918
		0.50	0.002	0.118	0.120	0.953	−0.072	0.114	0.013	0.121	0.951	0.008	0.121	0.947
	2000	0.10	0.007	0.193	0.189	0.952	−0.057	0.186	0.028	0.198	0.950	0.019	0.199	0.936
		0.25	−0.012	0.118	0.118	0.948	−0.080	0.117	0.004	0.124	0.906	0.014	0.121	0.945
		0.50	−0.002	0.082	0.084	0.953	−0.078	0.079	0.005	0.084	1.000	0.005	0.082	1.000

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the average of the standard error estimator; CP is the coverage probability of the 95% confidence interval. RE is the relative efficiency of the estimator to the SMLE, but relative efficiency of the complete-case estimator to SMLE is not reported since it was biased under this sampling scheme. Each entry is based on 1000 replicates. The SMLE had greater than 96% convergence rates in all settings. The Monte Carlo simulation errors were ≤ 0.009 for bias and $\leq 0.8\%$ for CP.

5%–10% biased because the case-control sampling was based on an outcome subject to differential misclassification, but the HT and raking estimators remained unbiased. In general, efficiency was slightly better with naive case-control sampling than SRS, although the RE of the SMLE to the other estimators was similar.

4.3.2.2 Varying outcome misclassification rate

We also varied the misclassification rate in Y^* by changing the intercept and regression coefficient of Y, denoted by γ_0 and γ_1, respectively, in its generation model $P(Y^*|X_b^*, Y, X_b, X_a) = [1 + \exp\{-(\gamma_0 + X_b^* + \gamma_1 Y + X_b - 0.5X_a)\}]^{-1}$. The values of γ_0 and γ_1 were determined by the "underlying" sensitivity and specificity of Y^* when it depended on Y only, i.e., $\gamma_0 = -\log\left(\frac{\text{specificity}}{1-\text{specificity}}\right)$ and $\gamma_1 = -\gamma_0 - \log\left(\frac{1-\text{sensitivity}}{\text{sensitivity}}\right)$. The underlying sensitivity of Y^* was varied from 0.95 to 0.55 by decrements of 0.1, and the underlying specificity was set to be 0.05 lower than the underlying sensitivity. A Phase I sample of $N = 1000$ subjects was generated, and a validation subsample

Table 4.4: Simulation results for outcome misclassification with varied baseline sensitivity and specificity and an error-prone continuous covariate

		SMLE				HT			Raking		
Sensitivity	Specificity	Bias	SE	SEE	CP	Bias	SE	RE	Bias	SE	RE
					SRS						
0.95	0.90	−0.007	0.156	0.160	0.953	0.018	0.178	0.768	0.021	0.170	0.842
0.85	0.80	−0.006	0.170	0.175	0.961	0.020	0.180	0.892	0.022	0.176	0.933
0.75	0.70	0.001	0.177	0.182	0.965	0.020	0.183	0.936	0.015	0.189	0.877
0.65	0.60	0.010	0.182	0.185	0.962	0.019	0.184	0.978	0.021	0.193	0.899
0.55	0.50	0.018	0.183	0.186	0.963	0.019	0.183	1.000	0.021	0.190	0.928
					1:1 case-control sampling based on Y^*						
0.95	0.90	−0.028	0.148	0.151	0.952	0.028	0.170	0.758	0.012	0.154	0.924
0.85	0.80	−0.016	0.164	0.168	0.950	0.021	0.176	0.868	0.021	0.172	0.909
0.75	0.70	−0.003	0.178	0.177	0.951	0.019	0.185	0.926	0.009	0.180	0.978
0.65	0.60	0.018	0.182	0.184	0.956	0.028	0.185	0.968	0.025	0.184	0.978
0.55	0.50	0.020	0.188	0.186	0.952	0.021	0.188	1.000	0.026	0.194	0.939

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the average of the standard error estimator; CP is the coverage probability of the 95% confidence interval; RE is the relative efficiency of the estimator to the SMLE. Each entry is based on 1000 replicates. The SMLE had greater than 99% convergence rates in all settings. The Monte Carlo simulation errors for the bias and CP of the SMLE were ≤ 0.006 and $\leq 0.7\%$, respectively.

of $n = 250$ subjects was selected via SRS or naive case-control. The error variance was fixed at $\sigma_U^2 = 0.1$, and all other variables were generated as in Section 4.3.2.1.

The results are shown in Table 4.4. The largest efficiency gains of the SMLE over the HT estimator under SRS (equivalent to the complete-case analysis) and naive case-control were seen when the sensitivity and specificity of Y^* were highest. In fact, the RE decreased with these diagnostic measures until it was approximately equal to one at 0.55 sensitivity and 0.5 specificity. This was expected because for there to be an efficiency gain of the SMLE from incorporating information in unvalidated subjects, there needs to be a fair degree of correlation between Y and Y^*. Sensitivity and specificity of Y^* near 0.5 resulted in near random misclassification, in which case the unvalidated subjects were not very informative about the relationship between Y and X_b. The SMLE was always more efficient than the raking estimator, which also incorporates information in unvalidated subjects.

4.3.2.3 Other simulations with an error-prone continuous covariate

In Section 4.6, we include comparisons of the SMLE to regression calibration (RC) (Prentice, 1982) and generalized raking under the classical measurement error setting with covariate error only (Section 4.6.2.3). The robustness of the SMLE to different covariate error mechanisms, including non-zero mean additive errors and

multiplicative errors, was illustrated in Sections 4.6.2.5 and 4.6.2.6, respectively. In these simulations, the errors in X_b^* depended on the error-free covariate X_a. The SMLE continued to perform well in these settings.

4.4 Application to the CCASAnet Dataset

We now apply our method to the CCASAnet dataset. As in Giganti et al. (2019), the risk of developing an ADE after initiating ART was of primary interest. Specifically, we were interested in estimating the relative odds of developing an ADE within two years of ART initiation for two risk factors, CD4 count and prior AIDS diagnosis, conditional on other covariates. Both risk factors were measured at ART initiation. Specifically, CD4 count was defined as the lab measurement closest to the ART initiation date but no more than six months prior to or thirty days after ART initiation. Prior AIDS diagnosis was any evidence of a clinical AIDS event before ART initiation. Because variables were derived based on error-prone ART initiation date, errors could be correlated. Other error-free covariates included clinical site, age at baseline, sex, and year of ART initiation. CD4 count was rescaled to units of ten cells per microliter before being square root transformed, age was rescaled to ten-year increments, and year of ART initiation was centered at the median, 2004.

Clinical data from five sites (anonymously labeled as sites A–E) were compiled into the CCASAnet research database. Each site underwent an on-site audit by VDCC investigators between 2013–2014. Approximately 30 patient records were randomly selected from each site for auditing. Pre-audit records were compared with clinical source documents, including paper-based patient charts or electronic medical records; see Giganti et al. (2019) for details about the audit protocol and findings. The values found in clinical source documents were treated as the reference standard and are assumed to be more correct than the database.

To be included in our analysis, patients needed to (1) initiate ART while in care at a CCASAnet clinic, (2) be at least 18 years old at cohort enrollment, (3) have a valid CD4 measurement at time of ART initiation, and (4) remain in care for at least two years after initiating ART. Based on the unvalidated data, these inclusion criteria resulted in a Phase I sample of 5109 subjects from the CCASAnet research database, of whom 117 were audited. The number of audited records meeting these criteria varied between 16–36 per site. There were 510 unvalidated ADE (10% prevalence) and 13 validated ADE (11% prevalence). Giganti et al. (2019) noted that risk of an ADE was higher in the audited data than in the pre-audit data over a ten-year

follow-up period.

In these audits, the VDCC identified 6% misclassification in the ADE, all of which were false negatives. AIDS prior to ART initiation had 6% misclassification, with a higher FPR (13%) than FNR (3%). CD4 count had an error rate of 8%, with mean magnitude of -0.11 and variance of 2.51 on the square root scale. Errors in CD4 count were assumed to be additive on the square root scale, so magnitude was calculated by subtracting error-prone from validated values. No subject had errors in both their outcome and covariates and only one had errors in both CD4 count and AIDS status, suggesting little evidence of error correlation. Sites A, B, and C had five or six erroneous records while sites D and E had two or three. The low error rates and small audit size led us to choose the histogram basis for the SMLE. Specifically, we used separate histogram bases with one interior knot for subjects with and without unvalidated AIDS at ART initiation. Thus, errors in AIDS and CD4 count were assumed to be independent of other error-free covariates. Further stratification by site did not noticeably alter the results (see Table 4.13).

Results are presented in Table 4.5. The naive analysis using only Phase I data indicated that both CD4 count (log OR $= -0.28$; 95% CI: $(-0.34, -0.22)$) and prior AIDS (log OR $= 1.54$; 95% CI: $(1.32, 1.77)$) were strongly associated with ADE during the first two years after initiating ART. The complete-case and HT analyses, which only used Phase II data, yielded larger point estimates for the CD4 count association but point estimates closer to the null for the prior AIDS association; confidence intervals for the complete-case and HT analyses were quite wide due to the small audit size and included zero for the prior AIDS association. The CI for AIDS in the raking analysis was narrower than those in the HT and complete-case analyses, but still contained zero. Using the SMLE, estimates for CD4 count (log OR $= -0.48$; 95% CI: $(-0.73, -0.24)$) and AIDS (log OR $= 1.39$; 95% CI: $(0.58, 2.19)$) were significant and fell between those of the naive and complete-data-based analyses, capturing the information from the validated data while harnessing the statistical power of the full cohort.

Our analyses excluded 283 unaudited and 5 audited subjects who died within two years of initiating ART and thus did not meet inclusion criterion (4). Analyses were repeated including these patients and using the composite endpoint of death or ADE; results were largely similar (Section 4.6.3.1).

Table 4.5: log OR estimates and 95% confidence intervals from the analysis of the CCASAnet dataset

	Naive		Complete-Case		HT		Raking		SMLE	
Covariate	log OR	95% CI	log OR	95% CI	log OR	95% CI	log OR	95% CI	log OR	95% CI
$\sqrt{\text{CD4}/10}$	−0.280	(−0.343, −0.217)	−0.688	(−1.164, −0.212)	−0.755	(−1.154, −0.356)	−0.620	(−0.922, −0.318)	−0.482	(−0.725, −0.240)
AIDS	1.543	(1.317, 1.770)	0.243	(−1.166, 1.653)	0.579	(−0.850, 2.009)	0.093	(−1.131, 1.318)	1.388	(0.582, 2.194)
Site: A	−1.399	(−1.755, −1.042)	−0.396	(−2.433, 1.642)	−0.357	(−2.601, 1.887)	−0.289	(−1.945, 1.368)	1.129	(0.278, 1.980)
Site: C	0.409	(0.154, 0.664)	0.561	(−1.368, 2.491)	0.658	(−1.447, 2.764)	0.543	(−1.099, 2.184)	0.184	(0.003, 0.365)
Site: D	−0.991	(−1.412, −0.570)	−2.416	(−5.027, 0.194)	−2.638	(−5.015, −0.261)	−2.548	(−5.226, 0.131)	−1.225	(−2.394, −0.056)
Site: E	−0.225	(−0.581, 0.131)	−0.688	(−3.353, 1.976)	−0.686	(−3.615, 2.244)	−0.542	(−2.855, 1.772)	−0.732	(−1.725, 0.260)
Male	0.073	(−0.169, 0.316)	−0.728	(−2.330, 0.874)	−0.823	(−2.395, 0.749)	−1.195	(−2.669, 0.280)	−0.703	(−1.933, 0.527)
Age/10 years	0.014	(−0.091, 0.119)	0.354	(−0.296, 1.003)	0.310	(−0.315, 0.935)	0.223	(−0.311, 0.756)	−0.690	(−1.644, 0.263)
Year of ART	−0.023	(−0.051, 0.006)	0.092	(−0.144, 0.327)	0.155	(−0.206, 0.516)	0.081	(−0.134, 0.297)	−0.508	(−1.225, 0.210)

Note: 95% CI is the 95% confidence interval.

4.5 Discussion

Measurement error is a wide-reaching problem in biomedical research. As error-prone observational data are increasingly supporting decision-making in health policy and patient care, there is a demonstrated need for statistical methods that can retain the high power lent by large cohorts while accounting for data errors. We proposed a new SMLE method that can address dependent errors in binary outcomes and categorical or continuous covariates, and we illustrated its performance in our simulations and CCASAnet data application. The SMLE is robust, efficient, and can handle measurement error settings not yet addressed by the MLE of Tang et al. (2015). We note that other methods, including multiple imputation approaches proposed by Edwards et al. (2013) and Giganti et al. (2020), could be adapted to handle the same problem. Because these approaches rely on proper specification of the error-generating mechanisms, we expect them to perform similarly to the MLE.

The SMLE has limitations. First, it can only accommodate two or three continuous covariates in the B-spline basis because the dimension of the basis grows exponentially fast as the number of continuous covariates increases. This is a manifestation of the curse of dimensionality. There are workarounds: 1) error-free covariates that can be assumed to be independent of the error-prone covariates can be omitted or 2) dimension reduction techniques can summarize the covariates into a few representative features on which the basis is constructed. Second, although the logistic regression model $P_{\boldsymbol{\gamma}}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})$ seems to be fairly robust (Section 4.6.2.2), proper specification is still desirable. One may include additional covariates that affect Y^* but not Y and additional interaction terms or splines to facilitate flexible modeling of the outcome error mechanism. Third, $\boldsymbol{X}^*$ is assumed to be a surrogate for $\boldsymbol{X}$ such that $P(Y|\boldsymbol{X}, \boldsymbol{X}^*) = P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})$. Relaxing this assumption is straightforward, but changes the marginal interpretation of the estimates.

The proposed SMLE allows the Phase II sample selection to depend on the Phase I data in any manner. An interesting topic worth further investigation is efficient

design under outcome misclassification and covariate measurement error. Because the outcome is subject to misclassification, traditional case-control sampling may not be ideal. Multi-wave designs like those proposed by McIsaac and Cook (2015) and Chen and Lumley (2020) are promising because one can use validated data obtained from earlier waves to gain insights about error mechanisms and then use this knowledge to optimally allocate audit efforts in later waves.

4.6 Appendix C

4.6.1 Asymptotic Properties of the SMLE

Our asymptotic theory extends that in Tao et al. (2017) to allow for the outcome Y to be a Phase II variable. First, the joint cumulative distribution function of $\boldsymbol{X}$ and $\boldsymbol{X}^*$, $F(\boldsymbol{X}, \boldsymbol{X}^*)$, can be estimated by

$$\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*) = N^{-1} \sum_{k=1}^{m} \sum_{i=1}^{N} I(\boldsymbol{x}_k \leq \boldsymbol{x}, \boldsymbol{X}_i^* \leq \boldsymbol{x}^*) \sum_{j=1}^{s_N} B_j^q(\boldsymbol{X}_i^*) p_{kj}. \tag{4.11}$$

Let $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}$ denote the parameter spaces of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, respectively, which are bounded open sets in the domains of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, respectively. Let $\mathcal{F}$ denote the space of joint distributions of $(\boldsymbol{X}, \boldsymbol{X}^*)$. Let $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ denote the true value of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}_0 \in \boldsymbol{\Gamma}$ denote the true value of $\boldsymbol{\gamma}$, and $F_0 \in \mathcal{F}$ denote the true value of $F(\boldsymbol{X}, \boldsymbol{X}^*)$. We impose the following regularity conditions:

(C1) The set of covariates $(\boldsymbol{X}, \boldsymbol{X}^*)$ has bounded support.

(C2) If there exist two sets of parameters $(\boldsymbol{\theta}_1, \boldsymbol{\gamma}_1, F_1)$ and $(\boldsymbol{\theta}_2, \boldsymbol{\gamma}_2, F_2)$ such that

$$P_{\boldsymbol{\theta}_1}(Y|\boldsymbol{X}) P_{\boldsymbol{\gamma}_1}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X}) F_1(\boldsymbol{X}, \boldsymbol{X}^*) = P_{\boldsymbol{\theta}_2}(Y|\boldsymbol{X}) P_{\boldsymbol{\gamma}_2}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X}) F_2(\boldsymbol{X}, \boldsymbol{X}^*),$$

where $(Y, \boldsymbol{X}, Y^*, \boldsymbol{X}^*) \in \mathcal{C} \equiv \{(y, \boldsymbol{x}, y^*, \boldsymbol{x}^*) : P(V = 1|y^*, \boldsymbol{x}^*) > 0\}$, then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$, and $F_1 = F_2$. Further, if there exist a vector of constants $\boldsymbol{c}$ such that

$$\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{P_{\boldsymbol{\theta}_0}(y_1|\boldsymbol{x})}{P_{\boldsymbol{\theta}_0}(y_2|\boldsymbol{x})} + \frac{\partial}{\partial \boldsymbol{\gamma}} \log \frac{P_{\boldsymbol{\gamma}_0}(y_1^*|\boldsymbol{x}^*, y_1, \boldsymbol{x})}{P_{\boldsymbol{\gamma}_0}(y_2^*|\boldsymbol{x}^*, y_2, \boldsymbol{x})} \right\}^T \boldsymbol{c} = 0$$

for any $(y_i^*, \boldsymbol{x}^*, y_i, \boldsymbol{x}) \in \mathcal{C}$, $i = 1, 2$, then $\boldsymbol{c} = \boldsymbol{0}$.

(C3) The distribution function F_0 is positive in its support and q-times differentiable with respect to a suitable measure.

(C4) As $N \to \infty$, $s_N \to \infty$, and $N^{1/2} s_N^{-q/d} \to 0$.

(C5) The function $E(V|\boldsymbol{X}, \boldsymbol{X}^*)$ is q-times continuously differentiable with respect to $\boldsymbol{X}$ and $\boldsymbol{X}^*$.

With conditions (C1)–(C5), we state and prove the following asymptotic results.

Theorem 4.2.1. Under conditions (C1)−(C5),

$$||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|| + ||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| + \sup_{\boldsymbol{x}, \boldsymbol{x}^*} |\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*) - F_0(\boldsymbol{x}, \boldsymbol{x}^*)| \to 0$$

almost surely.

Proof. Since estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ are bounded and $\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*)$ is a distribution function with bounded support, it follows from Helly's Selection Theorem that, for any subsequence of $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\gamma}}$, and $\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*)$, there exists a further subsequence (still denoted $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\gamma}}$, and $\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*)$) such that $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ converge to some vectors $\breve{\boldsymbol{\theta}}$ and $\breve{\boldsymbol{\gamma}}$, respectively, and $\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*)$ converges weakly to some function $\breve{F}(\boldsymbol{x}, \boldsymbol{x}^*)$. For Theorem 4.2.1 to hold, we need to show that $\breve{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, $\breve{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$, and $\breve{F} = F_0$.

Recall that the B-spline coefficients $\hat{p}_{kj}$ are defined to maximize the observed-data log-likelihood $l_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \{p_{kj}\})$ in expression (4.4). Differentiating $l_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \{p_{kj}\})$ with respect to p_{kj}, we have

$$\begin{aligned}\hat{\mu}_j =& \sum_{i=1}^{N} V_i \frac{I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)}{p_{kj}} \\ &+ \sum_{i=1}^{N} (1 - V_i) \frac{\sum_{y=0}^{1} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_k) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)}{\sum_{y=0}^{1} \sum_{k'=1}^{m} \sum_{j'=1}^{s_N} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) B_{j'}^q(\boldsymbol{X}_i^*) p_{k'j'}}, \end{aligned} \quad (4.12)$$

where $\hat{\mu}_j$ is the Lagrange multiplier for the constraint $\sum_{k=1}^{m} p_{kj} = 1$. By multiplying both sides of equation (4.12) by p_{kj} and summing over k ($k = 1, \ldots, m$), we can show that

$$\begin{aligned}\hat{\mu}_j =& \sum_{i=1}^{N} V_i B_j^q(\boldsymbol{X}_i^*) \\ &+ \sum_{i=1}^{N} (1 - V_i) \frac{\sum_{y=0}^{1} \sum_{k'=1}^{m} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) B_j^q(\boldsymbol{X}_i^*) p_{k'j}}{\sum_{y=0}^{1} \sum_{k'=1}^{m} \sum_{j'=1}^{s_N} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) B_{j'}^q(\boldsymbol{X}_i^*) p_{k'j'}}. \end{aligned} \quad (4.13)$$

From equation (4.12), it follows that $\hat{p}_{kj}$ can be expressed as

$$\hat{p}_{kj} = \frac{\sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)}{\hat{\mu}_j - \sum_{i=1}^{N} (1 - V_i) \frac{\sum_{y=0}^{1} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_k) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)}{\sum_{y=0}^{1} \sum_{k'=1}^{m} \sum_{j'=1}^{s_N} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) B_{j'}^q(\boldsymbol{X}_i^*) \hat{p}_{k'j'}}}. \quad (4.14)$$

Now if we plug $\hat{\mu}_j$ from equation (4.13) into the form for $\hat{p}_{kj}$ from equation (4.14) we have the alternate form

$$\hat{p}_{kj} = \frac{\sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)}{\sum_{i=1}^{N} \{V_i + (1 - V_i)(a_{ij} - b_{ikj})\} B_j^q(\boldsymbol{X}_i^*)}, \tag{4.15}$$

where

$$\begin{aligned} a_{ij} =& \frac{\sum_{y=0}^{1} \sum_{k'=1}^{m} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) \hat{p}_{k'j}}{\sum_{y=0}^{1} \sum_{k'=1}^{m} \sum_{j'=1}^{s_N} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) B_{j'}^q(\boldsymbol{X}_i^*) \hat{p}_{k'j'}}, \\ b_{ikj} =& \frac{\sum_{y=0}^{1} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_k) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k)}{\sum_{y=0}^{1} \sum_{k'=1}^{m} \sum_{j'=1}^{s_N} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_{k'}) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_{k'}) B_{j'}^q(\boldsymbol{X}_i^*) \hat{p}_{k'j'}}. \end{aligned}$$

Recall that the B-spline basis is constructed such that $\hat{P}(\boldsymbol{x} = \boldsymbol{x}_k|\boldsymbol{x}^*) = \sum_{j=1}^{s_N} B_j^q(\boldsymbol{x}^*)\hat{p}_{kj}$. With the form for $\hat{p}_{kj}$ from equation (4.15), we have

$$\hat{P}(\boldsymbol{x} = \boldsymbol{x}_k|\boldsymbol{x}^*) = \sum_{j=1}^{s_N} B_j^q(\boldsymbol{x}^*) \left[\frac{\sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)}{\sum_{i=1}^{N} \{V_i + (1 - V_i)(a_{ij} - b_{ikj})\} B_j^q(\boldsymbol{X}_i^*)} \right]. \tag{4.16}$$

Because B-spline basis functions have local support, there is only a narrow region where $B_j^q(\boldsymbol{x}^*)$ takes on a value other than 0. This property can be expressed in the following inequality:

$$|B_j^q(\tilde{\boldsymbol{x}}^*) - B_j^q(\boldsymbol{x}^*) I(||\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*|| \leq \xi_N)| \lesssim \xi_N \tag{4.17}$$

where $\lesssim$ means less than or equal to up to a constant and $\xi_N = (1 + b_N)^{-1}$. By condition (C4), we have $s_N \to \infty$ as $N \to \infty$. Since $s_N = (b_N + q)^d$ and parameters q and d are fixed, it must be that $b_N \to \infty$ as $N \to \infty$. Now, from inequality (4.17) it follows that asymptotically

$$B_j^q(\tilde{\boldsymbol{x}}^*) \approx B_j^q(\boldsymbol{x}^*) I(||\tilde{\boldsymbol{x}}^* - \boldsymbol{x}^*|| \leq \xi_N) \tag{4.18}$$

for a general pair of values $\tilde{\boldsymbol{x}}^*$ and $\boldsymbol{x}^*$ in the domain of $\boldsymbol{X}^*$. From expression (4.18) and equation (4.16), we have that $\hat{P}(\boldsymbol{X} = \boldsymbol{x}_k|\boldsymbol{x}^*)$ is asymptotically equivalent to

$$\frac{\sum_{j=1}^{s_N} \sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{x}^*) I(||\boldsymbol{X}_i^* - \boldsymbol{x}^*|| \leq \xi_N)}{g_{1N}(\boldsymbol{x}_k, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})} \tag{4.19}$$

where

$$g_{1N}(\boldsymbol{x}_k, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F}) = \sum_{j=1}^{s_N} \sum_{i=1}^{N} \{V_i + (1 - V_i)(a_{ij} - b_{ikj})\} B_j^q(\boldsymbol{x}^*) I(||\boldsymbol{X}_i^* - \boldsymbol{x}^*|| \leq \xi_N). \tag{4.20}$$

With (4.19) and the definition in equation (4.11), we have that $\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*)$ is asymptotically equivalent to

$$N^{-1} \sum_{k=1}^{m} \sum_{i=1}^{N} I(\boldsymbol{x}_k \leq \boldsymbol{x}, \boldsymbol{X}_i^* \leq \boldsymbol{x}^*) \frac{\sum_{j=1}^{s_N} \sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{x}^*) I(||\boldsymbol{X}_i^* - \boldsymbol{x}^*|| \leq \xi_N)}{g_{1N}(\boldsymbol{x}_k, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})}.$$

Next, we show that $(Ns_N)^{-1} g_{1N}(\boldsymbol{x}, \boldsymbol{x}^*, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})$ is bounded away from zero for sufficiently large N. By the approximation theory of B-splines (Schumaker, 1981) and Glivenko-Cantelli theorem,

$$\begin{aligned} & N^{-1} \sum_{y=0}^{1} \sum_{k=1}^{m} \sum_{j=1}^{s_N} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_k) P_{\hat{\boldsymbol{\gamma}}}(y^*|\boldsymbol{x}^*, y, \boldsymbol{x}_k) B_j^q(\boldsymbol{x}^*) \hat{p}_{kj} \\ & = \sum_{y=0}^{1} \int_{\tilde{\boldsymbol{x}}} P_{\hat{\boldsymbol{\theta}}}(y|\tilde{\boldsymbol{x}}) P_{\hat{\boldsymbol{\gamma}}}(y^*|\boldsymbol{x}^*, y, \tilde{\boldsymbol{x}}) \hat{F}(d\tilde{\boldsymbol{x}}, \boldsymbol{x}^*) \to \sum_{y=0}^{1} \int_{\tilde{\boldsymbol{x}}} P_{\breve{\boldsymbol{\theta}}}(y|\tilde{\boldsymbol{x}}) P_{\breve{\boldsymbol{\gamma}}}(y^*|\boldsymbol{x}^*, y, \tilde{\boldsymbol{x}}) \breve{F}(d\tilde{\boldsymbol{x}}, \boldsymbol{x}^*) \end{aligned} \tag{4.21}$$

uniformly in $(\boldsymbol{x}^*, y^*)$. With the asymptotic results from expression (4.21) and equation (4.20), it follows that $(Ns_N)^{-1} g_{1N}(\boldsymbol{x}, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})$ converges to $g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})$ for $(\boldsymbol{x}, \boldsymbol{x}^*)$ in the support of $(\boldsymbol{X}, \boldsymbol{X}^*)$, where $g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})$ is defined as

$$E\left[\left\{1 - (1 - V)\frac{\sum_{y=0}^{1} P_{\breve{\boldsymbol{\theta}}}(y|\boldsymbol{x}) P_{\breve{\boldsymbol{\gamma}}}(y^*|\boldsymbol{x}^*, y, \boldsymbol{x}) \int_{\tilde{\boldsymbol{x}}} \breve{F}(d\tilde{\boldsymbol{x}}, \boldsymbol{x}^*)}{\sum_{y=0}^{1} \int_{\tilde{\boldsymbol{x}}} P_{\breve{\boldsymbol{\theta}}}(y|\tilde{\boldsymbol{x}}) P_{\breve{\boldsymbol{\gamma}}}(y^*|\boldsymbol{x}^*, y, \tilde{\boldsymbol{x}}) \breve{F}(d\tilde{\boldsymbol{x}}, \boldsymbol{x}^*)}\right\} f_{\boldsymbol{x}^*}(\boldsymbol{X}^*) \middle| \boldsymbol{X}^* = \boldsymbol{x}^*\right] \geq 0, \tag{4.22}$$

and $f_{\boldsymbol{x}^*}(\boldsymbol{X}^*)$ is the density function of $\boldsymbol{X}^*$. Recall from equation (4.19) that

$$\begin{aligned} 1 &= \sum_{k=1}^{m} \hat{P}(\boldsymbol{X} = \boldsymbol{x}_k | \boldsymbol{x}^*) \\ &= \sum_{k=1}^{m} \frac{\sum_{j=1}^{s_N} \sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{x}^*) I(||\boldsymbol{X}_i^* - \boldsymbol{x}^*|| \leq \xi_N)}{g_{1N}(\boldsymbol{x}_k, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})}. \end{aligned}$$

By equation (4.22) and the approximation theory of B-splines, we have

$$\sum_{k=1}^{m} \frac{\sum_{j=1}^{s_N}\sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{x}^*) I(||\boldsymbol{X}_i^* - \boldsymbol{x}^*|| \le \xi_N)}{g_{1N}(\boldsymbol{x}_k, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})} \to \int_{\boldsymbol{x}} \frac{E\{V f_{\boldsymbol{x}^*}(\boldsymbol{x}^*)|\boldsymbol{X}^* = \boldsymbol{x}^*\}}{g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})} d\boldsymbol{x}.$$

Now, if $g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})$ is not bounded away from zero, then there must exist $\boldsymbol{x}_0$ in the support of $\boldsymbol{X}$ such that $g_1(\boldsymbol{x}_0, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F}) = 0$. Because $g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})$ is a smooth function on the continuous components of $\boldsymbol{x}$, there exists a positive constant δ such that for any $\epsilon > 0$,

$$\begin{aligned} 1 &\ge \int_{\boldsymbol{x}} \frac{E\{V f_{\boldsymbol{x}^*}(\boldsymbol{X}^*)|\boldsymbol{X}^* = \boldsymbol{x}^*\}}{g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F}) + \epsilon} d\boldsymbol{x} \ge \int_{||\boldsymbol{x}-\boldsymbol{x}_0||\le\delta} \frac{E\{V f_{\boldsymbol{x}^*}(\boldsymbol{X}^*)|\boldsymbol{X}^* = \boldsymbol{x}^*\}}{|g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})| + \epsilon} d\boldsymbol{x} \\ &\gtrsim \int_{||\boldsymbol{x}-\boldsymbol{x}_0||\le\delta} \frac{E\{V f_{\boldsymbol{x}^*}(\boldsymbol{X}^*)|\boldsymbol{X}^* = \boldsymbol{x}^*\}}{||\boldsymbol{x} - \boldsymbol{x}_0|| + \epsilon} d\boldsymbol{x}, \end{aligned} \tag{4.23}$$

where $\gtrsim$ means greater than or equal to up to a constant. Because $\int_{||\boldsymbol{x}-\boldsymbol{x}_0||\le\delta}(1/||\boldsymbol{x} - \boldsymbol{x}_0||)d\boldsymbol{x}$ is infinite, the last integration in expression (4.23) also goes to ∞ when $\epsilon \to 0$. Thus, we have the following contradiction:

$$1 \gtrsim \int_{||\boldsymbol{x}-\boldsymbol{x}_0||\le\delta} \frac{E\{V f_{\boldsymbol{x}^*}(\boldsymbol{X}^*)|\boldsymbol{X}^* = \boldsymbol{x}^*\}}{||\boldsymbol{x} - \boldsymbol{x}_0|| + \epsilon} d\boldsymbol{x} \to \infty,$$

and it must be true that $g_1(\boldsymbol{x}, \boldsymbol{x}^*; \breve{\boldsymbol{\theta}}, \breve{\boldsymbol{\gamma}}, \breve{F})$ is bounded away from zero for $(\boldsymbol{x}, \boldsymbol{x}^*)$ in the support of $(\boldsymbol{X}, \boldsymbol{X}^*)$. Further, the same is true for $(Ns_N)^{-1} g_{1N}(\boldsymbol{x}, \boldsymbol{x}^*; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F})$ when N is sufficiently large.

Finally, through the Kullback-Leibler inequality we prove that $\breve{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, $\breve{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$, and $\breve{F} = F_0$. Specifically, let

$$\tilde{p}_{kj} = \frac{\sum_{i=1}^{N} V_i I(\boldsymbol{X}_i = \boldsymbol{x}_k) B_j^q(\boldsymbol{X}_i^*)/P(V_i = 1|\boldsymbol{X}_i^*, Y_i^*)}{\sum_{i=1}^{N} V_i B_j^q(\boldsymbol{X}_i^*)/P(V_i = 1|\boldsymbol{X}_i^*, Y_i^*)}$$

and

$$\tilde{F}(\boldsymbol{x}, \boldsymbol{x}^*) = N^{-1} \sum_{k=1}^{m}\sum_{i=1}^{N} I(\boldsymbol{x}_k \le \boldsymbol{x}, \boldsymbol{X}_i^* \le \boldsymbol{x}^*) \sum_{j=1}^{s_N} B_j^q(\boldsymbol{X}_i^*)\tilde{p}_{kj}. \tag{4.24}$$

By the approximation theory of B-splines (Schumaker, 1981), $\tilde{F}(\boldsymbol{x}, \boldsymbol{x}^*) \to F_0(\boldsymbol{x}, \boldsymbol{x}^*)$ uniformly. From the definitions of $\hat{F}$ and $\tilde{F}$ in equations (4.11) and (4.24), respectively, it follows that $\hat{F}$ is absolutely continuous with respect to $\tilde{F}$, and thus $d\hat{F}/d\tilde{F}$ converges uniformly to $d\breve{F}/dF_0$. By condition (C3), $\breve{F}$ is continuously differentiable

with respect to $\boldsymbol{x}$ and $\boldsymbol{x}^*$.

By definition, the SMLE $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\gamma}}$, and $\{\hat{p}_{kj}\}$ maximize the observed-data log-likelihood (expression (4.4)), i.e. $l_N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \{\hat{p}_{kj}\}) = \sup_{\boldsymbol{\theta}, \boldsymbol{\gamma}, \{p_{kj}\}} l_N(\boldsymbol{\theta}, \boldsymbol{\gamma}, \{p_{kj}\})$. Thus, the following is true

$$l_N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \{\hat{p}_{kj}\}) \geq l_N(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, \{\tilde{p}_{kj}\})$$

That is,

$$\begin{aligned}
0 \geq & N^{-1} l_N(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, \{\tilde{p}_{kj}\}) - N^{-1} l_N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \{\hat{p}_{kj}\}) \\
= & - N^{-1} \sum_{i=1}^{N} V_i \left\{ \log \frac{P_{\hat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{X}_i)}{P_{\boldsymbol{\theta}_0}(Y_i|\boldsymbol{X}_i)} + \log \frac{P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i)}{P_{\boldsymbol{\gamma}_0}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i)} \right\} \\
& - N^{-1} \sum_{i=1}^{N} V_i \sum_{k=1}^{m} \sum_{j=1}^{s_N} I(\boldsymbol{X}_i = \boldsymbol{x}_k) \log \frac{\hat{p}_{kj}}{\tilde{p}_{kj}} B_j^q(\boldsymbol{X}_i^*) \\
& - N^{-1} \sum_{i=1}^{N} (1 - V_i) \log \left\{ \frac{\sum_{y=0}^{1} \sum_{k=1}^{m} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x}_k) P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k) \sum_{j=1}^{s_N} \hat{p}_{kj} B_j^q(\boldsymbol{X}_i^*)}{\sum_{y=0}^{1} \sum_{k=1}^{m} P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x}_k) P_{\boldsymbol{\gamma}_0}(Y_i^*|\boldsymbol{X}_i^*, y, \boldsymbol{x}_k) \sum_{j=1}^{s_N} \tilde{p}_{kj} B_j^q(\boldsymbol{X}_i^*)} \right\}.
\end{aligned} \tag{4.25}$$

The first and second terms in expression (4.25) converge as follows:

$$-N^{-1} \sum_{i=1}^{N} V_i \log \frac{P_{\hat{\boldsymbol{\theta}}}(Y_i|\boldsymbol{X}_i)}{P_{\boldsymbol{\theta}_0}(Y_i|\boldsymbol{X}_i)} \to -E \left\{ V \log \frac{P_{\hat{\boldsymbol{\theta}}}(Y|\boldsymbol{X})}{P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X})} \right\}, \tag{4.26}$$

and

$$-N^{-1} \sum_{i=1}^{N} V_i \log \frac{P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i)}{P_{\boldsymbol{\gamma}_0}(Y_i^*|\boldsymbol{X}_i^*, Y_i, \boldsymbol{X}_i)} \to -E \left\{ V \log \frac{P_{\hat{\boldsymbol{\gamma}}}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})}{P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X})} \right\}. \tag{4.27}$$

In the third term, the approximation $\sum_{j=1}^{s_N} \log \frac{\hat{p}_{kj}}{\tilde{p}_{kj}} B_j^q(\boldsymbol{X}_i^*)$ is asymptotically equivalent to

$$\log \frac{\sum_{j=1}^{s_N} \hat{p}_{kj} B_j^q(\boldsymbol{X}_i^*)}{\sum_{j=1}^{s_N} \tilde{p}_{kj} B_j^q(\boldsymbol{X}_i^*)} = \log \left. \frac{d\hat{F}(\boldsymbol{x}, \boldsymbol{x}^*)}{d\tilde{F}(\boldsymbol{x}, \boldsymbol{x}^*)} \right|_{\boldsymbol{x} = \boldsymbol{x}_k},$$

following from the approximation theory of B-splines (Schumaker, 1981). Given this, it follows that

$$\sum_{j=1}^{s_N} \log \frac{\hat{p}_{kj}}{\tilde{p}_{kj}} B_j^q(\boldsymbol{X}_i^*) \to \log \left. \frac{d\breve{F}(\boldsymbol{x}, \boldsymbol{x}^*)}{dF_0(\boldsymbol{x}, \boldsymbol{x}^*)} \right|_{\boldsymbol{x} = \boldsymbol{x}_k}$$

uniformly. Thus, we have

$$-N^{-1}\sum_{i=1}^{N} V_i \sum_{k=1}^{m}\sum_{j=1}^{s_N} I(\boldsymbol{X}_i = \boldsymbol{x}_k)\log\frac{\hat{p}_{kj}}{\tilde{p}_{kj}} B_j^q(\boldsymbol{X}_i^*) \to -E\left\{V\log\frac{d\breve{F}(\boldsymbol{X},\boldsymbol{X}^*)}{dF_0(\boldsymbol{X},\boldsymbol{X}^*)}\right\}. \quad (4.28)$$

By definition, the last term in (4.25) is equivalent to

$$-N^{-1}\sum_{i=1}^{N}(1-V_i)\log\frac{\sum_{y=0}^{1}\int_{\boldsymbol{x}} P_{\hat{\boldsymbol{\theta}}}(y|\boldsymbol{x})P_{\hat{\boldsymbol{\gamma}}}(Y_i^*|\boldsymbol{X}_i^*,y,\boldsymbol{x})\hat{F}(d\boldsymbol{x},\boldsymbol{X}_i^*)}{\sum_{y=0}^{1}\int_{\boldsymbol{x}} P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x})P_{\boldsymbol{\gamma}_0}(Y_i^*|\boldsymbol{X}_i^*,y,\boldsymbol{x})\tilde{F}(d\boldsymbol{x},\boldsymbol{X}_i^*)},$$

which converges to

$$\to -E\left\{(1-V)\log\frac{\sum_{y=0}^{1}\int_{\boldsymbol{x}} P_{\breve{\boldsymbol{\theta}}}(y|\boldsymbol{x})P_{\breve{\boldsymbol{\gamma}}}(Y^*|\boldsymbol{X}^*,y,\boldsymbol{x})\breve{F}(d\boldsymbol{x},\boldsymbol{X}^*)}{\sum_{y=0}^{1}\int_{\boldsymbol{x}} P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x})P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*,y,\boldsymbol{x})F_0(d\boldsymbol{x},\boldsymbol{X}^*)}\right\}. \quad (4.29)$$

Substituting expressions (4.26)−(4.29) into inequality (4.25), we have

$$\begin{aligned} 0 \geq & -E\left[V\left\{\log\frac{P_{\breve{\boldsymbol{\theta}}}(Y|\boldsymbol{X})P_{\breve{\boldsymbol{\gamma}}}(Y^*|\boldsymbol{X}^*,Y,\boldsymbol{X})d\breve{F}(\boldsymbol{X},\boldsymbol{X}^*)}{P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X})P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*,Y,\boldsymbol{X})dF_0(\boldsymbol{X},\boldsymbol{X}^*)}\right\}\right] \\ & -E\left\{(1-V)\log\frac{\sum_{y=0}^{1}\int_{\boldsymbol{x}} P_{\breve{\boldsymbol{\theta}}}(y|\boldsymbol{x})P_{\breve{\boldsymbol{\gamma}}}(Y^*|\boldsymbol{X}^*,y,\boldsymbol{x})\breve{F}(d\boldsymbol{x},\boldsymbol{X}^*)}{\sum_{y=0}^{1}\int_{\boldsymbol{x}} P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x})P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*,y,\boldsymbol{x})F_0(d\boldsymbol{x},\boldsymbol{X}^*)}\right\} \end{aligned} \quad (4.30)$$

Based on expression (4.30), we conclude that the Kullback-Leibler information of the density indexed by $\breve{\boldsymbol{\theta}}$, $\breve{\boldsymbol{\gamma}}$, and $\breve{F}$ with respect to the true density indexed by $\boldsymbol{\theta}_0$, $\boldsymbol{\gamma}_0$, and F_0 is non-positive and thus must be zero. Therefore, the two densities are identical almost surely. For $V = 1$, this implies that

$$P_{\breve{\boldsymbol{\theta}}}(Y|\boldsymbol{X})P_{\breve{\boldsymbol{\gamma}}}(Y^*|\boldsymbol{X}^*,Y,\boldsymbol{X})\breve{F}(\boldsymbol{X},\boldsymbol{X}^*) = P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X})P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*,Y,\boldsymbol{X})F_0(\boldsymbol{X},\boldsymbol{X}^*).$$

It follows from condition (C2) that $\breve{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, $\breve{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0$, and $\breve{F} = F_0$. Thus, Theorem 4.2.1 holds. □

Theorem 4.2.2. Under conditions (C1)−(C5), $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ converge in distribution to zero-mean normal random vectors whose covariances attain their corresponding semiparametric efficiency bounds.

Proof. Denote the product of the two parametric models from the joint density as

$$P_{\boldsymbol{\theta}}(Y|\boldsymbol{X})P_{\boldsymbol{\gamma}}(Y^*|\boldsymbol{X}^*,Y,\boldsymbol{X}) \equiv P_{\boldsymbol{\lambda}}(Y,Y^*|\boldsymbol{X}^*,\boldsymbol{X}),$$

where $\boldsymbol{\lambda} = (\boldsymbol{\theta},\boldsymbol{\gamma})$. Let $\hat{\boldsymbol{\lambda}} = (\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$. The score function for the true values $\boldsymbol{\lambda}_0 = (\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)$

breaks into the sum of the separate score functions for $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$, i.e.,

$$U_{\boldsymbol{\lambda}} = \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(Y, Y^*|\boldsymbol{X}^*, \boldsymbol{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log P_{\boldsymbol{\theta}_0}(Y|\boldsymbol{X}) + \frac{\partial}{\partial \boldsymbol{\gamma}} \log P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*, Y, \boldsymbol{X}).$$

Let $U_F(h)$ denote the score function along the submodel $\{1 + \epsilon h(\boldsymbol{x}, \boldsymbol{x}^*)\}dF_0(\boldsymbol{x}, \boldsymbol{x}^*)$ based on one complete observation $(Y, \boldsymbol{X}, Y^*, \boldsymbol{X}^*)$, where $h \in L_2(\mathcal{P})$ for $\mathcal{P}$ the probability measure indexed by $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, F_0)$ and $E\{h(\boldsymbol{X}, \boldsymbol{X}^*)\} = 0$. The two-phase score operators, based on observed data, are defined as

$$U^o_{\boldsymbol{\lambda}} = VU_{\boldsymbol{\lambda}} + (1 - V)E\left(U_{\boldsymbol{\lambda}}|Y^*, \boldsymbol{X}^*\right), \text{ and } U^o_F = VU_F + (1 - V)E\left(U_F|Y^*, \boldsymbol{X}^*\right).$$

The information operator is

$$\begin{bmatrix} U^{o*}_{\boldsymbol{\lambda}} U^o_{\boldsymbol{\lambda}} & U^{o*}_{\boldsymbol{\lambda}} U^o_F \\ U^{o*}_F U^o_{\boldsymbol{\lambda}} & U^{o*}_F U^o_F \end{bmatrix},$$

where $U^{o*}_{\boldsymbol{\lambda}}$ and U^{o*}_F are the adjoint operators of $U^o_{\boldsymbol{\lambda}}$ and U^o_F, respectively. Elements of the information operator are calculated as

$$\begin{aligned} U^{o*}_{\boldsymbol{\lambda}} l^o_{\boldsymbol{\lambda}} &= E\left\{VU^{\otimes 2}_{\boldsymbol{\lambda}} + (1 - V)E(U_{\boldsymbol{\lambda}}|Y^*, \boldsymbol{X}^*)^{\otimes 2}\right\}, \\ U^{o*}_{\boldsymbol{\lambda}} U^o_F(h) = U^{o*}_F(h) U^{oT}_{\boldsymbol{\lambda}} &= E\left[E\left\{VU_{\boldsymbol{\lambda}} + (1 - V)E(U_{\boldsymbol{\lambda}}|Y^*, \boldsymbol{X}^*)|\boldsymbol{X}, \boldsymbol{X}^*\right\} h(\boldsymbol{X}, \boldsymbol{X}^*)\right], \\ U^{o*}_F(h) U^o_F(h) &= E\left(V|\boldsymbol{X}, \boldsymbol{X}^*\right) h(\boldsymbol{X}, \boldsymbol{X}^*) + E\left[(1 - V)E\left\{h(\boldsymbol{X}, \boldsymbol{X}^*)|Y^*, \boldsymbol{X}^*\right\}|\boldsymbol{X}, \boldsymbol{X}^*\right], \end{aligned}$$

where $\boldsymbol{c}^{\otimes} = \boldsymbol{c}\boldsymbol{c}^T$.

We aim to show the asymptotic normality and asymptotic efficiency of $\hat{\boldsymbol{\lambda}}$. In order to do that, we need to show that the information operator above is invertible. We note that the information operator is the sum of an invertible operator and a compact operator from the space $\mathbb{M} \equiv \mathbb{R}^{d_{\boldsymbol{\lambda}}} \times BV(\mathbb{D}_{\boldsymbol{x}, \boldsymbol{x}^*})$ to itself, where $d_{\boldsymbol{\lambda}}$ is the dimension of $\boldsymbol{\lambda}$ and $BV(\mathbb{D}_{\boldsymbol{x}, \boldsymbol{x}^*})$ is the space of functions with bounded total variation in the support of $(\boldsymbol{X}, \boldsymbol{X}^*)$. Using Theorem 4.7 of Rudin (1973), we can show that the information operator is invertible by demonstrating that the Fisher information along any nontrivial submodel is nonzero.

Suppose that the Fisher information is zero along some submodel

$$[\boldsymbol{\lambda}_0 + \epsilon \boldsymbol{c}, dF_0(\boldsymbol{x}, \boldsymbol{x}^*)\{1 + \epsilon h(\boldsymbol{x}, \boldsymbol{x}^*)\}],$$

where $\boldsymbol{c} \in \mathcal{C}$ is a vector of constants. The two-phase score along this submodel is $U^{oT}_{\boldsymbol{\lambda}} \boldsymbol{c} + U^o_F(h)$, and must also be zero. Consider the score contribution for a single

validated subject with complete data $(Y, \boldsymbol{X}, Y^*, \boldsymbol{X}^*) \in \mathcal{C}$ and constants $\boldsymbol{c}$ such that

$$U_{\boldsymbol{\lambda}}^{oT}\boldsymbol{c} + U_F^o(h) = 0.$$

For any pair of complete observations $(y_1^*, \boldsymbol{x}^*, y_1, \boldsymbol{x})$ and $(y_2^*, \boldsymbol{x}^*, y_2, \boldsymbol{x}) \in \mathcal{C}$, we have

$$\left\{\frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(y_1, y_1^*|\boldsymbol{x}^*, \boldsymbol{x})\right\}^T \boldsymbol{c} + h(\boldsymbol{x}, \boldsymbol{x}^*) = \left\{\frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(y_2, y_2^*|\boldsymbol{x}^*, \boldsymbol{x})\right\}^T \boldsymbol{c} + h(\boldsymbol{x}, \boldsymbol{x}^*),$$

which can be rewritten as a linear equation on the vector of constants $\boldsymbol{c}$:

$$\left\{\frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(y_1, y_1^*|\boldsymbol{x}^*, \boldsymbol{x}) - \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(y_2, y_2^*|\boldsymbol{x}^*, \boldsymbol{x})\right\}^T \boldsymbol{c} = 0.$$

By condition (C2), for this to be true it must be that $\boldsymbol{c} = \boldsymbol{0}$ and $h = 0$. Thus, the information must be nonzero along any submodel and the information operator is invertible. This further implies that there exists a function h such that $U_F^{o*}U_F^o(h) = U_F^{o*}U_{\boldsymbol{\lambda}}^o$, i.e.,

$$\begin{aligned} &E\left(V|\boldsymbol{X}, \boldsymbol{X}^*\right)h + E\left\{(1-V)E(h|\boldsymbol{X}^*, Y^*)|\boldsymbol{X}, \boldsymbol{X}^*\right\} \\ =&E\left\{VU_{\boldsymbol{\lambda}} + (1-V)E(U_{\boldsymbol{\lambda}}|Y^*, \boldsymbol{X}^*)|\boldsymbol{X}^*, \boldsymbol{X}\right\}, \end{aligned}$$

meaning that the least favorable direction for $\boldsymbol{\lambda}_0$ exists. Also, it can be shown that h is q-times continuously differentiable by using similar arguments as in the proof of Theorem 3.4 of Zeng (2005) in conjunction with conditions (C3) and (C4).

Recall that the SMLE $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{F}) = (\hat{\boldsymbol{\lambda}}, \hat{F})$ are defined to maximize the observed-data log-likelihood (expression (4.4)). Therefore, the derivatives of this function with respect to ϵ along the submodel $(\hat{\boldsymbol{\lambda}} + \epsilon\boldsymbol{c}, d\hat{F})$ must be zero for any $\boldsymbol{c}$, as is the derivative along submodel $\{\hat{\boldsymbol{\lambda}}, d\hat{F}(1 + h_N)\}$, where h_N is the projection of h onto the tangent space of the B-spline sieve space. By the approximation theory of B-splines, we have $||h_N - h||_{L_2} \lesssim s_N^{-q/d}$. The SMLE $(\hat{\boldsymbol{\lambda}}, \hat{F})$ can therefore be found by solving the functional

$$\Psi_N(\boldsymbol{\lambda}, F) = \Psi_{1N}(\boldsymbol{\lambda}, F) - \Psi_{2N}(\boldsymbol{\lambda}, F) = 0,$$

where

$$\Psi_{1N}(\boldsymbol{\lambda}, F) = \mathcal{P}_N\left\{V\frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}}(Y, Y^*|\boldsymbol{X}, \boldsymbol{X}^*)\right\} +$$

$$\mathcal{P}_N\left\{(1-V)\sum_{y=0}^{1}\int_{\boldsymbol{x}}\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\boldsymbol{\lambda}}(y,Y^*|\boldsymbol{x},\boldsymbol{X}^*)g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda},F)F(d\boldsymbol{x},\boldsymbol{X}^*)\right\}$$

$$\Psi_{2N}(\boldsymbol{\lambda},F)=\mathcal{P}_N\left\{Vh_N(\boldsymbol{X},\boldsymbol{X}^*)\right\}+$$

$$\mathcal{P}_N\left\{(1-V)\sum_{y=0}^{1}\int_{\boldsymbol{x}}g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda},F)h_N(\boldsymbol{x},\boldsymbol{X}^*)F(d\boldsymbol{x},\boldsymbol{X}^*)\right\},$$

$\mathcal{P}_N$ is the empirical measure of the sample, and

$$g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda},F)=\frac{P_{\boldsymbol{\theta}}(y|\boldsymbol{x})P_{\boldsymbol{\gamma}}(Y^*|\boldsymbol{X}^*,y,\boldsymbol{x})}{\sum_{\tilde{y}=0}^{1}\int_{\tilde{\boldsymbol{x}}}P_{\boldsymbol{\theta}}(\tilde{y}|\tilde{\boldsymbol{x}})P_{\boldsymbol{\gamma}}(Y^*|\boldsymbol{X}^*,\tilde{y},\tilde{\boldsymbol{x}})F(d\tilde{\boldsymbol{x}},\boldsymbol{X}^*)}. \tag{4.31}$$

Replace the empirical measure with the true $\mathcal{P}$ and define $\Psi(\boldsymbol{\lambda},F)$ following the same form as $\Psi_N(\boldsymbol{\lambda},F)$. Since $\Psi_N(\hat{\boldsymbol{\lambda}},\hat{F})=0$, it follows that $\hat{\boldsymbol{\lambda}}$ satisfies the following equation:

$$N^{-1/2}\left\{\Psi_N(\hat{\boldsymbol{\lambda}},\hat{F})-\Psi(\hat{\boldsymbol{\lambda}},\hat{F})\right\}=-N^{-1/2}\Psi(\hat{\boldsymbol{\lambda}},\hat{F}). \tag{4.32}$$

The left-hand side of equation (4.32) is an empirical process of the following two classes of functions indexed by $(\hat{\boldsymbol{\lambda}},\hat{F})$:

$$\mathcal{F}_{1N}=\left\{V\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\boldsymbol{\lambda}}(Y,Y^*|\boldsymbol{X},\boldsymbol{X}^*)+(1-V)\sum_{y=0}^{1}\int_{\boldsymbol{x}}\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\boldsymbol{\lambda}}(y,Y^*|\boldsymbol{x},\boldsymbol{X}^*)\times\right.$$
$$\left.g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda},F)F(d\boldsymbol{x},\boldsymbol{X}^*):|\boldsymbol{\lambda}-\boldsymbol{\lambda}_0|+||F-F_0||\le\epsilon_0\right\}$$

$$\mathcal{F}_{2N}=\left\{Vh_N(\boldsymbol{X},\boldsymbol{X}^*)+(1-V)\sum_{y=0}^{1}\int_{\boldsymbol{x}}g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda},F)\times\right.$$
$$\left.h_N(\boldsymbol{X},\boldsymbol{X}^*)F(d\boldsymbol{x},X^*):|\boldsymbol{\lambda}-\boldsymbol{\lambda}_0|+||F-F_0||\le\epsilon_0\right\},$$

where $||F-F_0||$ is the supreme norm in $\mathbb{D}_{\boldsymbol{x},\boldsymbol{x}^*}$. By Theorem 4.2.1 and the approximation theory of B-splines (Schumaker, 1981), it is straightforward to verify that

$$V\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\hat{\boldsymbol{\lambda}}}(Y,Y^*|\boldsymbol{X},\boldsymbol{X}^*)$$
$$+(1-V)\sum_{y=0}^{1}\int_{\boldsymbol{x}}\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\hat{\boldsymbol{\lambda}}}(y,Y^*|\boldsymbol{x},\boldsymbol{X}^*)g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\hat{\boldsymbol{\lambda}},\hat{F})\hat{F}(d\boldsymbol{x},\boldsymbol{X}^*)$$

converges uniformly in $(Y, \boldsymbol{X}, Y^*, \boldsymbol{X}^*)$ to

$$
\begin{aligned}
& V \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(Y, Y^* | \boldsymbol{X}, \boldsymbol{X}^*) \\
& + (1-V) \sum_{y=0}^{1} \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_0}(y, Y^* | \boldsymbol{x}, \boldsymbol{X}^*) \frac{P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x}) P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*, y, \boldsymbol{x}) F_0(d\boldsymbol{x}, \boldsymbol{X}^*)}{\sum_{y=0}^{1} \int_{\boldsymbol{x}} P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x}) P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*, y, \boldsymbol{x}) F_0(d\boldsymbol{x}, \boldsymbol{X}^*)} \\
= & V U_{\boldsymbol{\lambda}} + (1-V) E\left(U_{\boldsymbol{\lambda}} | Y^*, \boldsymbol{X}^*\right) = U^o_{\boldsymbol{\lambda}}
\end{aligned}
$$

Following these same steps, we can verify that

$$
V h_N(\boldsymbol{X}, \boldsymbol{X}^*) + (1-V) \sum_{y=0}^{1} \int_{\boldsymbol{x}} g_2(Y^*, y, \boldsymbol{X}^*, \boldsymbol{x}; \hat{\boldsymbol{\lambda}}, \hat{F}) h_N(\boldsymbol{X}, \boldsymbol{X}^*) \hat{F}(d\boldsymbol{x}, \boldsymbol{X}^*)
$$

converges uniformly in $(Y, \boldsymbol{X}, Y^*, \boldsymbol{X}^*)$ to

$$
\begin{aligned}
& V h(\boldsymbol{X}, \boldsymbol{X}^*) + (1-V) \frac{\sum_{y=0}^{1} \int_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{X}^*) P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x}) P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*, y, \boldsymbol{x}) F_0(d\boldsymbol{x}, \boldsymbol{X}^*)}{\sum_{y=0}^{1} \int_{\boldsymbol{x}} P_{\boldsymbol{\theta}_0}(y|\boldsymbol{x}) P_{\boldsymbol{\gamma}_0}(Y^*|\boldsymbol{X}^*, y, \boldsymbol{x}) F_0(d\boldsymbol{x}, \boldsymbol{X}^*)} \\
= & V h(\boldsymbol{X}, \boldsymbol{X}^*) + (1-V) E\left\{h(\boldsymbol{X}, \boldsymbol{X}^*) | Y^*, \boldsymbol{X}^*\right\} = U^0_F(h).
\end{aligned}
$$

Using Theorem 2.11.22 from van der Vaart and Wellner (1996), we can show that the left-hand side of equation (4.32) equals

$$
- N^{-1/2} (\mathcal{P}_N - \mathcal{P}) \left\{U^o_{\boldsymbol{\lambda}} - U^o_F(h_N)\right\} + o_p(1). \tag{4.33}
$$

In order to use this theorem, we need to verify its conditions. Clearly, all functions in the classes $\mathcal{F}_{1N}$ and $\mathcal{F}_{2N}$ are uniformly bounded. Next, we check the uniform entropy condition. Let f_1 and f_2 be two arbitrary functions from class $\mathcal{F}_{1N}$ indexed by $(\boldsymbol{\lambda}_1, F_1)$ and $(\boldsymbol{\lambda}_2, F_2)$, respectively. The difference between them is bounded above by

$$
\begin{aligned}
& \left| \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_1}(Y, Y^* | \boldsymbol{X}, \boldsymbol{X}^*) - \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_2}(Y, Y^* | \boldsymbol{X}, \boldsymbol{X}^*) \right| \\
& + \left| \sum_{y=0}^{1} \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_1}(y, Y^* | \boldsymbol{x}, \boldsymbol{X}^*) g_2(Y^*, y, \boldsymbol{X}^*, \boldsymbol{x}; \boldsymbol{\lambda}_1, F_1)(F_1 - F_2)(d\boldsymbol{x}, \boldsymbol{X}^*) \right| \\
& + \left| \sum_{y=0}^{1} \int_{\boldsymbol{x}} \left\{ \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_1}(y, Y^* | \boldsymbol{x}, \boldsymbol{X}^*) - \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\boldsymbol{\lambda}_2}(y, Y^* | \boldsymbol{x}, \boldsymbol{X}^*) \right\} \right. \\
& \quad \left. \times g_2(Y^*, y, \boldsymbol{X}^*, \boldsymbol{x}; \boldsymbol{\lambda}_1, F_1) F_2(d\boldsymbol{x}, \boldsymbol{X}^*) \right|
\end{aligned}
$$

$$
\begin{aligned}
&+\left|\sum_{y=0}^{1}\int_{\boldsymbol{x}}\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\boldsymbol{\lambda}_2}(y,Y^*|\boldsymbol{x},\boldsymbol{X}^*)\Big\{g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda}_1,F_1)\right.\\
&\quad\left.-g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda}_2,F_2)\Big\}F_2(d\boldsymbol{x},\boldsymbol{X}^*)\right|\\
&=(i)+(ii)+(iii)+(iv).
\end{aligned}
$$

Because the denominator in expression (4.31) is bounded away from zero, we obtain that

$$
\text{term }(ii)\lesssim\int_{\boldsymbol{x}}|F_1(\boldsymbol{x},\boldsymbol{X}^*)-F_2(\boldsymbol{x},\boldsymbol{X}^*)|d\boldsymbol{x}\lesssim\int_{\boldsymbol{x}^*}\int_{\boldsymbol{x}}|F_1(\boldsymbol{x},\boldsymbol{x}^*)-F_2(\boldsymbol{x},\boldsymbol{x}^*)|d\boldsymbol{x}d\boldsymbol{x}^*.
$$

Furthermore, by the mean-value theorem, we have

$$
\begin{aligned}
\text{term }(i)\lesssim&||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||,\\
\text{term }(iii)\lesssim&||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||\sum_{y=0}^{1}\int_{\boldsymbol{x}}g_2(Y^*,y,\boldsymbol{X}^*,\boldsymbol{x};\boldsymbol{\lambda}_1,F_1)F_2(d\boldsymbol{x},\boldsymbol{X}^*)\lesssim||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||\\
\text{term }(iv)\lesssim&||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||+\int_{\boldsymbol{x}^*}\int_{\boldsymbol{x}}|F_1(\boldsymbol{x},\boldsymbol{x}^*)-F_2(\boldsymbol{x},\boldsymbol{x}^*)|d\boldsymbol{x}d\boldsymbol{x}^*.
\end{aligned}
$$

Combining these upper bounds for terms (i)-(iv), we have

$$
|f_1-f_2|\lesssim||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||+\int_{\boldsymbol{x}^*}\int_{\boldsymbol{x}}|F_1(\boldsymbol{x},\boldsymbol{x}^*)-F_2(\boldsymbol{x},\boldsymbol{x}^*)|d\boldsymbol{x}d\boldsymbol{x}^*.
$$

Consider an arbitrary finite measure $\mathcal{Q}$. It follows from the Cauchy-Schwartz inequality that

$$
\begin{aligned}
||f_1-f_2||_{L_2(\mathcal{Q})}&\lesssim||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||+\left\{\int_{\boldsymbol{x}^*}\int_{\boldsymbol{x}}|F_1(\boldsymbol{x},\boldsymbol{x}^*)-F_2(\boldsymbol{x},\boldsymbol{x}^*)|^2d\boldsymbol{x}d\boldsymbol{x}^*\right\}^{1/2}\\
&=||\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2||+||F_1(\boldsymbol{X},\boldsymbol{X}^*)-F_2(\boldsymbol{X},\boldsymbol{X}^*)||_{L_2(\tilde{\mathcal{Q}})},
\end{aligned}
\tag{4.34}
$$

where $\tilde{\mathcal{Q}}$ is the uniform measure on $\mathbb{D}_{\boldsymbol{x},\boldsymbol{x}^*}$. Based on expression (4.34), we conclude that the covering numbers $N(\cdot,\cdot,\cdot)$ are related as follow:

$$
\begin{aligned}
N\left(\epsilon,\mathcal{F}_{1N},\mathcal{L}_2(\mathcal{Q})\right)\lesssim&N\left(\epsilon/2,\left\{\boldsymbol{\lambda}:||\boldsymbol{\lambda}-\boldsymbol{\lambda}_0||<\epsilon_0\right\},|\cdot|\right)\times\\
&N\left(\epsilon/2,\left\{F:||F-F_0||_\infty<\epsilon_0\right\},\mathcal{L}_2(\tilde{\mathcal{Q}})\right).
\end{aligned}
\tag{4.35}
$$

The covering number

$$
N\left(\epsilon/2,\left\{\boldsymbol{\lambda}:||\boldsymbol{\lambda}-\boldsymbol{\lambda}_0||<\epsilon_0\right\},|\cdot|\right)
$$

from expression (4.35) is $O(1/\epsilon^d)$. Next, observe that $\{F : ||F - F_0||_\infty < \epsilon\}$ is in the symmetric convex hull of a Vapnik-Chervonenkis class

$$\left[I\{\boldsymbol{a} < (\boldsymbol{X}^T, \boldsymbol{X}^{*T})^T \leq \boldsymbol{b}\} : \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{2d} \right].$$

Following from Theorem 2.6.9 of van der Vaart and Wellner (1996), we have that the last covering number $N\left(\epsilon/2, \{F : ||F - F_0||_\infty < \epsilon_0\}, \mathcal{L}_2(\mathcal{Q})\right)$ is $O(\exp\{\epsilon^{-2B/(B+2)}\})$ for some positive index B. From these results and expression (4.35), $\mathcal{F}_{1N}$ has been shown to satisfy the uniform entropy condition in Theorem 2.11.22 of van der Vaart and Wellner (1996). By similar arguments and the fact that $||h_N||_{L_2} \lesssim ||h||_{L_2}$, we can show that $\mathcal{F}_{2N}$ also satisfies this condition.

Replacing the arbitrary measure $\mathcal{Q}$ with true $\mathcal{P}$ in expression (4.34), we see that the functions in $\mathcal{F}_{1N}$ and $\mathcal{F}_{2N}$ are Lipschitz continuous with respect to $(\boldsymbol{\lambda}, F)$ in the metric defined as

$$p\left\{(\boldsymbol{\lambda}_1, F_1), (\boldsymbol{\lambda}_2, F_2)\right\} = ||\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2|| + ||F_1 - F_2||_{L_2(\mathcal{P})}.$$

Lastly, the total boundedness of the index set $(\boldsymbol{\lambda}, F)$ holds due to the precompactness of $(\boldsymbol{\lambda}, F)$ under the uniform metric. Thus, we have now verified all of the conditions in Theorem 2.11.22 of van der Vaart and Wellner (1996), and therefore, following from this theorem, equation (4.33) holds as desired.

Together equations (4.32) and (4.33) give us that

$$-N^{-1/2}\left\{\Psi_1(\hat{\boldsymbol{\lambda}}, \hat{F}) - \Psi_2(\hat{\boldsymbol{\lambda}}, \hat{F})\right\} = N^{1/2}(\mathcal{P}_N - \mathcal{P})\{U^o_{\boldsymbol{\lambda}} - U^o_F(h_N)\} + o_p(1), \quad (4.36)$$

where $\Psi_1(\hat{\boldsymbol{\lambda}}, \hat{F})$ and $\Psi_2(\hat{\boldsymbol{\lambda}}, \hat{F})$ are the same as $\Psi_{1N}(\hat{\boldsymbol{\lambda}}, \hat{F})$ and $\Psi_{2N}(\hat{\boldsymbol{\lambda}}, \hat{F})$, respectively, replacing $\mathcal{P}_N$ with $\mathcal{P}$. We can linearize the left-hand side of equation (4.36) around the truth $(\boldsymbol{\lambda}_0,\ F_0)$ to show that $\Psi_1(\hat{\boldsymbol{\lambda}}, \hat{F})$ is equal to

$$\begin{aligned}
&\Psi_1(\boldsymbol{\lambda}_0, F_0) + \mathcal{P}\left\{V \frac{\partial^2}{\partial \boldsymbol{\lambda}^T \partial \boldsymbol{\lambda}} \log P_{\tilde{\boldsymbol{\lambda}}}(Y, Y^* | \boldsymbol{X}, \boldsymbol{X}^*)(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)\right\} \\
&+ \mathcal{P}\left[(1 - V)\sum_{y=0}^{1} \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{\lambda}} \left\{\frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\tilde{\boldsymbol{\lambda}}}(y, Y^* | \boldsymbol{x}, \boldsymbol{X}^*) g_2(Y^*, y, \boldsymbol{X}^*, \boldsymbol{x}; \tilde{\boldsymbol{\lambda}}, \tilde{F})\right\}\right. \\
&\left. \times (\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)\hat{F}(d\boldsymbol{x}, \boldsymbol{X}^*)\right] \\
&+ \mathcal{P}\left[(1 - V)\sum_{y=0}^{1} \int_{\boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{\lambda}} \log P_{\tilde{\boldsymbol{\lambda}}}(y, Y^* | \boldsymbol{x}, \boldsymbol{X}^*) \left\{\frac{\partial}{\partial F} g_2(Y^*, y, \boldsymbol{X}^*, \boldsymbol{x}; \tilde{\boldsymbol{\lambda}}, \tilde{F})\right.\right.
\end{aligned}$$

$$\times \tilde{F}(d\boldsymbol{x}, \boldsymbol{X}^*)\Bigg\}(\hat{F} - F_0)\Bigg],$$

where $\partial/\partial F$ denotes the pathwise derivative, and $(\tilde{\boldsymbol{\lambda}}, \tilde{F})$ lies between $(\hat{\boldsymbol{\lambda}}, \hat{F})$ and $(\boldsymbol{\lambda}_0, F_0)$. Similar expansions can be obtained for $\Psi_2(\hat{\boldsymbol{\lambda}}, \hat{F})$. By the approximation theory of B-splines (Schumaker, 1981), we can also show that the left-hand side of (4.36) equals

$$\begin{aligned}
&- N^{-1/2}\{1 + o_p(1)\}E\left\{U^o_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) + U^o_{\boldsymbol{\lambda}F}(\hat{F} - F_0) - U^o_{F\boldsymbol{\lambda}}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) - U^o_{FF}(h_N, \hat{F} - F_0)\right\} \\
&- N^{-1/2}\left\{\Psi_1(\boldsymbol{\lambda}_0, F_0) - \Psi_2(\boldsymbol{\lambda}_0, F_0)\right\}, \qquad (4.37)
\end{aligned}$$

where the derivatives in expression (4.37) are as follows: $U^o_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$ is the derivative of $U^o_{\boldsymbol{\lambda}}$ with respect to $\boldsymbol{\lambda}$; $U^o_{\boldsymbol{\lambda}F}(h)$ is the derivative of $U^o_{\boldsymbol{\lambda}}$ with respect to F along the direction h; $U^o_{F\boldsymbol{\lambda}}(h)$ is the derivative of $U^o_F(h)$ with respect to $\boldsymbol{\lambda}$; and $U^o_{FF}(h_1, h_2)$ is the derivative of $U^o_F(h_1)$ with respect to F along the direction h_2.

Since h was chosen to be the least-favorable direction for $\boldsymbol{\lambda}_0$ and $||h_N - h|| \lesssim s_N^{-q/d}$, it follows that

$$\begin{aligned}
E\left\{U^o_{FF}(h_N, \hat{F} - F_0)\right\} =& E\left\{U^o_{\boldsymbol{\lambda}F}(\hat{F} - F_0)\right\} + O(s_N^{-q/d}), \\
E\left\{U^o_{F\boldsymbol{\lambda}}(h_N)(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)\right\} =& E\left\{U^o_{F\boldsymbol{\lambda}}(h)(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)\right\} + O(N^{1/2}s_N^{-q/d}).
\end{aligned}$$

Thus, by condition (C4) the first term in expression (4.37) is

$$N^{-1/2}\boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) + O(N^{1/2}s_N^{-q/d}) = N^{1/2}\boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) + o(1),$$

where $\boldsymbol{\Sigma} = -E\left\{U^o_{\boldsymbol{\lambda}\boldsymbol{\lambda}} - U^o_{F\boldsymbol{\lambda}}(h)\right\}$. The covariance matrix $\boldsymbol{\Sigma}$ is invertible following from the invertability of the information for $(\boldsymbol{\lambda}_0, F_0)$. The last term in expression (4.37) equals zero because

$$\begin{aligned}
&\mathcal{P}\left\{V\frac{\partial}{\partial\boldsymbol{\lambda}}P_{\boldsymbol{\lambda}_0}(Y, Y^*|\boldsymbol{X}, \boldsymbol{X}^*)\right\} = 0, \\
&\mathcal{P}\left\{(1 - V)\sum_{y=0}^{1}\int_{\boldsymbol{x}}\frac{\partial}{\partial\boldsymbol{\lambda}}\log P_{\boldsymbol{\lambda}_0}(y, Y^*|\boldsymbol{x}, \boldsymbol{X}^*)\frac{P_{\boldsymbol{\lambda}_0}(y, Y^*|\boldsymbol{x}, \boldsymbol{X}^*)F_0(d\boldsymbol{x}, \boldsymbol{X}^*)}{\sum_{y=0}^{1}\int_{\boldsymbol{x}}P_{\boldsymbol{\lambda}_0}(y, Y^*|\boldsymbol{x}, \boldsymbol{X}^*)F_0(d\boldsymbol{x}, \boldsymbol{X}^*)}\right\} = 0.
\end{aligned}$$

Finally, it follows from equation (4.36) that

$$N^{1/2}\left\{1 + o_p(1)\right\}\boldsymbol{\Sigma}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0) + o_p(1) = N^{1/2}(\mathcal{P}_N - \mathcal{P})\left\{U^o_{\boldsymbol{\lambda}} - U^o_F(h)\right\},$$

thus establishing the asymptotic normality in Theorem 4.2.2. Further, $\boldsymbol{\Sigma}^{-1}\{U^o_{\boldsymbol{\lambda}} - U^o_F(h)\}$ is the efficient influence function for $\boldsymbol{\lambda}_0$, so the limiting covariance matrix attains the semiparametric efficiency bound. □

4.6.2 Additional Simulation Studies

4.6.2.1 Validity checks in a larger Phase I sample

All variables were generated following Section 4.3.1. From $N = 5000$ Phase I subjects, proportions of $p_v = 0.1$, 0.25, or 0.5 were selected for validation through SRS or naive case-control sampling. Results in Table 4.6 are consistent with what we saw with increasing $N = 1000$ to 2000 in Table 4.1.

Table 4.6: Simulation results for validity checks with outcome misclassification and a binary error-prone covariate based on larger Phase I sample size

	SMLE		MLE			Complete-Case			HT			Raking		
p_v	Bias	SE	Bias	SE	RE	Bias	SE	RE	Bias	SE	RE	Bias	SE	RE
						SRS								
0.10	0.006	0.170	0.006	0.170	1.000	−0.013	0.197	0.744	−0.013	0.197	0.744	−0.009	0.183	0.861
0.25	−0.001	0.110	−0.001	0.110	1.002	0.004	0.122	0.808	0.004	0.122	0.808	0.003	0.115	0.916
0.50	−0.003	0.083	−0.003	0.083	1.001	−0.007	0.089	0.879	−0.007	0.089	0.879	−0.006	0.084	0.976
						1:1 case-control sampling based on Y^*								
0.10	−0.005	0.151	−0.005	0.151	1.004	0.005	0.185	0.668	−0.005	0.186	0.659	−0.005	0.173	0.763
0.25	−0.001	0.097	−0.001	0.097	1.003	0.008	0.114	0.718	−0.001	0.116	0.700	−0.001	0.106	0.832
0.50	−0.003	0.076	−0.003	0.076	0.999	0.008	0.082	0.870	−0.001	0.083	0.851	0.001	0.079	0.948

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; RE is the relative efficiency of the estimator to the SMLE. Each entry is based on 1000 replicates. The SMLE converged in all replications. The Monte Carlo simulation error for the bias of the SMLE was ≤ 0.006.

4.6.2.2 Robustness of the SMLE and MLE

A continuous, error-free covariate X_a was generated from a standard normal distribution with mean 0 and variance 1. Misclassification-prone X^*_b was generated from a Bernoulli distribution with $P(X^*_b = 1|X_b, Y, X_a) = [1 + \exp\{-(-1.1 + 2.2X_b + 0.5X_a + \delta_3 X^2_a)\}]^{-1}$. The coefficient δ_3 was varied in between -0.5 and 0.5. All other variables were generated as in Section 4.3.1, with the Phase I and II sample sizes fixed at $N = 1000$ and $n = 250$, respectively. The fully-parametric MLE assumed the covariate error model to be $P(X_b = 1|X^*_b, Y, X_a) = [1+\exp\{-(\delta_0+\delta_1 X_b+\delta_2 X_a)\}]^{-1}$, i.e., main effects only. Clearly, when $\delta_3 = 0$ the MLE will be correctly specified but it will be misspecified otherwise. The SMLE placed separate cubic B-splines on X_a with 10 each interior knots on subjects with $X_b = 0$ and $X_b = 1$. SRS and naive case-control were used to select Phase II. The bias of the MLE remained reasonably small

(Table 4.7), suggesting that logistic regression can be fairly robust in these settings, even for large choices of δ_3. The bias of the generalized raking estimator was around 5% in all settings, likely due to the small audit size of $n = 250$; it reduced to less than 1% when $n = 500$ (data not shown).

In a second set of simulations, error-prone Y^* was generated from a Bernoulli distribution with $P(Y^* = 1|X_b^*, Y, X_b, X_a) = [1 + \exp\{-(-2.2 - 0.2X_b^* + 5.14Y - 0.2X_b - 0.1X_a + \theta_5 X_a^2)\}]^{-1}$. The coefficient θ_5 was varied between -0.5 and 0.5; when $\theta_5 = 0$ the model reduces to main effects only. All other variables were generated as above with $\delta_3 = -0.1$, and the same B-spline specification was used for the SMLE. We compare the performance of the SMLE when the logistic regression model $P(Y^*|X_b^*, Y, X_b, X_a)$ included the quadratic term X_a^2, which was correctly specified for $\theta_5 \neq 0$, to the SMLE when this model assumed main effects only, which was misspecified for all $\theta_5 \neq 0$. The fully-parametric MLE assumed main effects only in the covariate and exposure error models such that $P(X_b = 1|X_b^*, Y, X_a) = [1+\exp\{-(\delta_0+\delta_1 X_b+ \delta_2 X_a)\}]^{-1}$ and $P(Y^*|X_b^*, Y, X_b, X_a) = [1+\exp\{-(\theta_0+\theta_1 X_b^*+\theta_2 Y+\theta_3 X_b+\theta_4 X_a)\}]^{-1}$, respectively. SRS and naive case-control sampling were used to select the Phase II subsample. While the SMLE was slightly more biased when the outcome error mechanism was misspecified (Table 4.8), it was still no more than 5% biased and coverage probabilities for the 95% CI remained reasonable. The bias tended to be larger when θ_5 was positive than negative, even for the same magnitude of the coefficient; for example there was 5% bias when $\theta_5 = 0.5$ and the SMLE assumed the main effects only but 2.5% bias when $\theta_5 = -0.5$ and the same model was used. As expected, the SMLE assuming main effects only performed similarly to the MLE. Performance of the raking estimator was similar to what was seen in Table 4.7, with low bias but less efficiency than the SMLE or MLE. Like those in Table 4.7, these simulations suggest that logistic regression can be fairly robust to model misspecification.

4.6.2.3 Classical covariate measurement error

Continuous covariate X_b was generated from a standard normal distribution. Error-free covariate X_a and outcome Y were generated from Bernoulli distributions with $P(X_a = 1) = 0.25$ and $P(Y = 1|X_b, X_a) = [1 + \exp\{-(-1 + \beta X_b - 0.5X_a)\}]^{-1}$, respectively. Error-prone covariate X_b^* was generated by $X_b^* = X_b + U$, where U was a normal random variable with mean zero and variance $\sigma_U^2 = 0.5$. The outcome, Y, was assumed to be error-free. The Phase I and II sample sizes were $N = 1000$ and $n = 250$, respectively, and Phase II was selected by SRS and traditional case-control sampling. The effect β was varied between -2 and 2. When implementing the SMLE,

Table 4.7: Simulation results under complex specification of the covariate error mechanism

	SMLE		MLE		Raking	
δ_3	Bias	SE	Bias	SE	Bias	SE
			SRS			
−0.5	−0.001	0.246	0.009	0.240	−0.010	0.266
−0.2	0.001	0.242	0.008	0.238	−0.010	0.263
−0.1	0.002	0.243	0.007	0.238	−0.010	0.264
0.0	0.001	0.243	0.007	0.237	−0.011	0.265
0.1	0.001	0.243	0.007	0.236	−0.012	0.264
0.2	0.001	0.241	0.007	0.235	−0.013	0.265
0.5	0.005	0.246	0.009	0.240	−0.013	0.267
			1:1 case-control sampling based on Y^*			
−0.5	−0.015	0.241	−0.010	0.236	−0.013	0.252
−0.2	−0.004	0.235	−0.003	0.229	−0.002	0.253
−0.1	−0.014	0.240	−0.011	0.234	−0.003	0.253
0.0	−0.001	0.237	0.000	0.232	−0.007	0.260
0.1	−0.013	0.230	−0.009	0.224	−0.002	0.245
0.2	−0.004	0.241	0.000	0.235	−0.010	0.251
0.5	−0.012	0.240	−0.006	0.236	−0.002	0.241

Note: The error-prone exposure X_b^* was generated from a Bernoulli distribution with $P(X_b^* = 1|X_b, Y, X_a) = [1 + \exp\{-(-1.1 + 2.2X_b + 0.5X_a + \delta_3 X_a^2)\}]^{-1}$. The fully-parametric MLE specified this model with main effects only; the SMLE and raking estimator require no such model specification. Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator. Each entry is based on 1000 replicates. The SMLE had greater than 98% convergence rates in all settings. The Monte Carlo simulation error for the bias of the SMLE was ≤ 0.008.

Table 4.8: Simulation results under complex specification of the outcome and covariate error mechanisms

	SMLE											
	Includes quadratic term				Main effects only				MLE		Raking	
θ_5	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	Bias	SE
						SRS						
−0.5	0.008	0.242	0.242	0.954	−0.005	0.248	0.244	0.951	−0.006	0.247	−0.008	0.262
−0.2	0.005	0.241	0.241	0.945	0.001	0.242	0.242	0.947	−0.005	0.247	−0.009	0.263
−0.1	0.004	0.242	0.242	0.942	0.002	0.243	0.242	0.944	−0.005	0.248	−0.008	0.264
0.0	0.002	0.243	0.242	0.949	0.002	0.243	0.242	0.949	−0.007	0.249	−0.010	0.264
0.1	0.004	0.244	0.243	0.951	0.005	0.245	0.243	0.951	−0.007	0.249	−0.010	0.264
0.2	0.005	0.242	0.243	0.952	0.007	0.244	0.244	0.950	−0.007	0.251	−0.010	0.266
0.5	0.007	0.238	0.244	0.956	0.010	0.249	0.248	0.947	−0.011	0.254	−0.009	0.270
						1:1 case-control sampling based on Y^*						
−0.5	−0.003	0.243	0.230	0.938	−0.011	0.248	0.233	0.939	0.006	0.227	0.007	0.240
−0.2	−0.012	0.234	0.230	0.947	−0.014	0.234	0.230	0.945	0.000	0.233	−0.003	0.248
−0.1	−0.003	0.241	0.230	0.945	−0.004	0.242	0.230	0.943	0.003	0.230	0.003	0.245
0.0	−0.004	0.241	0.230	0.940	−0.004	0.241	0.230	0.941	−0.011	0.234	−0.009	0.252
0.1	0.003	0.233	0.230	0.946	0.004	0.233	0.231	0.947	−0.010	0.236	−0.005	0.252
0.2	−0.008	0.240	0.232	0.948	−0.007	0.241	0.232	0.946	−0.001	0.242	−0.001	0.258
0.5	0.003	0.245	0.235	0.944	0.007	0.255	0.239	0.936	−0.010	0.234	−0.007	0.251

Note: The error-prone outcome Y^* was generated from a Bernoulli distribution with $P(Y^* = 1|X_b^*, Y, X_b, X_a) = [1+\exp\{-(-2.2-0.2X_b^*+5.14Y-0.2X_b-0.1X_a+\theta_5 X_a^2)\}]^{-1}$. Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; SEE is the average of the standard error estimator; CP is the coverage probability of the 95% confidence interval. Each entry is based on 1000 replicates. The SMLE had greater than 95% convergence rates in all settings. The Monte Carlo simulation error for the bias and coverage probability of the SMLE were ≤ 0.008 and $\leq 0.8\%$, respectively, when the model included the quadratic and ≤ 0.009 and $\leq 0.9\%$, respectively, when the model did not.

Table 4.9: Simulation results under continuous covariate error with additive measurement error of varied effect size β

	SMLE		RC		Raking	
β	Bias	SE	Bias	SE	Bias	SE
			SRS			
-2	-0.069	0.215	0.268	0.137	-0.033	0.244
-1	-0.033	0.122	0.038	0.101	-0.018	0.149
0	0.000	0.093	-0.001	0.089	-0.002	0.107
1	0.045	0.126	-0.031	0.102	0.011	0.142
2	0.087	0.220	-0.260	0.138	0.035	0.247
			1:1 case-control sampling based on Y			
-2	-0.070	0.231	0.262	0.140	-0.050	0.241
-1	-0.036	0.117	0.019	0.098	-0.019	0.135
0	0.003	0.089	0.003	0.089	0.007	0.093
1	0.029	0.126	-0.025	0.106	0.009	0.141
2	0.075	0.220	-0.253	0.143	0.028	0.229

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator. Each entry is based on 1000 replicates. The SMLE converged in all replications. The Monte Carlo simulation error for the bias of the SMLE was ≤ 0.007.

we used the histogram B-spline basis and varied b_N from 16 to 52 to assess its effects on model fitting. As in the main text, we maintained a 3:1 ratio of the number of knots allocated to subjects with $X_a = 0$:$X_a = 1$. Results were very similar for $b_N \geq 32$, i.e., the maximum difference in the coverage probability of the 95% confidence interval was less than 0.5%. Consequently, separate histogram B-splines with 24 and 8 interior knots were used for subjects with $X_a = 0$ and $X_a = 1$, respectively. We note that this B-spline setup differs from that for the same Phase I sample size in Section 4.3.2.

Performance of the SMLE under classical covariate measurement error is compared to RC and generalized raking in Table 4.9. The SMLE and generalized raking estimator performed well, with the SMLE generally being slightly more efficient than the raking estimator. The RC estimator performed well when β was small to moderate but became biased when β was large. The findings were similar when either SRS or case-control sampling was used to select Phase II.

Table 4.10: Simulation results for the naive estimator under outcome misclassification and a continuous covariate with varied additive measurement error variance when the Phase II design is simple random sampling

		Errors in Y^* and X_b^*					
	Error in X_b^*	$\gamma_2 = -1$		$\gamma_2 = 0$		$\gamma_2 = 1$	
σ_U^2	Bias	Bias	FPR/FNR	Bias	FPR/FNR	Bias	FPR/FNR
0.10	−0.105	0.021	0.1/0.04	0.518	0.14/0.06	−0.385	0.09/0.06
0.25	−0.228	−0.125	0.1/0.04	0.357	0.14/0.06	−0.525	0.10/0.06
0.50	−0.372	−0.294	0.1/0.04	0.173	0.15/0.06	−0.691	0.11/0.07
0.75	−0.470	−0.408	0.1/0.04	0.049	0.15/0.07	−0.806	0.11/0.07
1.00	−0.542	−0.490	0.1/0.04	−0.039	0.15/0.07	−0.892	0.12/0.08
1.25	−0.596	−0.552	0.1/0.04	−0.104	0.16/0.07	−0.958	0.13/0.09
1.50	−0.639	−0.601	0.1/0.04	−0.155	0.16/0.08	−1.010	0.14/0.09
1.75	−0.674	−0.640	0.1/0.04	−0.195	0.17/0.08	−1.050	0.14/0.10
2.00	−0.703	−0.672	0.1/0.04	−0.227	0.17/0.08	−1.090	0.15/0.10

Note: Bias is the empirical bias of the naive estimator; FPR and FNR are the average false positive rate and false negative rate, respectively, for Y^*. Each entry is based on 1000 replicates.

4.6.2.4 Naive estimator

Intuitively, larger errors in X_b^* would result in larger bias for the naive estimator. This was not always the case, as could be seen from the simulation results in Table 4.10. When only X_b^* was error-prone, larger σ_U^2 led to larger negative bias. When both Y^* and X_b^* were error-prone, and the errors were correlated, the direction and magnitude of the bias depended on both σ_U^2 and γ_2, where γ_2 is the regression coefficient for X_b^* in $P(Y^* = 1|X_b^*, Y, X_b, X_a)$.

4.6.2.5 Systematically biased covariate error

In this set of simulation studies, the mean of the additive error U depended on the error-free covariate X_a, i.e.,

$$U \sim \begin{cases} N(\mu_0,\, \sigma^2 = 0.1), \text{ if } X_a = 0; \\ N(\mu_1,\, \sigma^2 = 0.1), \text{ if } X_a = 1. \end{cases}$$

The means μ_0 and μ_1 were chosen from $\{0, 1, 2\}$. All other variables were generated as in Section 4.3.2.1 with Phase I and II sample sizes of $N = 1000$ and $n = 250$, respectively. The Phase II sample was selected by naive case-control sampling. Results for the SMLE, HT analysis, and raking estimator are included in Table 4.11. All estimators remained unbiased under additive errors that were not centered at zero. The SMLE was much more efficient than the HT or raking estimators, but the relative efficiency of raking to the SMLE approached 1 for $\mu_0 = 2$ and larger choices of μ_1.

Table 4.11: Simulation results under outcome misclassification and a continuous covariate with additive measurement error that may not center at zero

		SMLE		HT			Raking		
μ_0	μ_1	Bias	SE	Bias	SE	RE	Bias	SE	RE
0	0	−0.028	0.148	0.028	0.170	0.758	0.016	0.166	0.795
	1	−0.023	0.153	0.025	0.179	0.731	0.017	0.167	0.839
	2	−0.020	0.155	0.015	0.174	0.794	0.019	0.168	0.851
1	0	−0.013	0.154	0.016	0.175	0.774	0.018	0.169	0.830
	1	−0.016	0.148	0.014	0.166	0.795	0.019	0.170	0.758
	2	−0.005	0.160	0.018	0.184	0.756	0.020	0.171	0.875
2	0	0.013	0.167	0.015	0.182	0.842	0.020	0.175	0.911
	1	0.015	0.173	0.027	0.186	0.865	0.021	0.175	0.977
	2	0.012	0.176	0.018	0.194	0.823	0.022	0.176	1.000

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; RE is the relative efficiency of the estimator to the SMLE. Each entry is based on 1000 replicates. The Monte Carlo simulation error for the bias of the SMLE was ≤ 0.006.

4.6.2.6 Multiplicative covariate error

We set $X_b^* = X_b U$, where the multiplicative error U was generated as follows:

$$U \sim \begin{cases} Unif(0, \eta_0), \text{ if } X_a = 0; \\ Unif(0, \eta_1), \text{ if } X_a = 1. \end{cases}$$

All other variables were generated as in Section 4.3.2.1 with Phase I and II sample sizes of $N = 1000$ and $n = 250$, respectively. The Phase II sample was selected by naive case-control sampling. Results for the SMLE, HT analysis, and the raking estimator are included in Table 4.12. The SMLE continued to perform well in this setting and was substantially more efficient than the other estimators.

4.6.3 Additional Results of the CCASAnet Data Analysis

4.6.3.1 Sensitivity analysis of death within two years of ART initiation

There were 288 subjects who met criteria (1)–(3) from Section 4.4 but were excluded from the main analysis because they died within two years of initiating ART. Here we investigate potential survivor bias resulting from these exclusions. Thus, a Phase I sample of 5379 subjects was extracted from the CCASAnet research database, of whom 122 (2.3%) were audited.

We define a composite endpoint of either ADE or death within two years of ini-

Table 4.12: Simulation results under outcome misclassification and a continuous covariate with multiplicative measurement error

		SMLE		HT			Raking		
η_0	η_1	Bias	SE	Bias	SE	RE	Bias	SE	RE
1	1	−0.007	0.146	0.024	0.176	0.688	0.011	0.162	0.815
	2	−0.004	0.144	0.029	0.177	0.662	0.009	0.168	0.732
	3	0.002	0.146	0.027	0.175	0.696	0.006	0.169	0.743
2	1	0.002	0.151	0.024	0.179	0.712	0.010	0.170	0.787
	2	−0.014	0.150	0.024	0.178	0.710	0.011	0.171	0.768
	3	−0.013	0.156	0.024	0.184	0.719	0.009	0.171	0.830
3	1	0.013	0.157	0.020	0.176	0.796	0.011	0.171	0.842
	2	−0.014	0.155	0.020	0.178	0.758	0.011	0.171	0.820
	3	−0.015	0.164	0.024	0.190	0.745	0.010	0.171	0.920

Note: Bias and SE are, respectively, the empirical bias and standard error of the parameter estimator; RE is the relative efficiency of the estimator to the SMLE. Each entry is based on 1000 replicates. The SMLE had $\geq 95\%$ convergence in all settings. The Monte Carlo simulation error for the bias of the SMLE was ≤ 0.005.

tiating ART, with 905 unvalidated events (17% prevalence) and 19 validated events (16% prevalence). During the audits, the VDCC identified 6% misclassification in the outcome, all false negatives. There was also 6% misclassification in AIDS at initiation, with 4% false positive rate and 2% false negative rate, and 7% error rate in CD4 count, with mean magnitude of −0.03 and variance 0.24 on the square root scale. No subject had errors in both the outcome and covariates and only one had errors in both CD4 count and AIDS status, suggesting that there was little evidence of error correlation.

In Table 4.14, results of the SMLE are presented with the naive, complete-case, HT, and generalized raking analyses for comparison. The clinical conclusions drawn based on these models are all in alignment with those from the main analysis: the naive analysis and SMLE found both CD4 count and AIDS at initiation to be significantly associated with poor prognosis (in this case, death or ADE), while the complete-case, HT, and raking analyses only found CD4 count to be.

Table 4.13: log OR estimates and 95% confidence intervals from the analysis of the CCASAnet dataset using the SMLE approach with various B-spline bases

	Not stratified by site group		Stratified by site group	
Predictor	log OR	95% CI	log OR	95% CI
	B-spline basis specification (a)			
$\sqrt{\text{CD4}/10}$	−0.377	(−0.580, −0.174)	−0.344	(−0.553, −0.135)
AIDS	0.196	(−0.484, 0.876)	0.191	(−0.516, 0.897)
Site: A	−1.653	(−2.395, −0.911)	−1.566	(−2.357, −0.775)
Site: C	0.038	(−0.088, 0.164)	0.057	(−0.065, 0.179)
Site: D	−2.354	(−3.244, −1.463)	−2.224	(−3.111, −1.337)
Site: E	−0.786	(−1.401, −0.171)	−0.741	(−1.331, −0.150)
Male	0.254	(−0.603, 1.112)	0.203	(−0.638, 1.045)
Age (/10 yrs)	−1.305	(−2.211, −0.399)	−1.278	(−2.201, −0.355)
Year of ART	0.439	(−0.195, 1.073)	0.443	(−0.201, 1.087)
	B-spline basis specification (b)			
$\sqrt{\text{CD4}/10}$	−0.482	(−0.724, −0.239)	−0.471	(−0.817, −0.125)
AIDS	1.386	(0.580, 2.193)	0.792	(−0.170, 1.754)
Site: A	1.124	(0.272, 1.977)	−2.165	(−3.325, −1.004)
Site: C	0.183	(0.002, 0.365)	0.014	(−0.142, 0.170)
Site: D	−1.222	(−2.389, -0.055)	−1.788	(−3.129, −0.447)
Site: E	−0.733	(−1.725, 0.258)	−1.170	(−1.926, −0.413)
Male	−0.709	(−1.943, 0.524)	0.094	(−1.058, 1.247)
Age (/10 yrs)	−0.694	(−1.648, 0.261)	−1.068	(−2.521, 0.385)
Year of ART	−0.505	(−1.224, 0.214)	0.293	(−0.595, 1.180)
	B-spline basis specification (c)			
$\sqrt{\text{CD4}/10}$	−0.482	(−0.725, −0.240)	−0.433	(−0.683, −0.182)
AIDS	1.388	(0.582, 2.194)	1.384	(0.508, 2.260)
Site: A	1.129	(0.278, 1.980)	1.218	(0.284, 2.152)
Site: C	0.184	(0.003, 0.365)	0.179	(−0.030, 0.388)
Site: D	−1.225	(−2.394, −0.056)	−1.157	(−2.268, −0.045)
Site: E	−0.732	(−1.725, 0.260)	−0.484	(−1.487, 0.519)
Male	−0.703	(−1.933, 0.527)	−0.793	(−2.025, 0.440)
Age (/10 yrs)	−0.690	(−1.644, 0.263)	−0.706	(−1.703, 0.290)
Year of ART	−0.508	(−1.225, 0.210)	−0.489	(−1.202, 0.224)

Note: B-spline basis specification (a) corresponds to zero and zero interior knot for subjects with and without AIDS at initiation, respectively; (b) corresponds to one and zero interior knot for subjects with and without AIDS at initiation, respectively; (c) to corresponds to one and one interior knot for subjects with and without AIDS at initiation, respectively. 95% CI is the 95% confidence interval.

Table 4.14: log OR estimates and 95% confidence intervals from the analysis of the composite outcome (death or ADE within two years) in the CCASAnet dataset

	Naive		Complete-Case		HT		Raking		SMLE	
Covariate	log OR	95% CI	log OR	95% CI	log OR	95% CI	log OR	95% CI	log OR	95% CI
$\sqrt{\text{CD4}/10}$	−0.285	(−0.334, −0.236)	−0.621	(−1.032, −0.211)	−0.860	(−1.273, −0.488)	−0.549	(−0.868, −0.230)	−0.378	(−0.717, −0.040)
AIDS	1.371	(1.202, 1.540)	0.750	(−0.416, 1.917)	0.520	(−0.872, 1.912)	0.613	(−0.472, 1.698)	1.472	(0.623, 2.321)
Site: A	−1.343	(−1.604, −1.083)	−0.610	(−2.363, 1.143)	−0.831	(−2.953, 1.291)	−0.254	(−1.724, 1.215)	0.582	(−0.447, 1.611)
Site: C	0.109	(−0.105, 0.323)	−0.075	(−1.778, 1.627)	−0.232	(−2.211, 1.746)	0.034	(−1.437, 1.505)	0.058	(−0.739, 0.852)
Site: D	−0.458	(−0.729, −0.187)	−1.957	(−3.794, −0.119)	−2.395	(−4.300, −0.489)	−1.665	(−3.567, 0.238)	−1.722	(−2.760, −0.685)
Site: E	−0.333	(−0.620, −0.046)	−1.344	(−3.844, 1.157)	−1.627	(−4.570, 1.316)	−0.869	(−3.229, 1.491)	−1.220	(−2.258, −0.182)
Male	0.073	(−0.113, 0.259)	−0.967	(−2.275, 0.341)	−0.956	(−2.320, 0.409)	−1.250	(−2.562, 0.064)	−0.436	(−1.065, 0.194)
Age/10 years	0.077	(−0.005, 0.158)	0.215	(−0.358, 0.788)	0.135	(−0.589, 0.859)	0.166	(−0.351, 0.683)	−0.128	(−0.392, 0.136)
Year of ART	−0.028	(−0.050, −0.005)	0.089	(−0.104, 0.282)	0.186	(−0.080, 0.452)	0.072	(−0.116, 0.260)	−0.074	(−0.146, −0.001)

Note: 95% CI is the 95% confidence interval.

CHAPTER 5

CONCLUSION

This dissertation and the research within it were driven by a demonstrated need for methods to promote the practical and statistical efficiency of two-phase designs for data quality. We know first-hand from our collaborations with two large-scale, observational HIV cohorts, the Carribean, Central, and South America network for HIV Epidemiology (CCASAnet) and the Vanderbilt Comprehensive Care Clinic (VCCC), that, while the benefits of data auditing are clear, they are resource-intensive undertakings. Therefore, simply put, we sought methods to make the most of the investment in two-phase studies with respect to the design, implementation, and analysis. The proposed methods from Chapters 2–4 can be used together, so that this entire dissertation can serve as a start-to-finish guide for likelihood-based analyses in error-prone binary response data.

We began by answering the question "who should we audit?" In Chapter 2, we derived the optimal, i.e., lowest-variance, two-phase design for likelihood-based analysis of data with binary outcome and exposure misclassification. This setting was not yet addressed in the literature; to our knowledge, the proposed optimal design is the first to capture complex outcome misclassification in addition to covariate error. With this design as our gold standard, we introduced a multi-wave design which can be implemented to approximate it in practice without knowledge of modeling parameters, unlike the optimal design. Since the variance of the MLE cannot be minimized directly, we developed a novel adaptive grid search algorithm to solve for the optimal design. We demonstrated the superior efficiency of the proposed designs through extensive simulations and in EHR data from the VCCC. We illustrated real-world implementation of the optimal designs in CCASAnet.

Now since we know who best to audit, the question became "how should we audit?" The typical data audit procedure first used in clinical trials (Weiss, 1998) and later adopted to observational studies like Duda et al. (2012), Kiragga et al. (2011), and Mphatswe et al. (2012), involved on-site audits conducted by external auditors. While improvements to this protocol were made through the development of more advanced audit capture tools (e.g., a REDcap database instead of a paper audit form as in Duda et al. (2012)), sending trained auditors to clinical sites for these "travel-audits" is expensive, especially in a multi-national cohort like CCASAnet. This imposes heavy constraints on how many sites can be visited, how often, and on

the number of patient records and variables that can be reviewed. In Chapter 3, we investigated the efficacy of data audits conducted by clinical staff at the sites (called "self-audits") as alternatives to conventional travel-auditing. We analyzed a sample of doubly-audited data, i.e, records that were reviewed by both self- and travel-auditors, and found similar overall error rates: self- and travel-auditors reported that 93% and 92% of data entries, respectively, were entered correctly in the original database. In addition, the auditors agreed on 94% of point-by-point audit findings. Therefore, data audits conducted by trained local investigators could provide a cost-effective alternative to on-site audits by external auditors to ensure continued data quality. Our findings led us to consider the proposed self-audit procedure for future projects; it is particularly useful as plans are underway for audits in 2021, in the midst off the COVID-19 pandemic, when travel is especially difficult.

Lastly, with the error-prone and error-free data collected we ask ourselves "how should we incorporate this audit data in analyses?" After great care was taken in the planning and execution of the audit, a statistical method was needed to analyze the data with the same level of precision. Thus, in Chapter 4 we proposed a new full-likelihood approach, the sieve maximum likelihood estimator (SMLE), for logistic regression under complex error settings. The SMLE has desirable statistical properties - namely that it is consistent, asymptotically efficient, and asymptotically normal - and it handles outcome misclassification with continuous covariate error, a setting not yet addressed by other likelihood approaches in the literature (Tang et al., 2015). The utility of this approach was demonstrated through extensive simulations and in data from CCASAnet. With a binary misclassified covariate, the added robustness of the SMLE came at little cost in its efficiency relative to the MLE, and with continuous covariate error there were clear efficiency gains for the SMLE over existing approaches. Finally, the SMLE allows selection of Phase II to depend on the Phase I data in any way, so it can be paired with our optimal design from Chapter 2 for even greater efficiency gains.

All computation in this dissertation was carried out in R Statistical Software (R Core Team, 2019). The *auditDesignR* software accompanies Chapter 2, including an R package and Shiny application. The *auditDesignR* R package (available on GitHub at https://github.com/sarahlotspeich/auditDesignR) contains functions to compute the MLE and optimal design, in addition to analysis code from the chapter. The *auditDesignR* Shiny application is accessible at https://sarahlotspeich.shinyapps.io/auditDesignR/ and can be used to find the optimal design, as well. The R package *logreg2ph* that implemented the SMLE in Chapter 4, along with all simulation and

analysis code, is available on GitHub (https://github.com/sarahlotspeich/logreg2ph).

We focused on binary outcomes in Chapters 2 and 4, but these methods could be extended to other types of response data. First, future work could adapt the optimal designs for other generalized linear models; the objective function for the adaptive grid search would need to be altered accordingly, but the general plan would be similar. Second, future research could develop semiparametric SMLE approaches to other types of outcomes; one was proposed in Tao et al. (2021) for continuous error-prone outcomes, but none have been derived for Poisson or Cox regression.

Secondary use databases include large, diverse datasets that are convenient for biomedical analyses and often come at little-to-no additional cost to collect. As such, these data are being utilized to answer scientific and clinical questions in a broad range of disciplines. Given the error-prone nature of routinely collected data, it is important that these data are analyzed responsibly to avoid biased inference. Two-phase designs can be employed to ensure the quality of observational data for healthcare research, but they can be expensive initiatives. In this dissertation, we proposed new methods to select the most informative records for validation, conduct cost-effective on-site audits, and incorporate audit data into statistical analyses to maximize the investment in two-phase studies.

REFERENCES

Amorim, G., Tao, R., Lotspeich, S., Shaw, P. A., Lumley, T. and Shepherd, B. E. (2021), Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome, *Journal of the Royal Statistical Society, Series A* **in press**, xxx.

Barron, B. A. (1977), The effects of misclassification on the estimation of relative risk, *Biometrics* **33**(2), 414–418.

Beesley, L. J. and Mukherjee, B. (2020), Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification, *Biometrics* **n/a**(n/a), 1–13.

Breslow, N. E. and Cain, K. C. (1988), Logistic regression for two-stage case-control data, *Biometrika* **75**(1), 11–20.
URL: *http://www.jstor.org/stable/2336429*

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: a Modern Perspective*, Boca Raton: Chapman and Hall/CRC.

Chaulagai, C. N., Moyo, C. M., Koot, J., Moyo, H. B., Sambakunsi, T. C., Khunga, F. M. and Naphini, P. D. (2005), Design and implementation of a health management information system in Malawi: issues, innovations and results, *Health Policy and Planning* **20**, 375–384.

Chen, T. and Lumley, T. (2020), Optimal multiwave sampling for regression modeling in two-phase designs, *Statistics in Medicine* **39**(30), 4912–4921.

Chen, Y., Wang, J., Chubak, J. and Hubbard, R. A. (2019), Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence, *Pharmacoepidemiology and Drug Safety* **28**(2), 264–268.

Copeland, K. T., Checkoway, H., McMichael, A. J. and Holbrook, R. H. (1977), Bias due to misclassification in the estimation of relative risk, *American Journal of Epidemiology* **105**, 488–495.

Crabtree-Ramirez, B. E., Jenkins, C., Jayathilake, K., Carriquiry, G., Veloso, V. G., Padgett, D., Gotuzzo, E., Cortes, C., Mejia, F., McGowan, C. C., Duda, S., Shepherd, B. E. and Sterling, T. (2019), HIV-related tuberculosis: mortality risk in persons without vs. culture-confirmed disease, *International Journal of Tuberculosis and Lung Disease* **23**, 306–314.

De, S. (2011), Hybrid approaches to clinical trial monitoring: Practical alternatives to 100% source data verification, *Perspectives in Clinical Research* **2**(3), 100–104.

Dempster, A., Laird, N. and Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 1–38.

Deville, J. C., Särndal, C. E. and Sautory, O. (1993), Generalized raking procedures in survey sampling, *Journal of the American Statistical Association* **88**(423), 1013–1020.

Duda, S. N., Shepherd, B. E., Gadd, C. S., Masys, D. R. and C., M. C. (2012), Measuring the quality of observational study data in an international HIV research network, *PloS one* **7**(4), e33908.

Duda, S. N., Wehbe, F. H. and Gadd, C. S. (2011), Desiderata for a computer-assisted audit tool for clinical data source verification audits., *Studies in health technology and informatics* **160**(Pt 2), 894–898.

Edwards, J. K., Cole, S. R., Troester, M. A. and Richardson, D. B. (2013), Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data, *American Journal of Epidemiology* **177**(9), 904–912.

Funning, S., Grahnen, A., Eriksson, K. and Kettis-Linblad, A. (2009), Quality assurance within the scope of Good Clinical Practice (GCP)—what is the cost of GCP-related activities? A survey within the Swedish Association of the Pharmaceutical Industry (LIF)'s members., *Quality Assurance Journal* **12**, 3–7.

Giganti, M. J., Shaw, P. A., Chen, G., Bebawy, S. S., Turner, M. M., Sterling, T. R. and Shepherd, B. E. (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* **14**, 1045–1061.

Giganti, M. J., Shepherd, B. E., Caro-Vega, Y., Luz, P. M., Rebeiro, P. F., Maia, M., Julmiste, G., Cortes, C., McGowan, C. C. and Duda, S. N. (2019), The impact

of data quality and source data verification on epidemiologic inference: a practical application using HIV observational data, *BMC Public Health* **19**(1), 1748.

Green, S. (2013), Congruence of Disposition After Emergency Department Intubation in the National Hospital Ambulatory Medical Care Survey, *Annals of Emergency Medicine* **61**(4), 423–426.e8.

Grenander, U. (1981), *Abstract Inference*, New York: Wiley.

Han, K., Lumley, T., Shepherd, B. E. and Shaw, P. A. (2020), Two-phase analysis and study design for survival models with error-prone exposures, *Statistical Methods in Medical Research* **0**(0), 0962280220978500.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N. and Conde, J. G. (2009), Research Electronic Data Capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support., *Journal of Biomedical Informatics* **42**(2), 377–381.

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R., Bernstam, E. V., Lehmann, H. P., Hripcsak, G., Hartzog, T. H., Cimino, J. J. and Saltz, J. H. (2013), Caveats for the use of operational electronic health record data in comparative effectiveness research, *Medical care* **51**(8 Suppl 3), S30–S37.

HICDEP 1.110. (2017).
URL: *http://www.hicdep.org/wiki/Hicdep_1.110*

Holcroft, C. A. and Spiegelman, D. (1999), Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified, *Biometrics* **55**, 1193–1201.

Horvitz, D. G. and Thompson, J. D. (1952), A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**(260), 663–685.

Houston, L., Probst, Y. and Martin, A. (2018), Assessing data quality and the variability of source data verification auditing methods in clinical research settings, *Journal of Biomedical Informatics* **83**, 25–32.

IeDEA - Data Exchange Standard (DES) (2017).
URL: *http://iedeades.org*

Keogh, R. H., Shaw, P. A., Gustafson, P., Carroll, R. J., Deffner, V., Dodd, K. W., Küchenhoff, H., Tooze, J. A., Wallace, M. P., Kipnis, V. and Freedman, L. S. (2020), STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment, *Statistics in Medicine* **39**(16), 2197–2231.

Kim, E., Rubinstein, S. M., Nead, K. T., Wojcieszynski, A. P., Gabriel, P. E. and Warner, J. L. (2019), The evolving use of electronic health records (ehr) for research, *Seminars in Radiation Oncology* **29**(4), 354–361. Big Data in Radiation Oncology.

Kimaro, H. C. and Twaakyondo, H. M. (2005), Analysing the hindrance to the use of information and technology for improving efficiency of health care delivery system in Tanzania, *Tanzania Health Research Bulletin* **7**, 189–197.

Kiragga, A. N., Castelnuovo, B., Schaefer, P., Muwonge, T. and Easterbrook, P. J. (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**(3).

Koehler, E., Brown, E. and Haneuse, J. P. A. (2009), On the assessment of monte carlo error in simulation-based statistical analyses, *The American Statistician* **63**(2), 166–162.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc.

Lotspeich, S. C., Giganti, M. J., Maia, M., Vieira, R., Machado, D. M., Succi, R. C., Ribeiro, S., Pereira, M. S., Rodriguez, M. F., Julmiste, G., Luque, M. T., Caro-Vega, Y., Mejia, F., Shepherd, B. E., McGowan, C. C. and Duda, S. N. (2020), Self-audits as alternatives to travel-audits for improving data quality in the Caribbean, Central and South America network for HIV epidemiology, *Journal of Clinical and Translational Science* **4**(2), 125–132.

Lotspeich, S. C., Shepherd, B. E., Amorim, G. G. C., Shaw, P. A. and Tao, R. (2021), Efficient odds ratio estimation using error-prone data from a multi-national HIV research cohort, *Biometrics* **under review**, xxx.

Lumley, T. (2019), 'survey: analysis of complex survey samples'. R package version 3.35-1.

Lumley, T., Shaw, P. A. and Dai, J. Y. (2011), Connections between survey calibration estimators and semiparametric models for incomplete data, *International Statistical Review* **79**(2), 200–220.

Magder, L. S. and Hughes, J. P. (1997), Logistic regression when the outcome is measured with uncertainty, *American Journal of Epidemiology* **146**(2), 195–203.

Marshall, R. J. (1990), Validation study methods for estimating exposure proportions and odds ratios with misclassified data, *Journal of Clinical Epidemiology* **43**(9), 941–947.

McGowan, C. C., Cahn, P., Gotuzzo, E., Padgett, D., Pape, J. W., Wolff, M., Schechter, M. and Masys, D. R. (2007), Cohort Profile: Caribbean, Central and South America Network for HIV research (CCASAnet) collaboration within the International Epidemiologic Databases to Evaluate AIDS (IeDEA) programme, *International Journal of Epidemiology* **36**(5), 969–976.

McIsaac, M. A. and Cook, R. J. (2014), Response-dependent two-phase sampling designs for biomarker studies, *The Canadian Journal of Statistics* **42 (2)**, 268–284.

McIsaac, M. A. and Cook, R. J. (2015), Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis, *Statistics in Medicine* **34**(21), 2899–2912.

Mphatswe, W., Mate, K. S., Bennett, B., Ngidi, H., Reddy, J., Barker, P. M. and Rollins, N. (2012), Improving public health information: a data quality intervention in KwaZulu-Natal, South Africa, *Bulletin of the World Health Organization* **90**(3), 176–182.

Murphy, S. and Van der Vaart, A. (2000), On profile likelihood, *Journal of the American Statistical Association* **95**(450), 449–465.

Neuhaus, J. M. (1999), Bias and efficiency loss due to misclassified responses in binary regression, *Biometrika* **86**(4), 843–855.

Nordo, A. H., Levaux, H. P., Becnel, L. B., Galvez, J., Rao, P., Stem, K., Prakash, E. and Kush, R. D. (2019), Use of EHRs data for clinical research: historical progress and current applications, *Learning Health Systems* **3**(1), e10076.

Oh, E. J., Shepherd, B. E., Lumley, T. and Shaw, P. A. (2018), Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with

simex, *Statistics in Medicine* **37**(8), 1276–1289.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7554*

Prentice, R. L. (1982), Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**(2), 331–342.

Quake, D., Lachenbruch, P. A., Whaley, F. S., McClish, D. K. and Haley, R. W. (1980), Effects of misclassifications on statistical inferences in epidemiology, *American Journal of Epidemiology* **111**, 503–515.

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/*

Reilly, M. and Pepe, M. S. (1995), A mean score method for missing and auxiliary covariate data in regression models, *Biometrika* **82**(2), 299–314.
URL: *http://www.jstor.org/stable/2337409*

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994), Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association* **89**(427), 846–866.

Rudin, W. (1973), *Functional Analysis*, New York: McGraw-Hill.

Safran, C., Bloomsrosen, M., Hammond, E., Labkoff, S., Markel-Fox, S., Tang, P. C. and Detmer, D. E. (2007), Toward a national framework for the secondary use of health data: An American Medical Informatics Association White Paper, *Journal of the American Medical Informatics Association* **14**(1), 1–9.

Sarkar, M., Mallick, B. K. and Carroll, R. J. (2014), Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors, *Biometrics* **70**(4), 823–834.

Schumaker, L. (1981), *Spline Functions: Basic Theory*, New York: Wiley-Interscience.

Shepherd, B. E. and Yu, C. (2011), Accounting for data errors discovered from an audit in multiple linear regression, *Biometrics* **67**, 1083–1091.

Sinnott, J., Dai, W., Liao, K., Shaw, S., Ananthakrishnan, A., Gainer, V., Karlson, E., Churchill, S., Szolovits, P., Murphy, S., Kohane, I., Plenge, R. and Cai, T.

(2014), Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records, *Human Genetics* **133**(11), 1369–1382.

Staudenmayer, J., Ruppert, D. and Buonaccorsi, J. P. (2008), Density estimation in the presence of heteroscedastic measurement error, *Journal of the American Statistical Association* **103**(482), 726–736.

Tang, L., Lyles, R. H., King, C. C., Celentano, D. D. and Lo, Y. (2015), Binary regression with differentially misclassified response and exposure variables, *Statistics in Medicine* **34**(9), 1605–1620.

Tang, L., Lyles, R. H., Ye, Y., Lo, Y. and King, C. C. (2013), Extended matrix and inverse matrix methods utilizing internal validation data when both disease and exposure status are misclassified, *Epidemiologic Methods* **2**(1), 49–66.

Tannen, R. L., Weiner, M. G. and Xie, D. (2009), Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings, *BMJ* **338**.

Tao, R., Lotspeich, S. C., Amorim, G., Shaw, P. A. and Shepherd, B. E. (2021), Efficient semiparametric inference for two-phase studies with outcome and covariate measurement errors, *Statistics in Medicine* **40**(3), 725–738.

Tao, R., Zeng, D. and Lin, D. Y. (2017), Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies, *Journal of the American Statistical Association* **112**, 1468–1476.

Tao, R., Zeng, D. and Lin, D. Y. (2020), Optimal designs of two-phase studies, *Journal of the American Statistical Association* **115**(532), 1946–1959.

Tosteston, T. D. and Ware, J. H. (1990), Designing a logistic regression study using surrogate measures for exposure and outcome, *Biometrika* **77**(1), 11–21.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Wei, W.-Q. and Denny, J. C. (2015), Extracting research-quality phenotypes from electronic health records to support precision medicine, *Genome Medicine* **7**(1), 41.

Weiss, R. (1998), Systems of protocol review, quality assurance, and data audit., *Cancer Chemotherapy and Pharmacology* **42**(Suppl), S88–S92.

White, J. E. (1982), A two stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**(1), 119–128.

Zaniewski, E., Tymejczyk, O., Kariminia, A., Desmonde, S., Leroy, V., Ford, N., Sohn, A. H., Nash, D., Yotebieng, M., Cornell, M., Althoff, K. N., Rebeiro, P. F. and Egger, M. (2018), IeDEA-WHO Research-Policy Collaboration: contributing real-world evidence to HIV progress reporting and guideline development, *Journal of Virus Eradication* **4**(Suppl 2), 9–15.

Zeng, D. (2005), Likelihood approach for marginal proportional hazards regression in the presence of dependent censoring, *Annals of Statistics* **33**, 501–521.