Design and Analysis Considerations for Complex Longitudinal and Survey Sampling Studies

By

Nathaniel David Mercaldo

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

September 30, 2017

Nashville, Tennessee

Approved:

Frank E. Harrell Jr., Ph.D.

Jonathan S. Schildcrout, Ph.D.

Bryan E. Shepherd, Ph.D.

Matthew S. Shotwell, Ph.D.

Loren P. Lipworth, Sc.D.

To my girls, Sarah and Joey
and
To my parents, Ann and David

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the support of a few very special people.

I would like to express my deepest gratitude to my advisor, Dr. Jonathan Schildcrout, for his guidance, patience, and friendship throughout the years. Thank you for always making time to discuss research ideas, and constantly pushing me to becoming a better biostatistician. I would like to thank my dissertation committee members, Dr. Frank Harrell, Dr. Loren Lipworth, Dr. Bryan Shepherd, and Dr. Matthew Shotwell, for your time, assistance, and your invaluable insights throughout this dissertation process.

I would also like thank all the faculty, staff, and graduate students in the Department of Biostatistics at Vanderbilt University. I would especially like to thank Dr. Jeffrey Blume for his constant encouragement, and his efforts in creating such a positive environment for all graduate students.

I would like to thank my family. To my wife Sarah, I could not be here without your love and support, and I am so blessed that I can share this experience with you. To my parents, Ann and David, thank you for making me guffaw like only you know how to do and for your endless advice and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The United States health care system has become more reliant on health information technology and active data collection due in part to the Health Information Technology for Economic and Clinical Health Act of 2009. This Act provides financial incentives to institutions that are implementing and promoting the "meaningful use" of electronic health record (EHR) data. As the amount of EHR data proliferates, nationwide efforts (e.g., Project HealthDesign) have been initiated to generate novel secondary uses of EHR data to improve public health, such as combining EHR data with biorepositories to understand complex genotype and phenotype relationships. The need for efficient study designs is paramount due to resource constraints (e.g., financial, limited biospecimens). This dissertation consists of three chapters relating to the design and analysis of longitudinal and survey sampling studies when utilizing EHR data and biorepository data.

In Chapter two, we extend the class of outcome dependent sampling (ODS) designs for longitudinal binary data. These retrospective study designs are implemented when it is not feasible to collect an expensive exposure on an entire cohort. One subclass of ODS designs stratifies individuals into three possible strata according to the categorization of their response vector: those who did not experience the outcome, those that only experienced the outcome, and those that exhibited response variation. For time-varying covariate effects, it has been shown that sampling only those individuals with response variation results in nearly fully efficient estimation. If inference lies in a time-invariant effect, or a combined time-varying and time-invariant effect, then the choice of how to allocate resources is not obvious. We propose a class of two-stage ODS designs where data from both stages are collected using ODS designs. Two distinct sub-classes of these designs are explored. First, we extend standard (or single-stage) ODS designs to permit two waves of data collection using pre-specified sampling probabilities. Second, adaptive two-stage ODS designs are described whereby information from stage one is used to identify the "conditionally optimal" stage two design. These designs are applied to data from the Lung Health Study where it is of interest to identify genetic determinants of lung function decline among individuals with mild chronic obstructive pulmonary disease.

In Chapter three, we investigate the effects of utilizing an imperfect sampling frame on the design and analysis of complex survey data. This study is motivated by

a large multi-center survey developed to elicit perspectives on biobank participation among under-studied subgroups (e.g., racial and ethnic monitories). A disproportionate stratified sampling scheme is implemented to enrich the sample population with these less prevalent populations using a sampling frame primarily constructed from EHR data. Incomplete EHR data is imputed using geocode-derived census summaries which resulted in a well-defined, but imperfect sampling frame. Chapter five in the dissertation provides additional details of the construction of the sampling frame, and the targeted sampling approach that maximizes the entropy of the stratification information in the final sample. We provide analytic calculations of the expectation and variance of the design-based estimators of the mean and total under stratum misclassification. We explore the effects of stratum misclassification in a real-world example by analyzing a subset of the biobank survey data from Vanderbilt University Medical Center.

In Chapter four, the `MMLB` package is introduced and examples are provided to demonstrate how to estimate parameters from marginalized regression models for longitudinal binary data. Estimation of model parameters is described when data are collected prospectively under random sampling, and under a class of ODS designs. Using data from the Madras Longitudinal Schizophrenia Study, we demonstrate how this package can be used to fit three types of marginalized regression models, including: the marginalized latent variable model, the marginalized transition model, and the marginalized latent variable and transition model. Examples are provided to show how `MMLB` functions may be used to generate longitudinal binary outcomes under a pre-specified marginal model, and to demonstrate how it is used to estimate marginalized model parameters under single- and two-stage ODS sampling designs.

CHAPTER 2

TWO-STAGE FIXED & ADAPTIVE OUTCOME-DEPENDENT SAMPLING DESIGNS FOR LONGITUDINAL BINARY DATA

2.1 Abstract

Retrospective outcome dependent sampling (ODS) designs are an efficient class of study designs that may be implemented when resource constraints prohibit the ascertainment of an exposure on an entire cohort. One type of ODS design for longitudinal binary data stratifies individuals into three strata according to a categorization of their response vector: those who did not experience the outcome, those that only experienced the outcome, and those that exhibited response variation (Schildcrout and Heagerty, 2008). For time-varying covariate effects, it has been shown that sampling only those individuals with response variation results in nearly fully efficient estimation compared to the full cohort analysis. If inference lies in a time-invariant covariate effect, or a combined time-varying and time-invariant covariate effect, then the choice of how to allocate resources, or how to define sampling probabilities, is not obvious. We propose a class of two-stage ODS designs for longitudinal binary data. We extend standard (or single-stage) ODS designs to permit two waves of data collection. Fixed two-stage ODS designs utilize pre-specified sampling probabilities, and adaptive two-stage ODS designs use information from stage one to inform our choice of the stage two sampling probabilities. These designs are applied to data from the Lung Health Study where it is of interest to identify genetic determinants of lung function decline among individuals with mild chronic obstructive pulmonary disease.

2.2 Introduction

Cost-effective study designs in the health sciences have, and currently remains, an important research area (Zhou et al., 2013). The use of retrospective study designs to investigate novel scientific questions are becoming more common due to the proliferation of existing cohort data (e.g., electronic health records, biobanks). When exposure of interest is unavailable, it may not be financially possible or scientifically ethical (e.g., patient burden, limited biospecimens) to collect this information on an entire cohort (Schildcrout and Heagerty, 2008, 2011). If the outcome of interest is rare, then a study design which targets sampling to those individuals who are most informative is necessary to efficiently study the disease-outcome relationship (Schildcrout

and Heagerty, 2008, 2011; Schildcrout et al., 2015; Zhou et al., 2013). An outcome-dependent sampling (ODS) design is a type of retrospective study design whereby the probability of inclusion depends on the subject's outcome value(s). In the univariate setting, the most common ODS design is the case-control study. Standard ODS designs for correlated binary data define sampling strata using a summary of an individuals?? response vector (e.g., sum, or a categorization of the sum) (Neuhaus and Jewell, 1990; Schildcrout and Heagerty, 2008, 2011). Similarly, ODS designs for univariate and longitudinal continuous outcomes are typically defined by categorizing either the outcome or a low-dimensional subject-specific summary (e.g., intercept, slope, dfbeta) (Zhou et al., 2007; Schildcrout et al., 2013).

In the longitudinal binary data setting, the choice of ODS design depends on the inferential target. For example, consider the ODS design of Schildcrout and Heagerty (2011) where individuals are stratified into one of three possible groupings: those who did not experience the outcome (binary response vector only 0s), those that only experienced the outcome (binary response vector only 1s), and those that exhibited response variation. When the scientific question pertains only to a time-varying covariate effect, then it has been shown the most efficient ODS designs are those that devote resources to individuals exhibiting response variation (Schildcrout and Heagerty, 2008). When interest lies in a time-invariant covariate effect, or the joint effect between a time-varying and time-invariant covariate, then the choice of how to optimally define the triplet of sampling probabilities is not straightforward. Schildcrout and Heagerty (2011) outline one approach to prospectively compare candidate ODS designs. The authors assume that all information except the key exposure data is available on the entire cohort. Missing exposure data is imputed using the observed data, and positing assumptions regarding the exposure prevalence and the exposure-outcome relationship. These complete data sets may then be used to evaluate, and compare, different ODS designs. The results of these simulation-based comparative design analyses are sensitive to the assumptions regarding the prevalence of the exposure and thus "misspecification could potentially lead to overestimates or underestimates of design precision" (Schildcrout and Heagerty, 2011). One approach to reduce the dependencies on these assumptions is to collect data in more than one stage, and use information from the first stage (or internal pilot) to inform the decision making process at later stages.

Multi-stage study designs, such as the two-stage case-control study, may also be implemented to efficiently characterize the exposure-outcome relationship (Zhao and Lipsitz, 1992; Zhou et al., 2013). Two-stage designs typically collect outcome and

auxiliary exposure data at stage one. Using the cross-classification of these data, sampling strata are defined to identify the most informative subjects to sample at stage two. The expensive exposure/confounder is then only collected on the stage two subsample. Zhou and colleagues have extended ODS designs for continuous outcome data setting to permit data collection in two stages where an auxiliary variable is collected in stage one via random- or outcome-dependent sampling, and the exposure of interest is collected on a subsample during stage two (Song et al., 2009; Zhou et al., 2010; Xu and Zhou, 2012; Zhou et al., 2013). To date, the extension to longitudinal binary has yet to be performed.

We propose a class of two-stage ODS designs for longitudinal binary data where data from both stages are collected using ODS designs and we explore two distinct sub-classes of these designs. First, we extend standard (or single-stage) ODS designs to permit two waves of data collection using pre-specified sampling probabilities. Second, adaptive two-stage ODS designs are described whereby information from stage one is used to identify the stage two design. These designs are more flexible than their single- or two-stage fixed design counterparts by permitting mid-study modifications of the stage two sample size or sampling probabilities. The two-stage fixed design/adaptive sample size ODS design identifies the stage two sample size needed to attain a pre-specified level of precision for a time-varying covariate effect when implementing an extreme ODS design. The two-stage adaptive design/fixed sample size ODS design identifies the design that maximizes an optimality criterion (e.g., precision, determinant of the information matrix) for a pre-specified overall stage two sample size.

In Section 2, we describe a class of the marginalized regression models that are used to demonstrate our novel two-stage designs for longitudinal binary data. The ODS designs described are agnostic to the choice of model, but we chose these models since they permit population-average interpretations of all covariate effects, they separate the specification of the mean and dependence model, and provide the flexibility to model a wide range of dependence structures that are typically encountered in the health sciences. In Sections 3 and 4, we define the fixed two-stage ODS design, and the adaptive two-stage ODS designs, respectively. We apply these methods to data from the Lung Health Study in Section 5. The discussion and future directions are presented in Section 6.

2.3 Model

Marginalized regression models are defined by a pair of regression models that, with assumptions regarding random-effect distributions, fully specify the multivariate distribution of binary outcome vector given the observed design matrix (Heagerty, 1999, 2002; Schildcrout and Heagerty, 2007). First, a marginal mean model is constructed to relate covariates to the logit-transformed binary outcome. To capture second and higher order moments, a dependence model is defined to characterize serial and/or long-range dependence structures. The marginalized transition and latent-variable model (mTLV) allows the specification of both types of dependence structures simultaneously (Schildcrout and Heagerty, 2007).

To define the mTLV model, let Y_{ij} denote the binary outcome of subject i at observation j where $i = \{1, 2, \ldots, N\}$ and $j = \{1, 2, \ldots, n_i\}$. Let $\boldsymbol{X}_i$ denote a $n_i \times p$ design matrix, $\boldsymbol{X}_{ij}$ the corresponding p-dimensional design vector at time j and $\boldsymbol{\beta}^m$ the p-dimensional vector of parameters. Then, the marginal mean and dependence models are defined as:

$$\text{logit}\left(\mu_{ij}^m\right) = \boldsymbol{X}_{ij}\boldsymbol{\beta}^m \tag{2.1}$$

$$\text{logit}\left(\mu_{ij}^c\right) = \Delta_{ij} + \gamma(\boldsymbol{X}_i)Y_{ij-1} + b_i \quad \text{where } b_i \sim N\left(0, \sigma^2(\boldsymbol{X}_i)\right) \tag{2.2}$$

Δ_{ij} is the value that relates the marginal and conditional means via the convolution equation

$$\mu_{ij}^m = \int_{\boldsymbol{A}_{ij}} \mu_{ij}^c dF_{\boldsymbol{A}_{ij}} = \int_{\boldsymbol{A}_{ij}} \text{logit}^{-1}\left(\Delta_{ij} + \boldsymbol{A}_{ij}\boldsymbol{\alpha}\right) dF_{\boldsymbol{A}_{ij}} \tag{2.3}$$

where $\boldsymbol{A}_{ij}$ and $\boldsymbol{\alpha}$ denote the design matrix of the dependence model and the stacked parameter vector $(\gamma(\boldsymbol{X}_i), \sigma(\boldsymbol{X}_i))$, respectively. For the remainder of this paper, we assume that the dependence model parameters are not modified by covariates. This implies that $\gamma(\boldsymbol{X}_i) = \gamma$ and $b_i \sim N(0, \sigma^2)$ which can be rewritten as σZ_i where $Z_i \sim N(0, 1)$.

Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}^m, \boldsymbol{\alpha}\}$, then under random sampling, subject i's contribution to the likelihood function is defined as:

$$L_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i) = pr(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}) = \int_{z_i} \left[\prod_{j=1}^{n_i} \mu_{ij}^{c\ y_{ij}} \left(1 - \mu_{ij}^{c\ y_{ij}}\right)^{1-y_{ij}}\right] \phi(z_i)dz_i \equiv \int_{z_i} L_{i,z_i}\phi(z_i)dz_i. \tag{2.4}$$

where ϕ denotes the standard normal distribution.

2.4 Two-Stage Fixed Outcome Dependent Sampling Designs

Outcome-dependent sampling (ODS) is a class of retrospective stratified sampling schemes that are designed to enrich a sample with individuals that are the most informative (Schildcrout and Heagerty, 2008; Zhou et al., 2013; Song et al., 2009). The most notable ODS design for univariate binary response data is the case-control study whereby an individual's sampling probability is dependent on their response status (Anderson, 1972; Prentice and Pyke, 1979). Neuhaus and Jewell (1990) extended the case-control design to accommodate correlated binary response data by defining sampling strata as the sum of an individual's response vector, and modeling the exposure-response relationship using a random-intercept logistic model. Schildcrout and Heagerty (2011) propose a class of ODS study designs based on the coarse categorization of an individual's response vector, and describe an approach to estimate mTLV model parameters under these designs using the ascertainment-corrected likelihood.

Our proposed fixed two-stage ODS designs extend the ODS designs of Schildcrout and Heagerty (2011) by allowing data to be collected in two waves. Next, we review the single-stage ODS design, and then describe modifications associated with stage two ODS sampling probabilities. Finally, we describe how data from both stages are combined to estimate mTLV model parameters.

2.4.1 Fixed Stage One ODS Design

We consider the class of ODS designs where each individual is stratified into one of three groups based on their response vector. Let $V_i = g(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ denote the stratum membership for subject i. While V_i can depend on both Y_i and X_i, we define three strata with: those that did not experience the event of interest (non-responders, $\sum_j Y_{ij} = 0$; $V_i = 0$), those that exhibited response variation (any-responders, $0 < \sum_j Y_{ij} < n_i$; $V_i = 1$), and those that only experienced the event (all-responders, $\sum_j Y_{ij} = n_i$; $V_i = 2$). Let N_v denote the stratum sample sizes in the full cohort for sampling strata $v = (0, 1, 2)$, and let S_{1i} represent the indicator if subject i is sampled during stage one. Since S_{1i} is conditionally independent of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ given V_i (i.e., S_{1i} depends on (Y_i, X_i) only through the coarsened response sum), then the stratum-specific sampling probabilities are defined as:

$$pr(S_{1i} = 1 \mid \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i}) = pr(S_{1i} = 1 \mid V_i = v) \equiv \pi_1(v) \tag{2.5}$$

The stage one ODS design is defined as either a pre-specified triplet of sampling probabilities, $\pi_1(0)$,$\pi_1(1)$, and $\pi_1(2)$, or as the sampling probabilities that are induced from a pre-specified vector of expected stratum sample sizes n_{10}, n_{11} and n_{12}. For ease of interpretation, we define the stage one ODS design as $D_1[n_{10}, n_{11}, n_{12}]$ where D_1 denotes the design for stage one, and n_{1v} represents the expected stage one sample size for stratum v.

Via Bayes' theorem, subject i's contribution to the conditional, or ascertainment-corrected, stage one likelihood is

$$\begin{aligned} L^c_{1i}(\boldsymbol{\theta} \mid \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i}, S_{1i} = 1) = pr(\boldsymbol{y}_{1i} \mid \boldsymbol{x}_{1i}, S_{1i} = 1; \boldsymbol{\theta}) &= \frac{pr(S_{1i} = 1 \mid \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i})}{pr(S_{1i} = 1 \mid \boldsymbol{x}_{1i})} \cdot pr(\boldsymbol{y}_{1i} \mid \boldsymbol{x}_{1i}; \boldsymbol{\theta}) \\ &\equiv \frac{\pi_1(v_i)}{AC_{1i}} \cdot L_{1i}(\boldsymbol{\theta} \mid \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i}) \end{aligned} \tag{2.6}$$

where AC_{1i} denotes the ascertainment correction, and $L_{1i}(\boldsymbol{\theta} \mid \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i})$ is the unconditional likelihood contribution defined in Equation 4.4. The ascertainment correction is defined as:

$$\begin{aligned} AC_{1i} &= \sum_{v=0}^{2} pr(S_{1i} = 1, V_i = v \mid \boldsymbol{x}_{1i}) = \sum_{v=0}^{2} \pi_1(v) \cdot pr(V_i = v | \boldsymbol{x}_{1i}) \\ &= \pi_1(1) + [\pi_1(0) - \pi_1(1)]\, pr(V_i = 0 \mid \boldsymbol{x}_{1i}) + [\pi_1(2) - \pi_1(1)]\, pr(V_i = 2 \mid \boldsymbol{x}_{1i}) \end{aligned} \tag{2.7}$$

Note, $pr(V_i = v \mid \boldsymbol{x}_{1i})$ corresponds to the likelihood contribution of subject i when their response vector is either all 0s (v=0) or 1s (v=2), respectively.

2.4.2 Fixed Stage Two ODS Design

Let $\boldsymbol{q}_1 = \{\boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = \boldsymbol{s}_1\}$ denote the response vector, design matrix and the sampling indicator for the stage one cohort. The design matrix $\boldsymbol{x}_1$ includes the covariate information available at the beginning of the study *and* the exposure data for those sampled at stage one. Since sampling is independent within each stage, the stage two sampling probabilities are defined as

$$\begin{aligned} pr(S_{2i} = 1 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) &\equiv pr(S_{2i} = 1, S_{1i} = 0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \\ &= pr(S_{2i} = 1 \mid S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot pr(S_{1i} = 0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \\ &= \pi_2(v; \boldsymbol{q}_1) \cdot [1 - \pi_1(v)] \end{aligned} \tag{2.8}$$

where $S_{2i}, \boldsymbol{y}_{2i}$, and $\boldsymbol{x}_{2i}$ denote the stage two sampling indicator, response vector, and design matrix for subject i, respectively. Subject i's stage two sampling probability is defined as the product of the probability of being sampled at stage two given not being sampled at stage one, and the probability of not being sampled at stage one. For a fixed sample size, the stage two ODS design denoted as $D_2[n_{20}, n_{21}, n_{22}]$.

Similar to Equations 2.6 and 2.7, subject i's contribution to the stage two conditional likelihood is:

$$L^c_{2i}(\boldsymbol{\theta} \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, S_{2i} = 1) = \frac{\pi_2(v; \boldsymbol{q}_1) \cdot [1 - \pi_1(v)]}{AC_{2i}} \cdot L_{2i}(\boldsymbol{\theta} | \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \tag{2.9}$$

where the stage two ascertainment correction, AC_{2i}, is defined as $\sum_{v=0}^{2} \pi_2(v; \boldsymbol{q}_1)[1 - \pi_1(v)] \cdot pr(V_i = v \mid \boldsymbol{x}_{2i})$, and $L_{2i}(\boldsymbol{\theta} | \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i})$ is defined in Equation 4.4.

2.4.3 Conditional Two-Stage ODS Likelihood

Since $(\boldsymbol{Y}_i, \boldsymbol{X}_i) \perp\!\!\!\perp (\boldsymbol{Y}_k, \boldsymbol{X}_k)$ for all $i \neq k$, the combined two-stage conditional likelihood is defined as the product of individual likelihood contributions from each stage. Without loss of generality, we assume that subject identifiers have been re-ordered such that the first N_1^s subjects correspond to those individuals sampled at stage one, and the remaining $N_2^s \equiv N_2^s(\boldsymbol{q}_1)$ represent individuals sampled at stage two. The combined two-stage conditional likelihood is defined as

$$\begin{aligned} L^c(\boldsymbol{\theta} | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{S} = 1) &= pr(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{S} = 1) = pr(\boldsymbol{y}_1 | \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \cdot pr(\boldsymbol{y}_2 | \boldsymbol{x}_2, \boldsymbol{S}_2 = 1, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \\ &= \left[\prod_{i=1}^{N_1^s} \frac{\pi_1(v)}{AC_{1i}} \cdot L_{1i}(\boldsymbol{\theta} | \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i}) \right] \cdot \left[\prod_{i=N_1^s+1}^{N_1^s+N_2^s} \frac{\pi_2(v; \boldsymbol{q}_1)[1 - \pi_1(v)]}{AC_{2i}} \cdot L_{2i}(\boldsymbol{\theta} | \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \right] \end{aligned} \tag{2.10}$$

2.4.4 Ascertainment-Correct Maximum Likelihood Estimation

Equation 2.10 explicitly accounts for the biased sampling scheme in the definition of the conditional likelihood. To estimate parameters, the corresponding score equation for parameter $\theta \in \boldsymbol{\theta}$ is:

$$\begin{aligned} \frac{\partial}{\partial \theta} \log L^c(\theta | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{S} = 1) = &\left[\sum_{i=1}^{N_1^s} -\frac{1}{AC_{1i}} \frac{\partial}{\partial \theta} AC_{1i} + \frac{\partial}{\partial \theta} \log L_{1i}(\theta | \boldsymbol{y}_{1i}, \boldsymbol{x}_{1i}) \right] + \\ &\left[\sum_{i=N_1^s+1}^{N_1^s+N_2^s} -\frac{1}{AC_{2i}} \frac{\partial}{\partial \theta} AC_{2i} + \frac{\partial}{\partial \theta} \log L_{2i}(\theta | \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \right] \end{aligned} \tag{2.11}$$

where

$$\frac{\partial}{\partial\theta}AC_{1i} = \frac{\partial}{\partial\theta}pr(V_i = 0|\boldsymbol{x}_{1i})[\pi_1(0) - \pi_1(1)] + \frac{\partial}{\partial\theta}pr(V_i = 2|\boldsymbol{x}_{1i})[\pi_1(2) - \pi_1(1)]$$

$$\begin{aligned}\frac{\partial}{\partial\theta}AC_{2i} = &\frac{\partial}{\partial\theta}pr(V_i = 0|\boldsymbol{x}_{2i})[\pi_2(0;\boldsymbol{q}_1)[1 - \pi_1(0)] - \pi_2(1;\boldsymbol{q}_1)[1 - \pi_1(1)]] \\ &+ \frac{\partial}{\partial\theta}pr(V_i = 2|\boldsymbol{x}_{2i})[\pi_2(2;\boldsymbol{q}_1)[1 - \pi_1(2)] - \pi_2(1;\boldsymbol{q}_1)[1 - \pi_1(1)]]\end{aligned}$$

$$\frac{\partial}{\partial\theta}\log L_i(\theta|\boldsymbol{y}_i,\boldsymbol{x}_i) = \left[\int_{z_i} L_{i,z_i}\phi(z_i)dz_i\right]^{-1}\int_{z_i} L_{i,z_i}\left[\sum_{j=1}^{n_i}\left(y_{ij} - \mu_{ij}^c\right)\frac{\partial}{\partial\theta}\left(\Delta_{ij} + \gamma y_{ij-1} + \sigma z_i\right)\right]\phi(z_i)dz_i$$

2.4.5 Simulation

We first investigate the operating characteristics of the proposed two-stage fixed designs by generating data according to the following marginalized transition model:

$$\begin{aligned}\text{logit}\left(\mu_{ij}^m\right) &= \beta_0 + \beta_t t_{ij} + \beta_e X_{ei} + \beta_{et}X_{ei}\cdot t_{ij} \\ \text{logit}\left(\mu_{ij}^c\right) &= \Delta_{ij} + \gamma Y_{ij-1}\end{aligned}$$

where X_{ei} is the binary time-invariant exposure that can be retrospectively collected and $t_{ij} = \{0, 1, 2, 3, 4\}$ for subject i at time j. We assume $pr(X_e = 1) = 0.25$ and $\{\boldsymbol{\beta}, \gamma\} = \{-1.50, -0.25, 1.00, 0.25, 2.00\}$. When a population of 5000 is genereated, these parameters induce sampling strata with expected sizes (2360, 2476, 164) which correspond to no-responders, any-responders, and all-responders. We assume that three inferential targets are of interest: 1) β_{et}, 2) β_e and 3) the joint exposure effect (β_e, β_{et}).

Due to resource constraints, suppose only 500 individuals are sampled. We consider sampling 100 at stage one to mimic an internal pilot study, and permit an additional 400 to be sampled at stage two. One stage one and five stage two ODS designs are considered: $D_1[25, 50, 25]$, and $D_2[0, 400, 0]$, $D_2[25, 350, 25]$, $D_2[50, 300, 50]$, $D_2[75, 250, 75]$ and $D_2[100, 200, 100]$. We examine the operating characteristics of these two-stage ODS fixed designs and compared them to random sampling and to the corresponding single stage ODS designs (e.g., $D_1[25, 50, 25] + D_2[0, 400, 0] = D[25, 450, 25]$).

Maximum likelihood, and ascertainment-corrected maximum likelihood, is used to estimate marginalized model parameters under random and outcome-dependent sampling, respectively. Other approaches to validly estimate model parameters from a biased sample include weighted-estimating equations (Robins et al., 1995; Cai et al.,

2001) and missing data techniques (Schildcrout and Heagerty, 2011; Schildcrout et al., 2015), which are not currently explored.

We generate a population for each of the 1000 replications of the simulation. In so doing we implement random and ODS sampling schemes, as well as estimate model parameters. For all model parameters, percent bias and coverage probabilities are computed. The efficiency associated with each ODS design is compared to random sampling for three optimality criteria: the variance of β_{et}, the variance of β_e, and the d-efficiency of (β_e, β_{et}) which minimizes the confidence region associated with these parameter estimates. The relative efficiency estimates are defined as the average optimality value under random sampling divided by the average value under the single- or two-stage ODS designs.

Table 2.1 summarizes the percent bias and coverage probabilities for two-stage fixed ODS designs considered in this simulation. It is clear that estimation is performing as expected since all parameter estimates are unbiased and attain nominal coverage.

Table 2.1: Percent bias and coverage probabilities for two-stage fixed ODS designs. Rows correspond to the stage one ODS design, and columns correspond to the stage two ODS design. Percent bias is defined as 100 times the difference between the average estimate minus the true parameter value divided by the true value. The coverage probability is defined as the proportion of times the estimated 95% confidence interval contained the true parameter value.

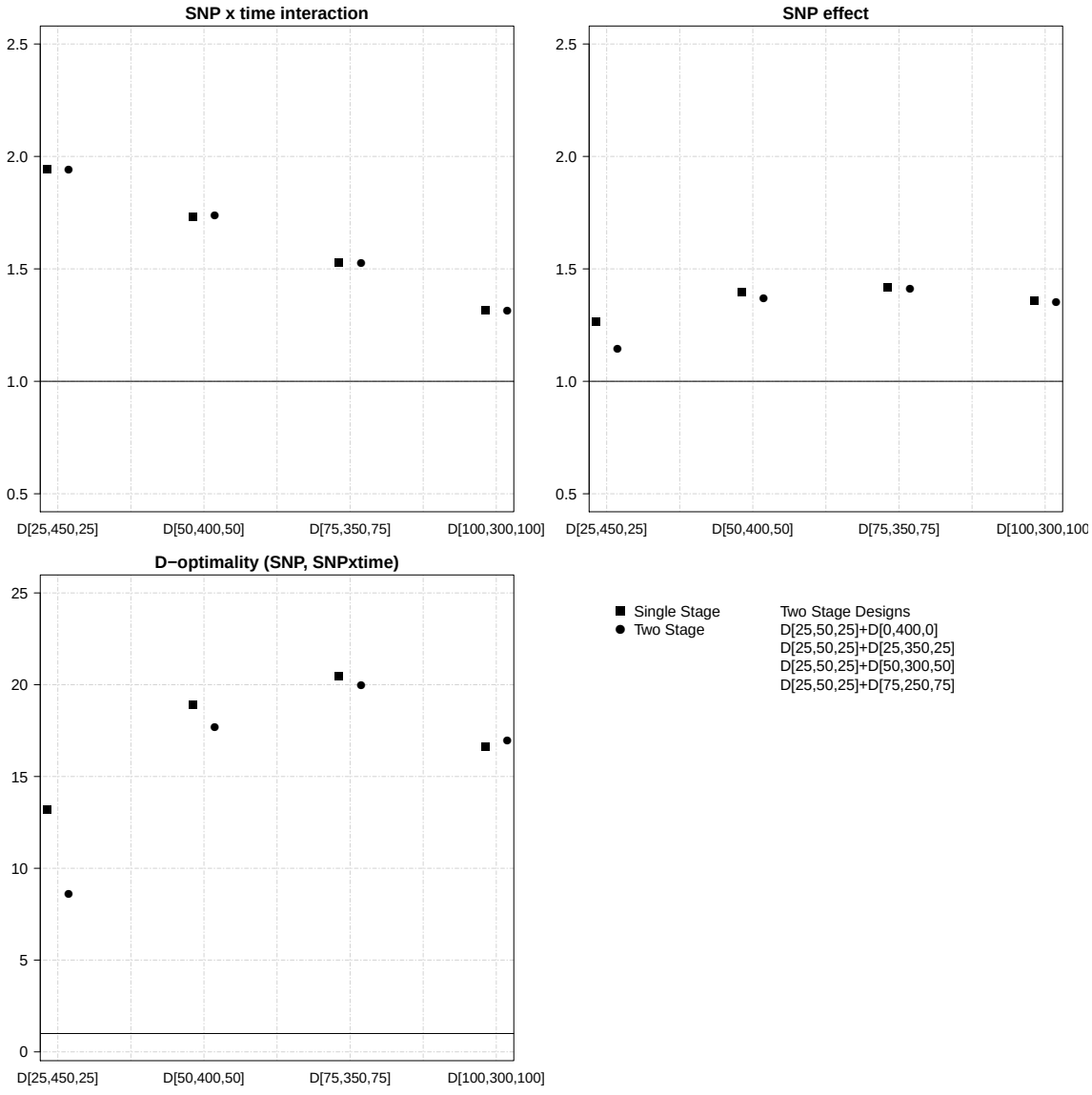
	$\mathbf{D}_2$**[25, 350, 25]**		$\mathbf{D}_2$**[50, 300, 50]**		$\mathbf{D}_2$**[75, 250, 75]**		$\mathbf{D}_2$**[100, 200, 100]**	
	Bias (%)	Coverage	Bias (%)	Coverage	Bias (%)	Coverage	Bias (%)	Coverage
$\mathbf{D}_1$**[25, 50, 25]**								
β_0	0	0.95	0	0.95	0	0.94	0	0.94
β_t	0	0.95	0	0.96	1	0.96	0	0.95
β_e	1	0.94	0	0.95	0	0.94	1	0.96
β_{et}	−1	0.94	0	0.95	1	0.95	−1	0.96
γ	0	0.96	0	0.94	0	0.96	0	0.95

Figure 2.1 summarize the relative efficiencies of the three optimality criteria when utilizing $D_1[25, 50, 25]$, respectively. For each figure, the x-axis represents the corresponding stage one ODS design. For example, $D[25, 450, 25]$ represents the single-stage design, as well as the $D_1[25, 50, 25] + D_2[0, 400, 0]$ two-stage design. For time-varying covariate effects, designs that oversample those with response variation result in more efficient estimates than those that do not (i.e., $D[25, 450, 25]$ vs others).

When estimating time-invariant or joint covariate effects, the choice of best design is not obvious (e.g., that which maximizes an optimality criterion). Depending on characteristics of the data, such as the magnitude of the exposure effect, differences in the efficiency of ODS designs can be significant.Our data generating model posited large effects for both the exposure and the exposure by time interaction, thus sampling those with response variation also enrich the sample with those with the exposure resulting in increased efficiency compared to random sampling (all RE>1). The stage two ODS design that maximizes the precision of the β_e is between $D_2[50, 300, 50]$ and $D_2[75, 250, 75]$. For joint effects, all ODS designs considered resulted in larger d-efficiency values compared to random sampling, and those that sampled between 50-75 no- or all-responders produced the greatest d-efficiency values.

We observe minimal efficiency loss when performing a two-stage fixed ODS design compared to the induced single-stage ODS design when interest lies in the precision of a single covariate effect. Differences are observed when utilizing the d-efficiency criterion due to the definition of this quantity (i.e., a function of the entire covariance matrix, not simply the diagonal elements), the complex relationship between the covariates, and the increased variation in the two-stage sampling weights compared to single-stage weights.

Figure 2.1: Efficiency relative to random sampling using $D_1[25, 50, 25]$.



2.5 Two-Stage Adaptive Outcome Dependent Sampling Designs

We introduce two adaptive procedures when constructing the second stage of a two-stage ODS design. The first utilizes stage one information to aid in determining the stage two sample size needed to estimate a time-varying covariate effect within a pre-specified level of precision. The second identifies the stage two design that maximizes an optimality criterion (e.g., precision, or d-efficiency) for a fixed stage two sample size. These scenarios provide the basis for additional extensions, such as an adaptive-design *and* adaptive-sample size two-stage design.

The key to both adaptive methods is the ability to compare candidate designs which requires the imputation of missing exposure data. We describe two approaches to estimate the marginal conditional exposure model for non-sampled subjects. The first only uses information from the sampled subjects, while the second also utilizes information from the non-sampled subjects. Once a complete data set is created, different designs are then compared to identify the stage two sample size or design.

Combining data from stages one and two proceeds in the same manner as described in Section 2.10. Data from the first stage of an adaptive two-stage ODS design is identical to that of a fixed two-stage ODS design (Section 2.4.1). We describe the estimation of sampling probabilities, and subject-specific likelihood contributions using two adaptive approaches.

2.5.1 Stage Two: Adaptive Sample Size

The defining feature of longitudinal studies is that temporal changes may be investigated directly due to the repeated measurements on study participants (Diggle et al., 2002). Time-varying covariate effects, such as a gene by time interaction (β_{et}), are often used to quantify these changes in the regression modeling framework. When the inferential target is strictly a time-varying covariate, Schildcrout and others have demonstrated that the single-stage ODS designs that restrict sampling to only those subjects with response variation are more efficient than those that do not, and are nearly as efficient as the full-cohort when all these subjects (i.e., those with $V_i = 1$) are sampled (Schildcrout and Heagerty, 2008). Using notation outlined in Section 2.4.1, this extreme sampling design is denoted as $D_1[0, n_{11}, 0]$ where n_{11} is the expected number of sampled subjects from the "any-responders" sampling stratum when using a single-stage ODS design.

The primary drawback to this design is that it requires the pre-specification of n_{11} which may result in the study being either under- or over-powered. n_{11} is deter-

mined either by resource limitations or by performing a sensitivity analysis in which candidate designs (i.e., different values of n_{11}) are compared. A comparative design analysis, one type of sensitivity analysis, is used to impute missing exposure data. This is accomplished by coupling the observed data with assumptions regarding the missing exposure variable (e.g., prevalence, relationship with other covariates). Unfortunately, the choice of design is sensitive to these assumptions (Schildcrout and Heagerty, 2011).

We define the stage two adaptive sample size design as $D_2[0, N_{21}^{\kappa}, 0]$ where N_{21}^{κ} denotes the expected stage two sample size, and κ represented a pre-defined threshold for the $Var(\beta_{et})$. This quantity is estimated using a what we refer to as a comparative design analysis (as described in Section 2.5.3). The comparative design analysis utilizes data from stage one, and modeling assumptions, to impute exposure data for the non-sampled subjects. The estimation of N_{21}^{κ} proceeds by repeatedly evaluating candidate designs using these complete data sets until $Var(\beta_{et}) \approx \kappa$ is obtained.

Since N_{21}^{κ} is now a random variable, the stage two sampling probability is defined as:

$$\begin{aligned} pr(S_{2i} = 1 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) &\equiv pr(S_{2i} = 1, N_{21}^{\kappa} = n_{21}^{\kappa}, S_{1i} = 0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \\ &= pr(S_{2i} = 1, N_{21}^{\kappa} = n_{21}^{\kappa} \mid S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot pr(S_{1i} = 0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \\ &= pr(S_{2i} = 1 \mid N_{21}^{\kappa} = n_{21}^{\kappa}, S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot \\ &\qquad pr(N_{21}^{\kappa} = n_{21}^{\kappa} \mid S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot pr(S_{1i} = 0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \\ &= \pi_2(v; \boldsymbol{q}_1) \cdot pr(N_{21}^{\kappa} = n_{21}^{\kappa} \mid S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot [1 - \pi_1(v)] \end{aligned} \tag{2.12}$$

where n_{21}^{κ} is the conditional sample size that satisfies $\text{Var}(\beta_{et}) < \kappa$.

Since $pr(N_{21}^{\kappa} = n_{21}^{\kappa} \mid S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) = pr(N_{21}^{\kappa} = n_{21}^{\kappa} \mid \boldsymbol{q}_1)$, subject i's stage two conditional likelihood contribution is

$$\begin{aligned} &L_{2i}^{c}(\boldsymbol{\theta} \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, S_{2i} = 1, n_{21}^{\kappa}) = \\ &\qquad [1 - L_{2i}(\boldsymbol{\theta} | V_i = 0, \boldsymbol{x}_{2i}, \boldsymbol{q}_1, n_{21}^{\kappa}) - L_{2i}(\boldsymbol{\theta} \mid V_i = 2, \boldsymbol{x}_{2i}, \boldsymbol{q}_1, n_{21}^{\kappa})]^{-1} \cdot L_{2i}(\boldsymbol{\theta} \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \end{aligned}$$

where $L_{2i}(\boldsymbol{\theta} \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i})$ is defined in Equation 4.4. For the remainder of this chapter, we assume that $pr(N_{21}^{\kappa} = n_{21}^{\kappa} \mid \boldsymbol{q}_1) = 1$ which implies that N_{21}^{κ} is fixed and known. We realize this assumption results in under-estimated standard errors when estimating parameters using Equation 2.13.

2.5.2 Stage Two: Adaptive Design

When the inferential target is not a single time-varying covariate, the choice of the conditionally-optimal stage two ODS design is not straightforward. Suppose it is of interest to evaluate the joint effect of a time-invariant and time-variant covariate effect (e.g., total genetic effect: β_e=main gene effect, and β_{et}=gene by time interaction). To be able to efficiently evaluate the two degree-of-freedom composite test, the stage two ODS design needs to involve the sampling of individuals from all sampling strata. As described in Section 2.5.1, the oversampling of any-responders is preferred for time-varying covariates, whereas the sampling of some of the all- or no-responders is preferred when it is of interest to estimate a time-invariant covariate effect. Since time-invariant covariate effects only vary between subjects, most information gain is obtained by sampling those with the predisposition of being healthy and those who tend to be sicklier. To identify the conditionally-optimal stage two design, a brute force approach may be taken, but depending on the stratum sizes, and the desired stage two sample size, it is likely to be computationally burdensome.

To reduce candidate design space, we only consider symmetric designs of the form $n_2 \cdot D_2[\alpha, 1-2\alpha, \alpha]$ where n_2 denotes the overall fixed stage two sample size (i.e., $n_2 = n_{20} + n_{21} + n_{22}$), and α denotes the proportion sampled in each of the extreme sample strata. The motivation of a symmetric design is that maximum variation is obtained by evenly sampling subjects from the most dissimilar strata. Using this reduced design space, a brute force search may be implemented to identify the conditionally-optimal balanced stage two ODS design.

Let D^α denote the stage two adaptive design, and the associated stage two sampling probabilities are defined as

$$\begin{aligned} pr(S_{2i}=1 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) &\equiv pr(S_{2i}=1, D^\alpha = d^\alpha, S_{1i}=0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \\ &= pr(S_{2i}=1, D^\alpha = d^\alpha \mid S_{1i}=0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot pr(S_{1i}=0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \\ &= pr(S_{2i}=1 \mid D^\alpha = d^\alpha, S_{1i}=0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot \\ &\qquad pr(D^\alpha = d^\alpha \mid S_{1i}=0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot pr(S_{1i}=0 \mid \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \\ &= \pi_2(v; \boldsymbol{q}_1) \cdot pr(D^\alpha = d^\alpha \mid S_{1i}=0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{q}_1) \cdot [1-\pi_1(v)] \end{aligned} \tag{2.13}$$

Similar to the adaptive sample size designs, we assume that $pr(D^\alpha = d^\alpha \mid \boldsymbol{q}_1) = 1$ which implies that D^α is fixed and known. Under this assumption, subject i's

contribution to the stage two adaptive design likelihood is:

$$L^c_{2i}(\boldsymbol{\theta}|\boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, S_{2i} = 1, d^\alpha) = \frac{\pi_2(v; d^\alpha, \boldsymbol{q}_1) \cdot [1 - \pi_1(v)]}{AC_{2i}} \cdot L_{2i}(\boldsymbol{\theta}|\boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \tag{2.14}$$

where the stage two ascertainment correction, AC_{2i}, is defined as $\sum_{v=0}^{2} \pi_2(v; d^\alpha, \boldsymbol{q}_1)[1-\pi_1(v)] \cdot pr(V_i = v|\boldsymbol{x}_{2i}, \boldsymbol{q}_1)$, and $L_{2i}(\boldsymbol{\theta}|\boldsymbol{y}_{2i}, \boldsymbol{x}_{2i})$ is defined in Equation 4.4.

2.5.3 Comparative Design Analysis

A comparative design analysis, like a power or sample size analysis, is used to estimate the stage two adaptive sample size or design using available data. The key to this analysis is that we need to be able to generate X_e from $[X_e \mid \boldsymbol{X}_o, \boldsymbol{Y}]$ for non-sampled subjects where X_o denotes the observed covariate matrix. Once we have this, we can generate the full cohort and then conduct designs and analysis procedures to explore which ones are likely to improve operating characteristics. We will describe two ways of estimating $[X_e \mid \boldsymbol{X}_o, \boldsymbol{Y}, \boldsymbol{S}_1 = 0]$.

Approach 1

Let $X_{ei}, \boldsymbol{X}_{oi}, \boldsymbol{Y}_i$ and S_{ki} denote subject i's time-invariant binary exposure of interest (e.g., SNP), observed design matrix, response, and indicator of being sampled during stage $k = \{1, 2\}$, respectively. Since the missing exposure variable is "missing by design" for those individuals not sampled at stage one, the sampling design is ignorable (Rubin, 1976). This implies that the conditional exposure model for a non-sampled subject is identical to that of a sampled subject irrespective of the sampling stage:

$$\begin{aligned} pr(X_{ei} = 1|\boldsymbol{x}_{oi}, \boldsymbol{y}_i, S_{1i} = 0) &= pr(X_{ei} = 1|\boldsymbol{x}_{oi}, \boldsymbol{y}_i) \\ &= pr(X_{ei} = 1|\boldsymbol{x}_{oi}, \boldsymbol{y}_i, S_{1i} = 1) \end{aligned} \tag{2.15}$$

Without loss of generality, we momentarily assume that subject i was sampled during stage one which implies $S_{1i} = S_i$. Via Bayes' formula, the conditional exposure

odds for non-sampled subjects is defined as

$$\begin{aligned}\frac{pr(X_{ei}=1|\boldsymbol{x}_{oi},\boldsymbol{y}_i,S_i=0)}{pr(X_{ei}=0|\boldsymbol{x}_{oi},\boldsymbol{y}_i,S_i=0)} &= \frac{pr(X_{ei}=1|\boldsymbol{x}_{oi},\boldsymbol{y}_i,S_i=1)}{pr(X_{ei}=0|\boldsymbol{x}_{oi},\boldsymbol{y}_i,S_i=1)} \\ &= \frac{pr(\boldsymbol{Y}_i \mid X_{ei}=1,\boldsymbol{x}_{oi},S_i=1;\boldsymbol{\theta})}{pr(\boldsymbol{Y}_i \mid X_{ei}=0,\boldsymbol{x}_{oi},S_i=1;\boldsymbol{\theta})}\frac{pr(X_{ei}=1|\boldsymbol{x}_{oi},S_i=1)}{pr(X_{ei}=0|\boldsymbol{x}_{oi},S_i=1)}\end{aligned} \tag{2.16}$$

Modeling the marginal exposure odds among sampled subjects may not be straightforward due to spurious associations induced by the sampling design. Alternatively, the conditional exposure odds may be factored into the product of the ascertainment-correction ratio and the marginal exposure odds

$$\frac{pr(X_{ei}=1|\boldsymbol{x}_{oi},S_i=1)}{pr(X_{ei}=0|\boldsymbol{x}_{oi},S_i=1)} = \frac{pr(S_i=1|X_{ei}=1,\boldsymbol{x}_{oi})}{pr(S_i=1|X_{ei}=0,\boldsymbol{x}_{oi})}\frac{pr(X_{ei}=1|\boldsymbol{x}_{oi})}{pr(X_{ei}=0|\boldsymbol{x}_{oi})} \tag{2.17}$$

The conditional exposure odds among non-sampled subjects is estimated by assuming the following functional form of the marginal population exposure model

$$\text{logit}\left[pr(X_{ei}=1|\boldsymbol{x}_{oi};\boldsymbol{\omega})\right] = \boldsymbol{x}_{oi}\boldsymbol{\omega}$$

and implementing the following five steps:

Among sampled subjects,

1. estimate $\hat{\boldsymbol{\theta}}^p$ and $\widehat{Cov}(\hat{\boldsymbol{\theta}}^p)$ by maximizing the combined two-stage ODS profile-likelihood defined in Equation 2.10 where we profile over parameters that are used in the identification of the stage two design. For example, we profile over β_{et} (i.e., fix its value with a hypothesized value) when it is of interest to find the sample size that minimizes $Var(\beta_{et})$. This reduces the inflation of type-I errors associated using stage one data to inform future design choices (Haneuse et al., 2012).
2. estimate $\hat{\boldsymbol{\omega}}$ and $\widehat{Cov}(\hat{\boldsymbol{\omega}})$ using an offsetted-logistic regression where the offsets are defined as the log-transformed ascertainment-correction ratios, as derived in Equation 2.17.

Among non-sampled subjects,

3. estimate the conditional likelihood ratio using Item 1,
4. estimate the conditional exposure model using Item 2, and

5. multiply quantities from Items 3 and 4 to estimate the conditional exposure odds among non-sampled subjects, as summarized in Equation 2.17.

Missing exposure information is estimated by performing independent Bernoulli sampling where the probability of success (or expsoure presence) is defined as $pr(X_{ei} = 1|\boldsymbol{x}_{oi}, \boldsymbol{y}_i, S_i = 0; \hat{\boldsymbol{\theta}}^p, \hat{\boldsymbol{\omega}})$.

Approach 2

The expectation-maximization algorithm utilizes a similar approach to estimate $pr(X_{ei} = 1|\boldsymbol{x}_{oi}, \boldsymbol{y}_i, S_i = 0)$ as defined in Section 2.5.3. The expectation steps at iteration m of this algorithm include:

1. computing $pr(\boldsymbol{Y}_i \mid X_{ei} = x, \boldsymbol{x}_{oi}; \hat{\boldsymbol{\theta}}^{p(m-1)})$ for $x = (0, 1)$, and
2. computing $p_i = pr(X_{ei} = 1|\boldsymbol{x}_{oi}, \boldsymbol{y}_i; \hat{\boldsymbol{\theta}}^{p(m-1)}, \hat{\boldsymbol{\omega}})$

and the maximization steps include:

3. computing

$$E(l) = \sum_{S_1=1} \log\left[pr(\boldsymbol{Y}_i \mid x_{ei}, \boldsymbol{x}_{oi})\right] + \sum_{S_1=0} p_i \log\left[pr(\boldsymbol{Y}_i \mid X_{ei} = 1, \boldsymbol{x}_{oi})\right] + (1 - p_i) \log\left[pr(\boldsymbol{Y}_i \mid x_{ei} = 0, \boldsymbol{x}_{oi})\right]$$

4. and maximizing $E(l)$ with respect to $\boldsymbol{\theta}^p$ and define $\hat{\boldsymbol{\theta}}^p = \hat{\boldsymbol{\theta}}^{p(m)}$

This process is repeated until a convergence criterion is achieved. Using Bayes' rule, the conditional exposure model for the non-sampled subjects is

$$pr(\boldsymbol{X}_{ei} = 1 \mid \boldsymbol{y}_i, \boldsymbol{x}_{oi}, S_i = 0) = \frac{pr(\boldsymbol{Y}_i \mid X_{ei} = 1, \boldsymbol{x}_{oi}; \hat{\boldsymbol{\theta}}^{p(m)})}{pr(\boldsymbol{Y}_i \mid X_{ei} = 0, \boldsymbol{x}_{oi}; \hat{\boldsymbol{\theta}}^{p(m)})} \frac{pr(X_{ei} = 1 \mid \boldsymbol{x}_{oi}; \hat{\boldsymbol{\omega}})}{pr(X_{ei} = 0 \mid \boldsymbol{x}_{oi}; \hat{\boldsymbol{\omega}})}$$

Missing exposure information is estimated by performing independent Bernoulli sampling. This approach differs from Approach 1 since the expected likelihood ratio is being estimated, and not the conditional likelihood for the sampled subjects only. Since the missingness mechanism is ignorable, we expect these quantities to be similar and only apply Approach 1 in these analyses.

2.5.4 Simulation

We investigate the operating characteristics of two-stage adaptive ODS designs by considering the same simulation set-up as described in Section 2.4.5. We assume that

it is of interest to define the stage one design as $\mathrm{D}_1[25, 50, 25]$ and to : 1) identify the stage two design so that $Var(\beta_{et}) \approx 0.005$, and 2) identify the stage two design of size 400 that maximizes the precision of the β_{et}, the precision of β_e, and the d-efficiency of (β_e,β_{et}). Even though the stage one design is not the optimal design for $Var(\beta_{et})$, it is of interest to estimate the main effect accurately which requires the sampling of those without response variation.

We perform 250 replicates of the simulation where we generate a population, and then implement the $\mathrm{D}_1[25, 50, 25]$ stage one design. We then use the imputation approach to estimate the exposure for all non-sampled subjects. With a complete data set, we identify the stage two sample size, n_{21}^{κ}, and the stage two design, d_2^{α}, to meet the two-stage ODS objectives. This process is repeated on a total of 25 imputed data sets, and the average sample size, and the design which maximizes the average optimality criterion of interest is used to define the stage two ODS design. Next, we describe one approach to estimating n_{21}^{κ} and d_2^{α}, and summarize the results of the 250 replications.

Adaptive sample size

The two-stage adaptive sample size ODS design in this simulation is defined as $\mathrm{D}_2[0, n_{21}^{0.005}, 0]$ where $n_{21}^{0.005}$ is the stratum sample size for any-responders such that $Var(\beta_{et}) \approx 0.005$ conditional on $\mathrm{D}_1[25, 50, 25]$. Our goal is to estimate the n_{21} and $Var(\beta_{et})$ relationship accurately in the vicinity of $Var(\beta_{et}) = 0.005$. Using the results from a comparative design analysis, one approach to estimating this relationship is to evaluate candidate designs at percentiles of the N_1 distribution (e.g., $\mathrm{D}_2[0, \frac{N_1}{2}, 0]$). Predicted n_{21} values can then be used to identify the value $n_{21}^{0.005} \approx 0.005$. Figure 2.2 is a realization of this approach. The numbers in the plotting region correspond to the iteration number (e.g., after 30 intial values, 40 is located near the optimal value). The subgraphic provides details regarding the variability of the estimate of $n_{21}^{0.005}$ after 250 iterations. For this example, 318 (95% CI: 278-361) is the optimal $n_{21}^{0.005}$. Table 2.2 summarizes the distribution of $n_{21}^{0.005}$ after 250 replications. A stage two sample size of $\approx 280 - 300$ is required such that $Var(\beta_{et}) \approx 0.005$.

Table 2.3 summarizes the bias, and coverage probability of the point estimates after performing this adaptive two-stage ODS design. Point estimates for all model parameters are unbiased, but this approach does under-estimate the standard error

Figure 2.2: Example of an algorithm to identify $\mathrm{D}_2[0, n_{21}^{\kappa}, 0]$ such that $Var(\beta_{et}) \approx 0.005$.

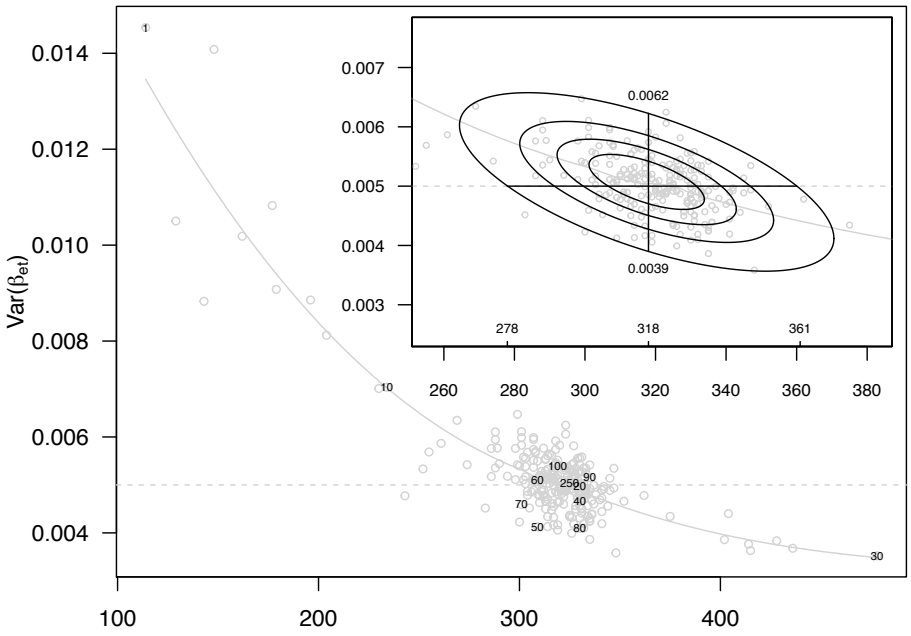


Table 2.2: Distribution of n_{21}^{κ} after 250 replications.

	κ	Mean	SD	Min	5	25	50	75	95	Max
D_1[25,50,25]	0.005	293	32	250	261	271	282	306	348	454

by up to 8% (i.e., intercept) in our simulation. This is not unexpected, since we assume that $n_2^{0.005}$ is fixed and do not acknowledge the variability associated with this random quantity. Regardless, for our estimation targets, we are still doing well. The average standard error for $\beta_{et} = 0.07$ which corresponds to a variance estimate of 0.0049.

Table 2.3: Parameter estimates using a two-stage adaptive sample size ODS design across 250 replications.

	True Value	Average Estimate	Estimate, Bias (%)	Average Std. Error	Std. Error, Bias (%)	Coverage
β_0	−1.50	−1.50	0	0.16	−8	0.95
β_t	−0.25	−0.24	−4	0.05	−6	0.91
β_e	1.00	1.02	2	0.22	−5	0.94
β_{et}	0.25	0.25	0	0.07	−5	0.94
γ	2.00	2.00	0	0.16	−1	0.95

Adaptive design

The two-stage adaptive design ODS design in this simulation is defined as $n_2\mathrm{D}_2[\alpha, (1-2\alpha), \alpha]$ where $n_2 = 400$ represents the pre-specified stage two sample size, and α denotes the proportion of individuals sampled from the no- and all-responder strata. We aim to identify the stage-two ODS design such that the precision of β_e, $Var(\beta_e)^{-1}$, or the precision of β_{et}, or the d-efficiency of (β_e, β_{et}) is maximized conditional on $\mathrm{D}_1[25, 50, 25]$. For the comparative design analysis, we fix $\beta_e = \beta_{et} = 0$. A brute force approach is adopted whereby all possible designs are identified such that $\mathrm{D}_1[25, 50, 25]+400\mathrm{D}_2[\alpha, (1-2\alpha), \alpha] \leq [N_0, N_1, N_2]$ where N_v denotes the number of subjects in the stratum v in the population. Each of the eligible designs is implemented using an imputed data set, and optimality criteria estimated. This process is repeated 250 times, and the design that maximizes the average criterion is chosen as the optimal stage two adaptive design.

Table 2.4 summarizes the percentage of times each candidate stage-two design is selected by the design implemented in stage one. Eligible designs ranged from $\mathrm{D}_2[0, 400, 0]$ to $\mathrm{D}_2[160, 80, 160]$ since the expected simulated population stratum size for the all-responder group is 160. Only those designs up to $\mathrm{D}_2[80, 240, 80]$ are presented since designs with stratum sizes greater than 80 in the no- or all-responder strata were never selected as optimal. When it is of interest to estimate a time-varying effect only then selecting the design that primarily samples those with response variation is optimal (96% of the selected designs sampled either 0 or 5 in the no-responder stratum). When it is of interest to estimate a time-invariant covariate effect only then selecting the design that samples approximately 60-75 in the no- or all-responder strata is optimal. If interest lies in the joint time-varying and time-invariant covariate effect, then designs that sample between 55-65 total individuals from the no- or all-responder strata results in the maximum d-efficiency estimate. Therefore, the chosen design depends on features of the model.

Table 2.5 presents the bias and coverage probability of all parameter estimates from a two-stage adaptive design ODS design. All point estimates are unbiased, and all attain nomial coverage. Stage two designs that only targeted β_{et} resulted in roughly 15% (0.06 vs 0.07) smaller standard error estimates than those that did not. Similarly, observations are made for designs that targeted β_e (e.g., 0.18 vs 0.20).

Table 2.4: Percentage of times each stage two design (rows) was selected as optimal by stage one design (columns) across 250 replications.

	$\mathbf{D_1}$**[25,50,25]**		
	$Var(\beta_{et})^{-1}$	$Var(\beta_e)^{-1}$	d-efficiency(β_e, β_{et})
D_2[0,400,0]	95.2	-	-
D_2[5,390,5]	4.6	-	-
D_2[10,380,10]	0.2	-	-
D_2[15,370,15]	-	-	-
D_2[20,360,20]	-	-	-
D_2[25,350,25]	-	-	0.2
D_2[30,340,30]	-	0.2	1.2
D_2[35,330,35]	-	1.6	7.4
D_2[40,320,40]	-	15.2	18.4
D_2[45,310,45]	-	29.8	19.6
D_2[50,300,50]	-	41.2	29.2
D_2[55,290,55]	-	9.8	15.4
D_2[60,280,60]	-	2.2	7.2
D_2[65,270,65]	-	-	1.4
D_2[70,260,70]	-	-	-
D_2[75,250,75]	-	-	-
D_2[80,240,80]	-	-	-

Table 2.5: Parameter estimates using a two-stage adaptive design, fixed sample size design

		$\mathbf{D_1}$**[25,50,25]**				
	True Value	Est	Bias (%)	Avg SE	Bias (%)	Coverage
$Var(\beta_{et})^{-1}$						
β_0	−1.50	−1.51	0	0.15	1	0.96
β_t	−0.25	−0.25	0	0.04	0	0.96
β_e	1.00	1.00	0	0.20	1	0.95
β_{et}	0.25	0.25	2	0.06	0	0.95
γ	2.00	2.00	0	0.14	−3	0.95
$Var(\beta_e)^{-1}$						
β_0	−1.50	−1.50	0	0.12	−2	0.95
β_t	−0.25	−0.25	0	0.04	−2	0.94
β_e	1.00	1.00	0	0.18	2	0.95
β_{et}	0.25	0.25	0	0.07	−1	0.95
γ	2.00	2.00	0	0.12	−1	0.96
d-efficiency(β_e, β_{et})						
β_0	−1.50	−1.50	0	0.12	3	0.94
β_t	−0.25	−0.25	0	0.04	2	0.94
β_e	1.00	1.00	0	0.18	5	0.95
β_{et}	0.25	0.25	0	0.07	0	0.95
γ	2.00	2.00	0	0.12	−1	0.95

2.6 Example: Lung Health Study

The Lung Health Study (LHS) was a multi-center study that evaluated the effectiveness of smoking cessation and inhaled bronchodilators on lung function in middle-aged smokers with mild to moderate chronic obstructive pulmonary disease (COPD) (Anthonisen, 2004; Connett et al., 1993). Annual spirometry measurements were collected for five years, and at the last visit blood samples were obtained to aid in the identification of genetic factors associated with lung function decline and lung cancer (Anthonisen, 2004). A sub-study, entitled the Genome-Wide Associations Environmental Interactions in the Lung Health Study, genotyped banked DNA on 4,287 European Americans from the LHS cohort (dbGaP (Mailman et al., 2007); access number phs000335.v2.p2). Hansel et al (2013) identified two genetic risk factors associated with lung function decline within this cohort including the single nucleotide polymorphism (SNP) rs177852.

We demonstrate the class of two-stage ODS designs when it is of interest to quantify the relationship between lung function decline and the presence of a T-allele in rs177852 under the constraint that only 750 subjects (20%) can be genotyped. For this analysis, lung function decline is defined as a 2-unit decrease in the percent predicted forced expiratory volume in one-second from baseline (Yuan et al., 2009). We assume the following marginalized transition and latent variable model:

$$\begin{aligned}
\text{logit}\left(\mu_{ij}^{m}\right) &= \beta_0 + \beta_t \text{time}_{ij} + \beta_e \text{SNP}_i + \beta_{et} \text{SNP}_i \cdot \text{time}_{ij} + \ldots \\
\text{logit}\left(\mu_{ij}^{c}\right) &= \Delta_{ij} + \gamma Y_{ij-1} + \sigma Z_i
\end{aligned}$$

where $\ldots$ denotes baseline FEV_1 percent predicted, gender, baseline BMI, age, and smoking status. Smoking status included the main effects cigarettes per/day, pack-years, and current smoking status that had been decomposed into its between- and within-subject components (i.e., between=$\bar{x}_i$, within=$x_{ij} - \bar{x}_i$).

We consider the following three two-stage fixed designs where 250 subjects are sampled in stage one (similar to an internal pilot study) and the remaining 500 are sampled in stage two:

1. D[50,650,50] $\equiv$ $\text{D}_1[25, 200, 25]$+$\text{D}_2[25, 450, 25]$,
2. D[100,550,100] $\equiv$ $\text{D}_1[50, 150, 50]$+$\text{D}_2[50, 400, 50]$, and
3. D[150,450,150] $\equiv$ $\text{D}_1[75, 100, 75]$+$\text{D}_2[75, 350, 75]$.

Both adaptive two-stage designs utilized a $\text{D}_1[50, 150, 50]$ stage one design, and identified the n_{21}^{κ} such that $\text{D}_2[0, n_{21}^{\kappa}, 0]$ resulted in $Var(\beta_{et}) \approx 0.04^2$, and the design

such that $500 \cdot \mathrm{D}_2[\alpha, (1-2\alpha), \alpha]$ resulted in maximizing $Var(\beta_e)^{-1}$, $Var(\beta_{et})^{-1}$, or d-efficiency(β_e, β_{et}). We do not consider the extreme design where only subjects with response variation are sampled since it is of interest to estimate the SNP effect.

We perform 250 replications of the simulation where each of the random and ODS sampling schemes are implemented on the complete LHS data set. For the two-stage adaptive ODS designs, we use imputation to conduct a comparative design analysis to estimate $n_{21}^{0.0016}$ and d_2^{α}. This process is repeated on a total of 25 imputed data sets, and the average sample size, and the design which maximizes the average optimality criterion of interest is used to define the stage two ODS design. The relative efficiency (RE) associated with each ODS design is compared to random sampling, and defined as the average optimality value under random sampling divided by the average value under the two-stage ODS designs.

Table 2.6 summarizes the demographics of the Lung Health Study cohort that had been genotyped. Fifty-five percent had at least one T allele on rs177852, and at baseline 50% had an FEV_1 percent predicted less than 80 indicating that half met the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria for moderate COPD. At the first follow-up visit, 30% had reduced lung function as defined as a 2% reduction in FEV_1 percent predicted from baseline (which we now refer to as "lung function decline"). This value decreased to 16% by the end of the fifth follow-up visit indicating that lung function decline stabilized for many of the study participants. Of the 3,771 subjects considered for this analysis, 2103, 1474, and 194 experienced lung function decline at none, at least one, all follow-up visits, respectively.

Table 2.6: Demographics of the Lung Health Study (genotyped) cohort. Categorical variables are summarized as proportions and frequencies, and baseline continuous measurements are summarized with $[5, 25, 50, 75, 95]^{th}$ percentiles.

	Summary
Baseline Measurements	
Number of observations	3771
Female	0.37
Age (years)	[37 : 43 : 49 : 54 : 58]
BMI (kg/m^2)	[20 : 23 : 26 : 29 : 33]
Pack-years	[17 : 28 : 37 : 50 : 75]
Cigarettes (per day)	[10 : 20 : 30 : 40 : 55]
Percent Predicted FEV_1	[62 : 73 : 79 : 86 : 92]
Any T allele	0.55
Longitudinal Measurements	
Number of follow-up observations	3671 - 3708 - 3714 - 3662 - 3751
Current smoker	0.70 - 0.69 - 0.67 - 0.65 - 0.64
Percent Predicted $\mathrm{FEV}_1 < -2$ from baseline	0.30 - 0.22 - 0.19 - 0.16 - 0.16
No $\mathrm{FEV}_1 < -2$	2103
Any $\mathrm{FEV}_1 < -2$	1474
All $\mathrm{FEV}_1 < -2$	194

Full Cohort Analysis

Based on the full cohort analysis, we did not detect an association between the main effect of SNP and of lung function decline (Table 2.7). We observe a significant SNP by time interaction in which the rate of lung function decline is greater among those individuals with the SNP compared to those without the SNP (exp(0.08)=1.08, 95%CI: 1.04-1.13). Other factors associated with lung function decline include baseline FEV_1 percent predicted, smoking (pack years, and current smoking status), and age.

Two-stage Fixed ODS Analyses

Each of the two-stage fixed ODS designs reproduced the full cohort estimates. When compared to random sampling, the D[50,650,50] design resulted in a more efficient estimate of the SNP by time interaction (0.036 vs 0.053, RE=2.17), but a less efficient estimate for the main SNP effect (0.177 vs 0.157, RE=0.78). These efficiency differences are due to oversampling those with response variation. As the number of sampled subjects with response variation decreased (sampling fewer individuals in the central stratum) the efficiency of the main effect of SNP increased (0.157 and 0.149 versus 0.177; RE=1.27-1.44).

Table 2.7: Regression results of the full cohort analysis, and average estimates [average standard errors] of 500 replications of each study design. Study designs considered, include: full cohort (FC, n=3771), and the sampling of 750 subjects using random sampling (RS), and two-stage fixed ODS designs.

	FC	RS	$D_1[25,200,25]$ $D_2[25,450,25]$	$D_1[50,150,50]$ $D_2[50,400,50]$	$D_1[75,100,75]$ $D_2[75,350,75]$
Mean					
Intercept	-1.80 [0.08]	-1.81 [0.19]	-1.74 [0.24]	-1.77 [0.20]	-1.81 [0.18]
SNP	-0.09 [0.069]	-0.11 [0.156]	-0.14 [0.177]	-0.12 [0.157]	-0.07 [0.149]
SNP x Visit	0.08 [0.023]	0.08 [0.053]	0.08 [0.036]	0.09 [0.039]	0.09 [0.042]
FEV, percent predicted (per 2)	0.02 [0.01]	0.02 [0.02]	0.01 [0.02]	0.00 [0.02]	0.00 [0.01]
Cigarettes/day (per 10)	-0.01 [0.02]	-0.01 [0.05]	-0.06 [0.07]	-0.05 [0.06]	-0.05 [0.05]
Packs/years (per 20)	0.11 [0.04]	0.11 [0.09]	0.15 [0.11]	0.14 [0.09]	0.14 [0.08]
Current smoking status (between)	1.22 [0.08]	1.23 [0.18]	1.20 [0.24]	1.22 [0.20]	1.23 [0.17]
Current smoking status (within)	0.44 [0.07]	0.45 [0.15]	0.42 [0.10]	0.44 [0.11]	0.45 [0.13]
Visit	-0.25 [0.02]	-0.25 [0.04]	-0.25 [0.03]	-0.25 [0.03]	-0.25 [0.03]
Female	0.01 [0.06]	0.01 [0.15]	-0.03 [0.18]	0.01 [0.15]	0.03 [0.14]
Baseline BMI (per 5 kg/m^2)	0.01 [0.04]	0.01 [0.09]	-0.03 [0.11]	-0.02 [0.09]	0.00 [0.08]
Age (per 10 years)	0.12 [0.05]	0.13 [0.11]	0.03 [0.15]	0.08 [0.12]	0.11 [0.11]
Dependence					
γ	1.01 [0.08]	1.00 [0.19]	1.00 [0.12]	1.01 [0.14]	1.02 [0.15]
$\log(\sigma)$	0.61 [0.04]	0.61 [0.09]	0.60 [0.09]	0.60 [0.08]	0.60 [0.08]

Two-stage Adaptive Sample Size Analysis: $D_2[0, n_{21}^{\kappa}, 0]$

The two-stage adaptive sample size design resulted in the sampling of an additional 490 subjects with response variation such that $Var(\beta_{et}) \approx 0.04^2$. Table 2.8 summarizes the estimated distribution of $n_{21}^{0.0016}$ across 500 replications. Sixty-six percent of the replications resulted in sample size estimates greater than 500, thus this approach may be used to assess if the continuation of the planned study is feasible, or if additional resources are needed to meet the study objectives. From Table 2.9, the resultant two-stage adaptive design, $\mathrm{D}_1[50, 150, 50]$+$\mathrm{D}_2[0, 490, 0]$, reproduced the full cohort point estimates. Point estimates and standard errors are similar to the two-stage fixed design $D[50, 650, 50]$ (Table 2.7).

Table 2.8: Distribution of n_{21}^{κ} after 500 replications.

	κ	Mean	SD	Min	5	25	50	75	95	Max
$\mathrm{D}_1[50, 150, 50]$	0.0016	489	36	373	434	463	486	514	547	598

Table 2.9: Regression results of the full cohort analysis, and average estimates [average standard errors] of 500 replications of each study design. Study designs considered, include: full cohort (FC, n=3771), and the two-stage adaptive sample size ODS designs.

	FC	$\mathrm{D}_2[0, 489, 0]$
Mean		
Intercept	-1.80 [0.08]	-1.72 [0.25]
SNP	-0.09 [0.069]	-0.14 [0.181]
SNP x Visit	0.08 [0.023]	0.08 [0.037]
FEV_1, percent predicted (per 2)	0.02 [0.01]	0.01 [0.02]
Cigarettes/day (per 10)	-0.01 [0.02]	-0.05 [0.07]
Packs/years (per 20)	0.11 [0.04]	0.15 [0.11]
Current smoking status (between)	1.22 [0.08]	1.17 [0.24]
Current smoking status (within)	0.44 [0.07]	0.43 [0.11]
Visit	-0.25 [0.02]	-0.25 [0.03]
Female	0.01 [0.06]	0.00 [0.19]
Baseline BMI (per 5 $\mathrm{kg/m^2}$)	0.01 [0.04]	-0.02 [0.12]
Age (per 10 years)	0.12 [0.05]	0.02 [0.15]
Dependence		
γ	1.01 [0.08]	1.00 [0.13]
$\log(\sigma)$	0.61 [0.04]	0.60 [0.10]

Two-stage Adaptive Design Analysis: $n_2 \cdot D_2[\alpha, (1-2\alpha), \alpha]$

From Table 2.10, the two-stage adaptive design that aimed to maximize the precision of β_{et} overwhelmingly chose $\mathrm{D}_2[0, 500, 0]$ (88%). We observed more variability in the two-stage adaptive designs that targeted the precision of β_e, and the d-efficiency of (β_e, β_{et}). The stage two designs most frequently identified using these optimality criteria were $\mathrm{D}_2[110, 280, 110]$ (33%) and $\mathrm{D}_2[90, 320, 90]$ (31%), respectively. Table 2.11

presents regression estimates and standard errors for the full cohort for each of the study designs. The two-stage design that targeted the precision of $Var(\beta_{et})$ resulted in efficiency gains compared to random sampling (0.039 versus 0.053; RE=1.85). The two-stage designs that targeted the precision of $Var(\beta_e)$ and the d-efficiency of (β_e, β_{et}) produced standard errors comparable to random sampling, but the later did result in improved efficiency of the β_{et} covariate effect.

Table 2.10: Percentage of times the stage two design maximized the optimality criteria: precision of β_{et} or β_e, and d-efficiency(β_e, β_{et}) based on 500 replications.

	$Var(\beta_{et})^{-1}$	$Var(\beta_e)^{-1}$	d-efficiency(β_e, β_{et})
0-500-0	88	0	0
10-480-10	11	0	0
20-460-20	1	0	0
30-440-30	0	0	0
60-380-60	0	0	3
70-360-70	0	0	11
80-340-80	0	0	21
90-320-90	0	8	31
100-300-100	0	28	21
110-280-110	0	33	8
120-260-120	0	26	4
130-240-130	0	5	0
140-220-140	0	0	0

Table 2.11: Regression results of the full cohort analysis, and average estimates [average standard errors] of 500 replications of each study design. Study designs considered, include: full cohort (FC, n=3771), and the sampling of 750 subjects using random sampling (RS), and two-stage adaptive design ODS designs.

	FC	RS	$Var(\beta_{et})^{-1}$	$Var(\beta_e)^{-1}$	d-efficiency(β_e, β_{et})
Mean					
Intercept	-1.80 [0.08]	-1.81 [0.19]	-1.76 [0.28]	-1.82 [0.19]	-1.82 [0.19]
SNP	-0.09 [0.069]	-0.11 [0.156]	-0.13 [0.191]	-0.07 [0.158]	-0.08 [0.159]
SNP x Visit	0.08 [0.023]	0.08 [0.053]	0.08 [0.039]	0.09 [0.049]	0.09 [0.047]
FEV_1, percent predicted (per 2)	0.02 [0.01]	0.02 [0.02]	0.01 [0.02]	0.00 [0.01]	0.00 [0.01]
Cigarettes/day (per 10)	-0.01 [0.02]	-0.01 [0.05]	-0.05 [0.07]	-0.06 [0.05]	-0.05 [0.05]
Packs/years (per 20)	0.11 [0.04]	0.11 [0.09]	0.15 [0.11]	0.14 [0.08]	0.14 [0.08]
Current smoking status (between)	1.22 [0.08]	1.23 [0.18]	1.18 [0.26]	1.24 [0.17]	1.24 [0.18]
Current smoking status (within)	0.44 [0.07]	0.45 [0.15]	0.43 [0.11]	0.45 [0.15]	0.43 [0.14]
Visit	-0.25 [0.02]	-0.25 [0.04]	-0.25 [0.03]	-0.26 [0.04]	-0.25 [0.04]
Female	0.01 [0.06]	0.01 [0.15]	0.00 [0.20]	0.03 [0.13]	0.04 [0.14]
Baseline BMI (per 5 kg/m^2)	0.01 [0.04]	0.01 [0.09]	-0.02 [0.12]	0.00 [0.08]	0.00 [0.09]
Age (per 10 years)	0.12 [0.05]	0.13 [0.11]	0.03 [0.15]	0.11 [0.11]	0.11 [0.11]
Dependence					
γ	1.01 [0.08]	1.00 [0.19]	1.02 [0.14]	1.02 [0.18]	1.00 [0.17]
$\log(\sigma)$	0.61 [0.04]	0.61 [0.09]	0.62 [0.12]	0.60 [0.08]	0.61 [0.08]

2.7 Discussion

We extended ODS designs for longitudinal binary data to permit data collection in two stages. We consider two sub-classes of designs: fixed designs where the sampling probabilities at each stage are pre-specified, and adaptive designs that utilize stage one data to improve design choice at stage two. We demonstrate that data from both stages can be aggregated to generate valid parameter estimates using ascertainment-corrected maximum likelihood methods. Efficiency gains are observed compared to random sampling, and in certain situations, as efficient as single-stage ODS sampling designs. Magnitudes of these efficiency gains depend on characteristics of the data, such as the prevalence of the exposure, and the (relative) magnitude of the effect size of interest.

These designs show promise based on these preliminary simulations, but a more thorough empirical study is needed to understand the operating characteristics of these two-stage ODS designs. For example, we assume the stage two sampling probabilities are fixed when performing an adaptive stage two design. Ignoring this source of variation likely results in under-estimated standard errors, but based on these simulations the coverage probabilities are adequate. We plan on investigating bootstrapped standard errors to incorporate this additional source of variation. All simulations are also based on correctly-specified mean and dependence models. The misspecification of either model may result in invalid inferences, and result in erroneous stage two decisions if performing a stage two adaptive design.

Regardless of these limitations, two-stage ODS designs do provide flexibility over single stage ODS designs. The cost associated with a two-stage design is increased variability in the sampling weights which leads to a loss in efficiency compared to the induced single stage design. If the inferential target is a time-invariant covariate effect, or the joint time-varying and a time-invariant covariate effect, then the choice of the optimal ODS design is not clear. It may be worthwhile to perform a less efficient two-stage study versus performing a single-stage stage design that is possibly under- or over-powered.

CHAPTER 3

SURVEY DESIGN AND ANALYSIS CONSIDERATIONS WHEN UTILIZING AN IMPERFECT SAMPLING FRAME

3.1 Abstract

We investigate the effects of utilizing an imperfect sampling frame on the design and analysis of complex survey data. This study is motivated by a large multi-center survey developed to elicit perspectives on biobank participation among under-studied subgroups (e.g., racial and ethnic monitories). A disproportionate stratified sampling scheme is implemented to enrich the sample population with these less prevalent populations using a sampling frame primarily constructed from electronic health record data (EHR). Incomplete EHR data is imputed using geocode-derived census summaries which resulted in a well-defined, but imperfect sampling frame. We determine, via analytic calculations and simulations, that in the presence of stratum misclassification: 1) complex study designs result in more diverse samples compared to random sampling, 2) the efficiency of design-based estimators change as a function of the relative size of the sampling strata, and 3) that analytic methods that account for the design are still required for valid inferences. We explore the effects of stratum misclassification in a real-world example by analyzing a subset of the biobank survey data from Vanderbilt University Medical Center.

3.2 Introduction

The Consent, Education, Regulation and Consultation (CERC) working group of the electronic Medical Records and Genomics (eMERGE) network conducted a large, eleven-site survey to examine patient concerns about, and barriers to, participating in biobank-derived research. Since most research has historically been based on individuals of northern European ancestry, this survey aimed to enrich their sample with ethnic and racial minorities, as well as younger adults, individuals of low socioeconomic status, low education, and rural residence (Garrison et al., 2016). The sampling of rare, or under-studied, subpopulations is typically accomplished by applying a disproportionate stratified sampling scheme to a well-defined sampling frame (Kalton, 2009). CERC researchers constructed a sampling frame using electronic health records (EHR) data at each of the sites. As is well known, EHR data are incomplete and accuracy varies by institution and by variable, and thus incomplete

EHR data were supplemented with geocode-derived, census summaries. Using this imperfect sampling frame, disproportionate stratified sampling was used to identify the final sample. Additional details regarding the survey development (Smith et al., 2016), the sampling approach, and the results (Sanderson et al., 2017) have been published previously.

Two broad classes of analysis strategies for complex survey data include design-based, and model-based methodologies. Design-based inference assumes population values are fixed, and all inferences are based on the randomization distribution, or on the probability of being sampled (Cochran, 1977). To account for the sampling scheme, design-based inferences weight by the inverse of being sampled. This results in unbiased point estimates, even in the presence of informative sampling (Rubin, 1976; Sugden and Smith, 1984; Kim and Skinner, 2013). In the presence of survey non-response, weights are commonly modified based on the available data (e.g., post-stratification, raking). Model-based inference, assumes the observed survey data are random, and are generated from a statistical model. Model-based approaches account for the survey design by incorporating design variables into a regression model. For an outcome Y, matrix of design variables X, and sampling indicator S, then a study design is ignorable if $[Y|X] = [Y|X, S = 1]$. Model-based approaches result in unbiased, and more efficient estimates than the weighting approach when the survey design is ignorable. Otherwise, model-based estimates are biased.

Because our sampling strata are based on EHR data, we are concerned that an individual's assigned stratum identifier may be misclassified. Stratum misclassification mechanisms are either non-differential or differential. Non-differential (differential) misclassification occurs when the probability of being misclassified is independent (dependent) of the outcome of interest. For example, suppose there are $h \in H$ sampling strata, and let $h^\star \in H^\star$ denote the mismeasured strata. In the CERC survey, H and $H^\star$ correspond to the true sampling strata defined by the survey responses, and mismeasured strata derived using EHR and census data, respectively. Table 3.1 describes the square $H^\star \times H$ misclassification matricies between the true and mismeasured stratification variables under non-differential misclassification, and differential misclassification for a binary outcome Y. Let $\alpha_{h|h^\star}$ denote the probability that an individual belongs to stratum h, but is assigned to stratum $h^\star$ where $\sum_h \alpha_{h|h^\star} = 1$. The added subscript in Table 3.1 denotes similar values under differential misclassification and it is assumed that $\alpha_{h|h^\star,0} \neq \alpha_{h|h^\star,1}$. Under no misclassification, $\alpha_{h|h^\star} = \alpha_{h|h^\star,y} = 1$ for all $h = h^\star$ and all levels of $Y = y$. Similarly, if $\alpha_{h|h^\star} = \alpha_{h|h^\star,y}$ for all $h \neq h^\star$, then the misclassification mechanism is symmetric.

Table 3.1: Misclassification matrix example. Rows correspond to misclassified stratification variables, and columns correspond to true stratification values.

Non-differential misclassification

	1	2	...	H
$1^\star$	$\alpha_{1\|1^\star}$	$\alpha_{2\|1^\star}$	...	$\alpha_{H\|1^\star}$
$2^\star$	$\alpha_{1\|2^\star}$	$\alpha_{2\|2^\star}$	...	$\alpha_{H\|2^\star}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$H^\star$	$\alpha_{1\|H^\star}$	$\alpha_{2\|H^\star}$	...	$\alpha_{H\|H^\star}$

Differential misclassification for $Y = y$

	1	2	...	H
$1^\star$	$\alpha_{1\|1^\star,y}$	$\alpha_{2\|1^\star,y}$	...	$\alpha_{H\|1^\star,y}$
$2^\star$	$\alpha_{1\|2^\star,y}$	$\alpha_{2\|2^\star,y}$	...	$\alpha_{H\|2^\star,y}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$H^\star$	$\alpha_{1\|H^\star,y}$	$\alpha_{2\|H^\star,y}$	...	$\alpha_{H\|H^\star,y}$

Misclassification mechanisms are sensitive to the coding of the design variables, and the response (Rothman et al., 2008). For example, grouping racial categories or dichotomizing the response may result in an originally non-differential mechanism becoming differential (Flegal et al., 1991). The effect of stratum misclassification has been well studied in the setting where the design variables are not recollected in the survey instrument. It has been shown that misclassification leads to arbitrary forms of bias, and adjustments to sampling weights are needed for valid inference (e.g., matrix adjustment, log-linear models) (Kuha and Skinner, 1997; Greenland, 1988).

The purpose of this paper is to investigate the effects of stratum misclassification on the design, and on the analysis of survey data when mismeasured design variables are recollected in the survey instrument. In this chapter, we address three main questions: 1) is performing a complex study design beneficial when utilizing an imperfect sampling frame, 2) what is the effect of stratum misclassification on the operating characteristics of design- and model-based estimators, and 3) does the study design, that uses an imperfect sampling frame, need to be acknowledged when analyzing data complex survey data? In Section 2, we analytically derive the variance of design-based estimator of the total under non-differential misclassification, present similar formulas for the mean, and show that extensions to linear regression parameters are straightforward. In Section 3, we conduct a simulation study and present results that address design and analysis considerations when dealing with an imperfect sampling frame. Data from the eMERGE survey is analyzed in Section 4. In Section 5, we discuss general findings and study limitations.

3.3 Methods

3.3.1 Means and Variances of Design-Based Descriptive Estimators Under No Misclassification

Two common types of design-based estimators include Horvitz-Thompson and ratio estimators. In this section, we define the Horvitz-Thompson estimators of the total and mean, and the ratio estimator of the mean under no stratum misclassification. We omit other estimators since our primary focus is on understanding the effects of stratum misclassification on analytic summaries (e.g., linear regression parameters are equivalent to a difference in means). Before defining these estimators, we introduce key notation. For stratum $h \in H$, let N_h and n_h denote the population size and sample size, respectively. Sampling weights are typically defined as the inverse of the probability of being sampled, $w_h = \frac{N_h}{n_h}$. Let y_{jh} denote a continuous response for subject $j \in h$, and let x_h represent an auxiliary variable related to y_h that can be collected on each sampled individual.

Horvitz-Thompson estimators are typically applied when stratum sizes are known or when auxiliary information is not available, whereas ratio estimates are used when stratum sizes are unknown or when adjusting sampling weights to reflect the respondent population (e.g., post-stratification). Ratio estimators utilize auxiliary information to aid in the valid estimation of design-based descriptive statistics, and can be more efficient than Horvitz-Thompson estimators depending on the correlation between y_h and x_h. Table 3.2 summarizes these estimates, and their variances, of the Horvitz-Thompson estimators of the total and mean, and the ratio estimator of the mean (Lohr, 2009).

Table 3.2: Design-based estimators of descriptive statistics under no stratum misclassification.

Quantity	Estimate	Variance
Total	$\sum_{j=1}^{n_h} w_h y_{hj}$	$N_h^2(1-\frac{n_h}{N_h})\frac{S_{hy}^2}{n_h}$
Mean	$\frac{1}{N_h}\sum_{j=1}^{n_h} w_h y_{hj}$	$(1-\frac{n_h}{N_h})\frac{S_{hy}^2}{n_h}$
Mean (ratio)	$B_h \cdot \bar{x}_{hp}$	$\frac{1}{n_h}\left(1-\frac{n_h}{N_h}\right)\left[S_{hy}^2 + B_h S_{hx}^2 - 2B_h R_h S_{hx} S_{yh} - \frac{1}{\bar{x}_{hp}}\left(B_h S_{hx}^2 + R_h S_{hx} S_{hy}\right)\right]$

where $B_h = \frac{\bar{y}_h}{\bar{x}_h}$, $\bar{x}_{hp}$ the known population mean for variable x, S_{hk}^2 the population variance for $k = \{x, y\}$, and $R_h = Corr(\bar{y}, \bar{x})$. S_{hy}^2 is estimated as $\frac{1}{n_h - 1}\sum_{j=1}^{n_h}(y_{hj} - \bar{y}_h)^2$. Other estimators of S_{hy}^2 have been derived, but are not explored further (Courbois and Urquhart, 2004).

The Horvitz-Thompson and ratio estimators of a mean are equivalent when the auxiliary information is fixed within a stratum (e.g., a stratum indicator). If $x_{hj} = 1$ for all $j \in h$, then $\bar{x}_{hp} = \bar{x}_h = 1$ and $S_{hx}^2 = S_{hx} = 0$.

Since a simple random sample of subjects is taken within each stratum, estimates of marginal descriptive statistics are simply the sum the quantities over all strata. For example, the estimate of the population total is $\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj} y_{hj}$, and the corresponding variance is $\sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{\hat{S}_h^2}{n_h}$ (Lohr, 2009). Similarly, the variances of differences between stratum means are the sum of the individual stratum variances.

For more complicated estimators, such as quantiles or generalized linear regression model parameters, variance formulas and the appropriate choice of weights are not straightforward. This is because functional forms of variances may be complex, or unknown, and data characteristics may require ad-hoc weight modifications. Two common approaches to estimating variances of analytic summaries include linearization (i.e., the delta-method) and replication methods (e.g., jack-knife, bootstrapping). Linearization involves rewriting the variance of the inferential target as a function of totals, and then approximating this functional using a first-order Taylor series. Variances of generalized linear regression model parameters fall under this framework, and are now the default variance estimation method in most survey software (Binder, 1983; Lumley, 2011). If the variance cannot be rewritten in this way, then replication methods are typically applied (Rust and Rao, 1996).

3.3.2 Non-Differential Stratum Misclassification & Sub-Domain Analysis

To investigate the effects of non-differential misclassification on analytic quantities, we first derive estimators of descriptive summaries under non-differential stratum misclassification. Once completed for a mean, we show that extensions to a simple linear regression parameter is trivial (e.g., difference in means). For other generalized linear models or for multivariate regressions, analytic extensions are not straightforward.

We consider the scenario when the design variables used in constructing the sampling strata and the true values for the survey respondents are available at the time of analysis. In this setting, the problem of stratum misclassification can then be reformulated as a sub-domain analysis problem. Sub-domain analyses are performed when interest lies in analyzing a well-defined subgroup of the original sampled population. Since each survey datum contains design information (e.g., a sampling weight, finite population correction), performing a naïve analysis by ignoring individuals not belonging to the subgroup of interest results in incorrect standard error estimates (Graubard and Korn, 1996; Lumley, 2011). To acknowledge the design information, sub-domain analyses define an indicator to denote subgroup membership. For ex-

ample, suppose it is of interest to summarize biobank views, $y_{biobank}$, among female respondents. Let $I_{j,female} = 1$ if subject j is female, then correct standard errors are obtained by analyzing $I_{j,gender} \cdot y_{j,biobank}$ for all N subjects.

To illustrate how stratum misclassification is related to sub-domain analysis, consider the estimation of the total for stratum h. Let $N_{h^\star}$, $n_{h^\star}$ and $w_{h^\star}$ as the population size, sample size and sampling weight for stratum $h^\star$, respectively. Define $I_{h|h^\star j}$ as the response and indicator that subject j belongs to stratum h given they were initially assigned to stratum $h^\star$. For each $h^\star$, the contribution to the estimate of the total of stratum h is $\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}$. Therefore, we are performing a sub-domain analysis for each value of $h^\star$ (e.g., our sub-domain of interest includes those individuals that belong to stratum h). The final estimate of the stratum total is then defined as the sum of these sub-domain analyses. This version of the sub-domain analysis is identical to post-stratification because we are conditioning on $\boldsymbol{I}_{h|h^\star}$, or equivalently, conditioning on the observed stratum sample sizes.

The expectation and variance of a stratum total under misclassification are (see Appendix 3.7 for the complete derivation):

$$\begin{aligned} E\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) &= E\left(\sum_{h^\star=1}^{H^\star} w_{h^\star} \sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} y_{h|h^\star j}\right) \\ &= \sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j} = \sum_{j=1}^{N_h} y_{hj} \end{aligned} \tag{3.1}$$

$$\begin{aligned} Var\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) &= Var\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) \\ &= \sum_{h^\star=1}^{H^\star} \frac{N_{h^\star}^2}{n_{h^\star}} \left(1 - \frac{n_{h^\star}}{N_{h^\star}}\right) \alpha_{h|h^\star} \left[S_{hy}^2 + (1 - \alpha_{h|h^\star})\bar{y}_{hp}^2\right] \end{aligned} \tag{3.2}$$

where $I_{h^\star j}$ denotes the random sampling indicator of subject j in stratum $h^\star$, $\alpha_{h|h^\star}$ represents the predictive probability (or calibration probability, Table 3.1) of belonging to stratum h conditioned on being initially assigned to stratum $h^\star$, and $\bar{y}_{hp}$ and S_{hy}^2 the stratum h population average and variance of y.

From Equation 3.1, it can be seen that this estimator is (design) unbiased when using the original sampling weights, $w_{h^\star}$. The variance formula in Equation 3.2 resembles its counterpart in Table 3.2, but now is also a function of the predictive probability, as well as the stratum mean. When $\alpha_{h|h^\star} = 1$, this variance estimate reduces to the variance of the standard Horvitz-Thompson estimator. When $\alpha_{h|h^\star} = 0$,

stratum $h^\star$ does not contribute to the variance of the mean of stratum h. Due to the complex functional form of Equation 3.2, it is not necessarily clear how misclassification effects the variance when $0 < \alpha_{h|h^\star} < 1$.

We focus on the ratio estimator of a mean since N_h is unknown under stratum misclassification. Under no misclassification, the variance of the ratio estimator can be rewritten as a function of residuals: $Var\left(B_h \cdot \bar{x}_{hp}\right) = \frac{1}{N_h^2} Var\left(\sum_{j=1}^{n_h} w_h e_{hj}\right)$ where $e_{hj} = y_{hj} - B_h x_{hj}$. Therefore, the ratio estimator may be rewritten in the form of a Horvitz-Thompson estimator. Additional details of this derivation are provided in Appendix 3.7. Using the results of Equations 3.1 and 3.2, the expectation and variance of the ratio estimator under stratum misclassification are:

$$\begin{aligned} E\left(\frac{1}{\tilde{N}_h}\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) &= E\left(\frac{1}{\tilde{N}_h}\sum_{h^\star=1}^{H^\star} w_{h^\star}\sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} y_{h|h^\star j}\right) \\ &= \frac{1}{\tilde{N}_h}\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j} = \frac{1}{\tilde{N}_h}\sum_{j=1}^{\tilde{N}_h} y_{hj} \end{aligned} \tag{3.3}$$

$$\begin{aligned} Var\left(\frac{1}{\tilde{N}_h}\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) &= \frac{1}{\tilde{N}_h^2} Var\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} w_{h^\star} e_{h|h^\star j}\right) \\ &= \sum_{h^\star=1}^{H^\star} \frac{N_{h^\star}^2}{n_{h^\star}} \frac{1}{\tilde{N}_h^2}\left(1 - \frac{n_{h^\star}}{N_{h^\star}}\right) \alpha_{h|h^\star} S_{he}^2 \end{aligned} \tag{3.4}$$

where $\tilde{N}_h = \sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h|h^\star} w_{h^\star}$. When $\alpha_{h|h^\star} = 1$, Equation 3.2 and Equation 3.4 are identical, since $\tilde{N}_h = N_h$ and $S_{hy}^2 = S_{he}^2$. As noted previously, it is not clear how $0 < \alpha_{h|h^\star} < 1$ effects the variance in Equation 3.4.

From Equation 3.4, the variance of the difference in means between strata a and b is:

$$Var(\bar{y}_{br}) - Var(\bar{y}_{ar}) = \sum_{h^\star=1}^{H^\star} \frac{N_{h^\star}^2}{n_{h^\star}}\left(1 - \frac{n_{h^\star}}{N_{h^\star}}\right)\left[\frac{S_{be}^2}{\tilde{N}_b^2}\alpha_{b|h^\star} + \frac{S_{ae}^2}{\tilde{N}_a^2}\alpha_{a|h^\star}\right] \tag{3.5}$$

This corresponds to the variance of a linear regression coefficient β_b when stratum a is the referent group.

We perform a simulation to investigate the effects of stratum misclassification on the efficiency of linear regression parameter estimates using Equation 3.5. We assume a population consists of three strata with stratum sizes ($N_1 = 8500, N_2 = 1000, N_3 =$

500). If it is of interest to sample 1000 individuals, then using a disproportionate stratified sampling scheme results in sampling, on average, 333 per stratum. We assume a non-differential, symmetric misclassification mechanism where the diagonal element of Table 3.1 ranges from 0.5 to 1 (no misclassification). We quantify the effect of misclassification on the efficiency of the estimator under no misclassification to that under misclassification. If the sample variances among the three strata are comparable, then we observe the relative efficiency changes presented in Figure 3.1. For the most prevalent subgroup, efficiency gains are observed; otherwise, efficiency losses are observed and the magnitude of the efficiency loss is relative to the sizes of the sampling strata. The second panel in Figure 3.1 summarizes the relative efficiencies when $(N_1 = 5000, N_2 = 4500, N_3 = 500)$. Changes in the stratum means, or variances resulted in similar patterns.

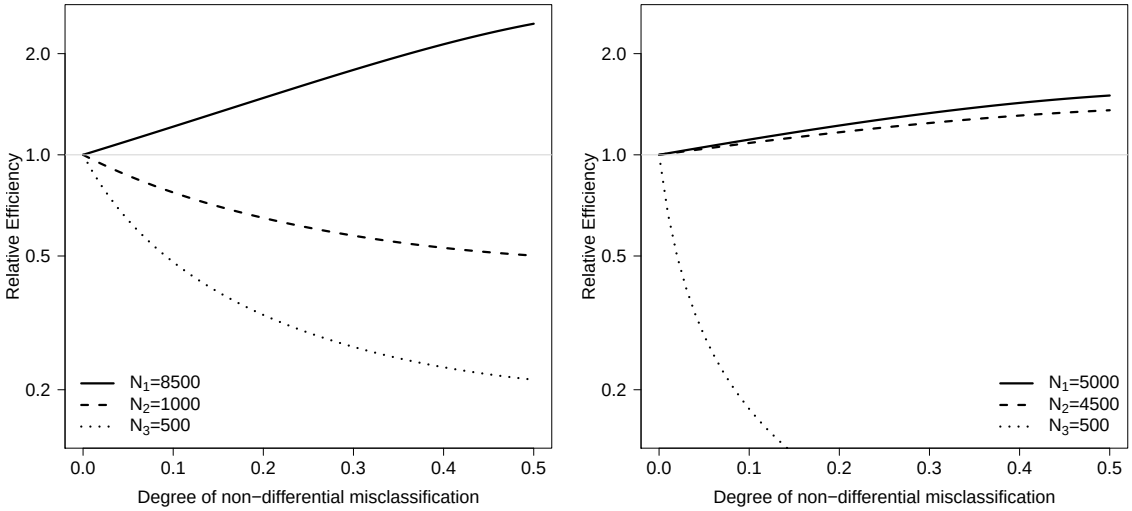


Figure 3.1: Relative efficiencies of design-based linear regression parameter estimates under non-differential symmetric stratum misclassification. Relative efficiency values are defined as the ratio of the variance of the parameter estimate under no misclassification versus that under misclassification, and are computed using Equation 3.5.

3.4 Simulation

We conducted a simulation study to investigate the effects of stratum misclassification on choices of study design and method of analysis. We assume our interest is to estimate the overall prevalence of trust in the healthcare system (trust), and the relationship between trust and race and ethnicity, poverty, and race and ethnicity and poverty. Our sampling frame was defined using only race and ethnicity which

is coded as: Hispanic and non-Hispanic White (reference group), Black, Asian, and Other. Trust was defined as 1 if the respondent answered either "agree" or "strongly agree" to the statement "I trust my healthcare system". Poverty was assigned 1 if self-reported income is less than the number of people in the household x 4,160 + 11,770 (Sanderson et al., 2017). The four regression models of interest include:

$$
\begin{aligned}
logit\left[pr(trust_i)\right] &= \beta_0 && \text{Model 1}\\
logit\left[pr(trust_i|race_i, ethnicity_i)\right] &= \beta_0 + \beta_B I(Black_i) + \beta_A I(Asian_i) \\
&\quad + \beta_O I(Other_i) + \beta_H I(Hispanic_i) && \text{Model 2}\\
logit\left[pr(trust_i|poverty_i)\right] &= \beta_0 + \beta_P I(poverty_i) && \text{Model 3}\\
logit\left[pr(trust_i|race_i, ethnicity_i, poverty_i)\right] &= \beta_0 + \beta_B I(Black_i) + \beta_A I(Asian_i) + \beta_O I(Other_i) \\
&\quad + \beta_H I(Hispanic_i) + \beta_P I(poverty_i) && \text{Model 4}
\end{aligned}
\tag{3.6}
$$

We created a single 'true' population to closely resemble the EHR and survey response dataset from Vanderbilt University Medical Center (VUMC). Self-reported race and ethnicity were estimated using EHR-reported race and ethnicity, along with the misclassification matrix of the respondent population (Table 3.3). For example, if a subject's EHR-reported race and ethnicity was non-Hispanic White, then the predicted probabilities of self-reporting as a non-Hispanic White, Black, Asian, and Other, or Hispanic were 94.3, 0.6, <0.1, <0.1, 9.5 and 0.9%, respectively. The following logistic regression models are used to predict poverty and trust:

$$
\begin{aligned}
logit\left[pr(poverty_i|race_i, ethnicity_i)\right] &= -2.00 + 1.25 I(Black_i) + 0.25 I(Asian_i) \\
&\quad + 1.75 I(Other_i) + 0.50 I(Hispanic_i) \\
logit\left[pr(trust_i|race_i, ethnicity_i, poverty_i)\right] &= -0.75 - 0.25 I(Black_i) - 0.50 I(Asian_i) \\
&\quad + 1.25 I(Other_i) - 1.50 I(Hispanic_i) + 1.00 I(poverty_i)
\end{aligned}
$$

Sub-model parameter estimates, such as the intercept-only model, are those that are induced by marginalizing over race and ethnicity and/or poverty of the full model.

With a complete dataset, misclassified versions of race and ethnicity are constructed by using both non-differential and differential misclassification mechanisms. Two types of non-differential misclassification matrices are utilized: 1) a symmetric matrix (Table 3.1, $MC_{1-\alpha}$; diagonal=α, off-diagonal=$\alpha/4$), and 2) a non-symmetric matrix based on the observed misclassification matrix (Table 3.3; $MC_{\kappa \times obs}$ for $\kappa = 0.5, 1$) where κ times the off-diagonal elements are redistributed to the diagonal of

the corresponding row. We show the differential misclassification mechanism that we explore in Table 3.4 where we stratify the observed misclassification matrix by the observed outcome trust $I(trust = 1)$. A misclassified version of race and ethnicity was then created using the estimated self-reported race and ethnicity and each of the misclassification matrices.

Table 3.3: Misclassification matrix among Vanderbilt University Medical Center respondents. Cell values in brackets respresent row percentages, and those in parentheses denote column percentages. Two degrees, κ, of non-symmetric misclassification are investigated. All non-diagonal elements are multiplied by κ and redistributed to the diagonal row element.

	Self report				
	White	Black	Asian	Other	Hispanic
EHR: $\kappa = 1$					
White	[94.3] (69.0)	[0.6] (0.9)	[0.0] (0.0)	[4.4] (9.5)	[0.6] (0.9)
Black	[0.0] (0.0)	[93.1] (73.6)	[0.0] (0.0)	[5.7] (6.8)	[1.1] (0.9)
Asian	[2.1] (0.9)	[1.0] (0.9)	[75.0] (79.1)	[18.8] (24.3)	[3.1] (2.8)
Other	[22.2] (8.3)	[2.5] (1.8)	[19.8] (17.6)	[43.2] (47.3)	[12.3] (9.3)
Hispanic	[26.7] (21.8)	[14.2] (22.7)	[1.7] (3.3)	[5.1] (12.2)	[52.3] (86.0)
EHR: $\kappa = 0.5$					
White	[97.2] (82.1)	[0.3] (0.5)	[0.0] (0.0)	[2.2] (4.5)	[0.3] (0.4)
Black	[0.0] (0.0)	[96.6] (85.3)	[0.0] (0.0)	[2.9] (3.2)	[0.6] (0.4)
Asian	[1.0] (0.5)	[0.5] (0.5)	[87.5] (89.8)	[9.4] (11.6)	[1.6] (1.1)
Other	[11.1] (4.8)	[1.2] (1.0)	[9.9] (8.6)	[71.6] (74.8)	[6.2] (3.5)
Hispanic	[13.4] (12.6)	[7.1] (12.7)	[0.9] (1.6)	[2.6] (5.8)	[76.1] (94.7)

Table 3.4: Misclassification matrix among Vanderbilt University Medical Center respondents by trust in the healthcare system. Cell values in brackets respresent row percentages, and those in parentheses denote column percentages.

	Self report				
	White	Black	Asian	Other	Hispanic
EHR: Trust=0					
White	[97.7] (70.0)	[0.0] (0.0)	[0.0] (0.0)	[0.0] (0.0)	[2.3] (2.2)
Black	[0.0] (0.0)	[87.5] (75.0)	[0.0] (0.0)	[8.3] (7.7)	[4.2] (2.2)
Asian	[0.0] (0.0)	[0.0] (0.0)	[74.3] (86.7)	[22.9] (30.8)	[2.9] (2.2)
Other	[24.1] (11.7)	[0.0] (0.0)	[10.3] (10.0)	[48.3] (53.8)	[17.2] (11.1)
Hispanic	[19.0] (18.3)	[12.1] (25.0)	[1.7] (3.3)	[3.4] (7.7)	[63.8] (82.2)
EHR: Trust=1					
White	[92.9] (69.1)	[0.9] (1.2)	[0.0] (0.0)	[6.2] (14.6)	[0.0] (0.0)
Black	[0.0] (0.0)	[95.2] (72.8)	[0.0] (0.0)	[4.8] (6.2)	[0.0] (0.0)
Asian	[1.8] (0.7)	[1.8] (1.2)	[75.0] (73.7)	[17.9] (20.8)	[3.6] (3.2)
Other	[21.2] (7.2)	[3.8] (2.5)	[25.0] (22.8)	[40.4] (43.8)	[9.6] (8.1)
Hispanic	[29.9] (23.0)	[15.4] (22.2)	[1.7] (3.5)	[6.0] (14.6)	[47.0] (88.7)

Two types of sampling designs are examined including disproportionate stratified

sampling and random sampling. Disproportionate sampling aims to enrich the final sample by applying unequal sampling probabilities to each sample stratum. One approach to defining these sampling probabilities is by applying the maximum entropy sampling algorithm (see Chapter 4). It aims to identify the number of subjects to sample per stratum such that the Shannon entropy of the stratification information within the sample is maximized, and is achieved when equal number of subjects are sampled from each stratum. We assume that it is of interest to sample 2,500 individuals. Under no misclassifiation and a random sampling design, the expected stratum sizes for the non-Hispanic White, Black, Asian, Other, and Hispanic are 2,056, 243, 27, 126, and 47, respectively. Under disproportionate stratified sampling, we expect to sample 500 from each stratum

We compare design-based and model-based analysis approaches based on the models described previously. Weighted analyses utilized the original design-based weights that are constructed using the misclassified race and ethnicity variables, and are estimated using the `survey` package in `R` (Lumley, 2011). Unweighted analyses ignored the design information (e.g., sampling weights, and finite-population corrections) and are estimated using `R`'s base generalized linear model functions. All weighted sub-models are valid (e.g., accounting for the design via weighting), while only the sub-models that adjusted for race and ethnicity are valid when the study design is ignorable.

For each of the misclassification matrices considered, a total of 2,500 replications of the simulation are performed. Each iteration consisted of generating misclassified 'EHR-reported' race and ethnicity values, identifying the sample population via disproportionate stratified or random sampling, and estimating all regression parameters using design-based and model-based analytic methods. Point estimates, and standard errors are stored for each replicate. Percent bias and coverage probabilities are computed to assess the operating characteristics of each estimation approach. For valid estimators (i.e., unbiased, 95% coverage), relative variances are calculated to compare designs, and analysis approaches under varying degrees of misclassification.

Next, we describe the simulation separately for those scenarios under non-differential, followed by those under differential misclassification.

3.4.1 Non-differential Misclassification

Table 3.5 displays the percent bias and coverage probabilities of the logistic regression parameter estimates for Models 1-4 (Equation 3.6) under random sampling.

Parameter estimates of the full cohort are also provided to aid in the interpretation of the percent bias values. Since the design is ignorable, all models are valid and produce estimates that are unbiased and attained nominal coverage regardless of the degree of misclassification.

Table 3.5: Bias and coverage probability estimates under random sampling. Parameter estimates are provided for the full cohort, and percent biases [coverage probabilities] are reported by type and degree of non-differential misclassification. Estimates are based on 2500 simulations when using model-based analytic methods. Symmetric misclassification mechanisms are denoted as MC_α where α represents the 1-diagonal element of the misclassification matrix. Non-symmetric misclassification matrices, $\text{MC}_{\kappa obs}$, reflect actually observed, or rescaled, misclassification matrices; see Table 3.3.

	Full Cohort	**No Misclass.**	**Symmetric**			**Non-symmetric**	
		MC_0	MC_5	MC_{20}	MC_{30}	$\text{MC}_{0.5obs}$	MC_{obs}
Model 1							
Intercept	0.86	0.4 [94.7]	0.4 [94.7]	-0.1 [94.7]	0.1 [94.7]	0.2 [94.7]	0.2 [94.7]
Model 2							
Intercept (White)	0.85	0.5 [94.6]	0.5 [94.6]	0.0 [94.6]	0.2 [94.6]	0.3 [94.6]	0.2 [94.6]
Black	-0.06	5.9 [94.8]	3.4 [94.8]	-2.4 [94.8]	-2.1 [94.8]	2.9 [94.8]	-4.8 [94.8]
Asian	-0.43	0.5 [95.8]	0.3 [95.8]	-1.3 [95.8]	1.0 [95.8]	-0.6 [95.8]	-2.0 [95.8]
Other	1.49	0.4 [95.4]	0.3 [95.4]	1.3 [95.4]	1.8 [95.4]	1.0 [95.4]	2.1 [95.4]
Hispanic	-1.44	0.2 [95.7]	0.4 [95.7]	0.8 [95.7]	0.7 [95.7]	0.8 [95.7]	0.8 [95.7]
Model 3							
Intercept	0.73	0.3 [94.4]	0.5 [94.4]	-0.1 [94.4]	0.1 [94.4]	0.1 [94.4]	0.1 [94.4]
Poverty	1.03	2.4 [95.6]	1.6 [95.6]	1.3 [95.6]	1.5 [95.6]	2.1 [95.6]	1.7 [95.6]
Model 4							
Intercept (White)	0.75	0.4 [94.4]	0.6 [94.4]	0.0 [94.4]	0.2 [94.4]	0.2 [94.4]	0.1 [94.4]
Black	-0.25	2.4 [95.0]	1.6 [95.0]	0.0 [95.0]	0.3 [95.0]	1.9 [95.0]	-0.2 [95.0]
Asian	-0.46	0.5 [95.3]	0.6 [95.3]	-1.4 [95.3]	1.2 [95.3]	-0.5 [95.3]	-1.8 [95.3]
Other	1.25	0.5 [95.0]	0.3 [95.0]	1.7 [95.0]	2.1 [95.0]	1.1 [95.0]	2.3 [95.0]
Hispanic	-1.53	0.4 [95.3]	0.6 [95.3]	0.9 [95.3]	0.9 [95.3]	0.9 [95.3]	1.0 [95.3]
Poverty	1.00	2.8 [95.7]	1.9 [95.7]	1.6 [95.7]	1.8 [95.7]	2.4 [95.7]	2.0 [95.7]

Tables 3.6 and 3.7 summarize percent bias and coverage probability under disproportionate stratified sampling when performing design-based and model-based analyses, respectively. Design-based analytic methods resulted in unbiased, and nominal coverage, for all model parameter estimates. Similar patterns are observed for Models 2 and 4 when performing a model-based analysis though the large biases observed for the race effect for Black respondents in Model 2 are an artifact of the negligible parameter estimate (i.e., -0.06). Models 2 and 4 adjust for race and ethnicity, and thus adequately account for the study design. Significant biases are observed for marginal effects (Models 1 and 3) which is expected since the design is not acknowledged. As misclassification increases, the magnitude of the bias does decrease (e.g., Model 1: -26.7% to 0%), but inadequate coverage remains.

Table 3.6: Bias and coverage probability estimates under disproportion stratified sampling. Parameter estimates are provided for the full cohort, and percent biases [coverage probabilities] are reported by type and degree of non-differential misclassification. Estimates are based on 2500 simulations when using design-based analytic methods. Symmetric misclassification mechanisms are denoted as MC_α where α represents the percentange misclassified (1-diagonal element of the misclassification matrix). Non-symmetric misclassification matrices, $MC_{\kappa obs}$, reflect actually observed, or rescaled, misclassification matrices; see Table 3.3.

	Full Cohort	**No Misclass.**	**Symmetric**			**Non-symmetric**	
		MC_0	MC_5	MC_{20}	MC_{30}	$MC_{0.5obs}$	MC_{obs}
Model 1							
Intercept	0.86	0.4 [95.3]	0.4 [94.8]	-0.1 [95.0]	0.1 [94.7]	0.2 [95.0]	0.2 [95.2]
Model 2							
Intercept (White)	0.85	0.5 [95.4]	0.5 [94.7]	0.0 [95.3]	0.2 [94.7]	0.3 [95.0]	0.2 [95.0]
Black	-0.06	5.9 [94.8]	3.4 [94.3]	-2.4 [95.6]	-2.1 [94.7]	2.9 [94.5]	-4.8 [95.0]
Asian	-0.43	0.5 [95.4]	0.3 [94.7]	-1.3 [94.6]	1.0 [95.9]	-0.6 [95.0]	-2.0 [94.4]
Other	1.49	0.4 [95.0]	0.3 [94.6]	1.3 [95.6]	1.8 [94.5]	1.0 [94.4]	2.1 [93.2]
Hispanic	-1.44	0.2 [95.0]	0.4 [95.2]	0.8 [94.9]	0.7 [94.2]	0.8 [95.6]	0.8 [95.0]
Model 3							
Intercept	0.73	0.3 [94.9]	0.5 [94.4]	-0.1 [94.9]	0.1 [94.6]	0.1 [94.6]	0.1 [94.4]
Poverty	1.03	2.4 [94.4]	1.6 [94.7]	1.3 [94.4]	1.5 [94.5]	2.1 [94.8]	1.7 [94.5]
Model 4							
Intercept (White)	0.75	0.4 [95.2]	0.6 [94.3]	0.0 [95.2]	0.2 [94.3]	0.2 [94.9]	0.1 [94.5]
Black	-0.25	2.4 [95.1]	1.6 [94.2]	0.0 [94.8]	0.3 [94.7]	1.9 [94.4]	-0.2 [94.5]
Asian	-0.46	0.5 [95.4]	0.6 [94.7]	-1.4 [94.6]	1.2 [95.8]	-0.5 [95.0]	-1.8 [94.3]
Other	1.25	0.5 [94.9]	0.3 [94.5]	1.7 [95.2]	2.1 [94.1]	1.1 [94.0]	2.3 [93.3]
Hispanic	-1.53	0.4 [95.4]	0.6 [95.0]	0.9 [94.6]	0.9 [94.7]	0.9 [95.6]	1.0 [94.7]
Poverty	1.00	2.8 [94.2]	1.9 [94.7]	1.6 [94.2]	1.8 [94.4]	2.4 [94.1]	2.0 [93.8]

Table 3.7: Bias and coverage probability estimates for model-based parameter estimates under disproportion stratified sampling. Parameter estimates are provided for the full cohort, and percent biases [coverage probabilities] are reported by type and degree of non-differential misclassification. Estimates are based on 2500 simulations when using model-based analytic methods. Symmetric misclassification mechanisms are denoted as MC_α where α represents the percentage misclassified (1-diagonal element of the misclassification matrix). Non-symmetric misclassification matrices, $\text{MC}_{\kappa obs}$, reflect actually observed, or rescaled, misclassification matrices; see Table 3.3.

	Full Cohort	**No Misclass.**	**Symmetric**			**Non-symmetric**	
		MC_0	MC_5	MC_{20}	MC_{30}	$\text{MC}_{0.5obs}$	MC_{obs}
Model 1							
Intercept	0.86	-26.7 [0.0]	-11.0 [41.0]	-1.5 [93.3]	0.0 [96.0]	5.8 [80.5]	7.6 [69.8]
Model 2							
Intercept (White)	0.85	0.5 [95.4]	-1.1 [94.9]	-0.5 [94.3]	0.3 [95.6]	1.5 [95.5]	0.9 [95.3]
Black	-0.06	5.9 [94.8]	-16.9 [95.0]	-13.4 [95.6]	-2.5 [95.0]	29.5 [94.9]	10.0 [95.6]
Asian	-0.43	0.5 [96.0]	-1.4 [95.6]	-4.9 [95.1]	1.8 [96.1]	3.9 [95.8]	-7.6 [95.3]
Other	1.49	0.4 [95.3]	1.3 [95.3]	0.7 [95.8]	1.1 [94.9]	-1.8 [94.9]	-0.6 [95.2]
Hispanic	-1.44	0.2 [95.8]	-1.0 [95.4]	0.3 [95.4]	0.5 [95.2]	2.5 [95.8]	2.0 [95.3]
Model 3							
Intercept	0.73	-45.2 [0.0]	-23.2 [5.3]	-6.8 [81.0]	-3.6 [92.0]	-2.5 [93.9]	1.3 [95.0]
Poverty	1.03	10.2 [86.8]	7.7 [91.0]	2.1 [95.1]	3.1 [95.1]	7.5 [93.2]	5.4 [94.6]
Model 4							
Intercept (White)	0.75	1.1 [95.2]	-1.2 [95.0]	-0.3 [95.1]	0.3 [96.0]	1.8 [95.0]	1.1 [95.2]
Black	-0.25	-2.6 [95.1]	-7.0 [95.4]	-4.9 [95.0]	-0.1 [94.7]	5.9 [94.7]	-0.4 [95.8]
Asian	-0.46	0.0 [96.2]	-1.8 [95.4]	-4.7 [94.9]	1.9 [96.0]	3.1 [95.9]	-8.1 [95.1]
Other	1.25	1.5 [95.0]	2.2 [95.3]	1.2 [96.3]	1.3 [94.9]	-1.4 [94.8]	-0.2 [95.6]
Hispanic	-1.53	-0.3 [95.9]	-1.2 [95.2]	0.2 [95.2]	0.6 [95.3]	2.0 [95.8]	1.2 [95.2]
Poverty	1.00	-5.9 [92.4]	-2.1 [94.4]	-2.1 [94.4]	0.7 [95.7]	-2.1 [95.1]	-2.8 [94.3]

Figure 3.2 and Table 3.8 summarize the relative efficiencies of logistic regression parameter estimates under varying degrees of non-differential misclassification comparing disproportionate stratified sampling to random sampling. We estimate relative efficiency (RE) of a regression model parameter as 100 times the average variance under random sampling divided by average variance under stratified sampling. Values greater than one indicate that the stratified sampling approach is more efficient (i.e., smaller variance) than that under random sampling.

Regardless of the analysis method, utilizing a disproportionate stratified sampling scheme resulted in more efficient estimation of design parameters for rare subgroups ($RE > 1$). Up to 10-times as many Asians are sampled when using the stratified design compared to random sampling under no stratum misclassification. Among non-Hispanic Whites, the stratified design resulted in less precise point estimates compared to random sampling due to the drastic differences in stratum sizes. As misclassification increased, the efficiency gains associated with stratified sampling decreased. Even under 30% misclassification, disproportionate stratified sampling results in a more diverse sample than under random sampling indicating that this design should still be performed if the goal is to learn about rare subgroups. For non-design variables (poverty), efficiency gains are observed when using model-based estimation which is likely due to the correlation between poverty and race/ethnicity. Since race and ethnicity was used to construct the study design, we expect sample enrichment on variables related to these design variables.

Table 3.8: Relative efficiencies of logistic regression parameter estimates under non-differential stratum misclassification comparing disproportionate and random sampling. Estimates in brackets correspond to unweighted analyses; values are omitted for Models 1 and 3 since these models did not adjust for the design variables. Degree of symmetric non-differential misclassification ranged from 5% misclassified (MC_5; or 95% correctly classified) to 30% misclassified, and degree of non-symmetric non-differential misclassification included $MC_{0.5obs}$=the observed non-differential misclassification matrix where the off-diagonal elements are reduced by 0.5, and MC_{obs}=the observed misclassification matrix.

$\frac{Var(\mathbf{rs})}{Var(\mathbf{ss})}$	**No Misclassification**	**Symmetric**			**Non-symmetric**	
	MC_0	MC_5	MC_{20}	MC_{30}	$MC_{0.5obs}$	MC_{obs}
Model 1						
Intercept	0.29 []	0.32 []	0.42 []	0.51 []	0.39 []	0.51 []
Model 2						
Intercept (White)	0.24 [0.24]	0.27 [0.50]	0.36 [0.76]	0.45 [0.84]	0.33 [0.58]	0.44 [0.67]
Black	1.15 [1.14]	1.14 [1.57]	1.09 [1.55]	1.08 [1.43]	1.38 [1.74]	1.36 [1.81]
Asian	10.11 [9.47]	5.42 [7.57]	2.36 [3.16]	1.80 [2.25]	6.54 [8.07]	4.19 [5.56]
Other	3.25 [3.18]	2.72 [3.21]	1.89 [2.31]	1.54 [1.85]	2.12 [3.32]	1.64 [2.92]
Hispanic	5.69 [5.47]	3.95 [5.24]	2.11 [2.82]	1.58 [2.06]	1.57 [1.74]	0.95 [1.17]
Model 3						
Intercept	0.28 []	0.31 []	0.41 []	0.50 []	0.38 []	0.50 []
Poverty	0.36 []	0.40 []	0.51 []	0.60 []	0.48 []	0.61 []
Model 4						
Intercept (White)	0.25 [0.25]	0.27 [0.52]	0.37 [0.78]	0.46 [0.86]	0.33 [0.59]	0.45 [0.68]
Black	1.10 [1.16]	1.09 [1.57]	1.07 [1.54]	1.06 [1.42]	1.32 [1.73]	1.32 [1.80]
Asian	10.06 [9.51]	5.37 [7.57]	2.35 [3.16]	1.79 [2.24]	6.46 [8.08]	4.18 [5.57]
Other	3.06 [3.16]	2.60 [3.18]	1.84 [2.29]	1.52 [1.84]	2.06 [3.28]	1.62 [2.89]
Hispanic	5.42 [5.51]	3.81 [5.23]	2.07 [2.81]	1.57 [2.06]	1.54 [1.75]	0.95 [1.18]
Poverty	0.35 [1.51]	0.38 [1.35]	0.50 [1.19]	0.59 [1.12]	0.46 [1.27]	0.59 [1.23]

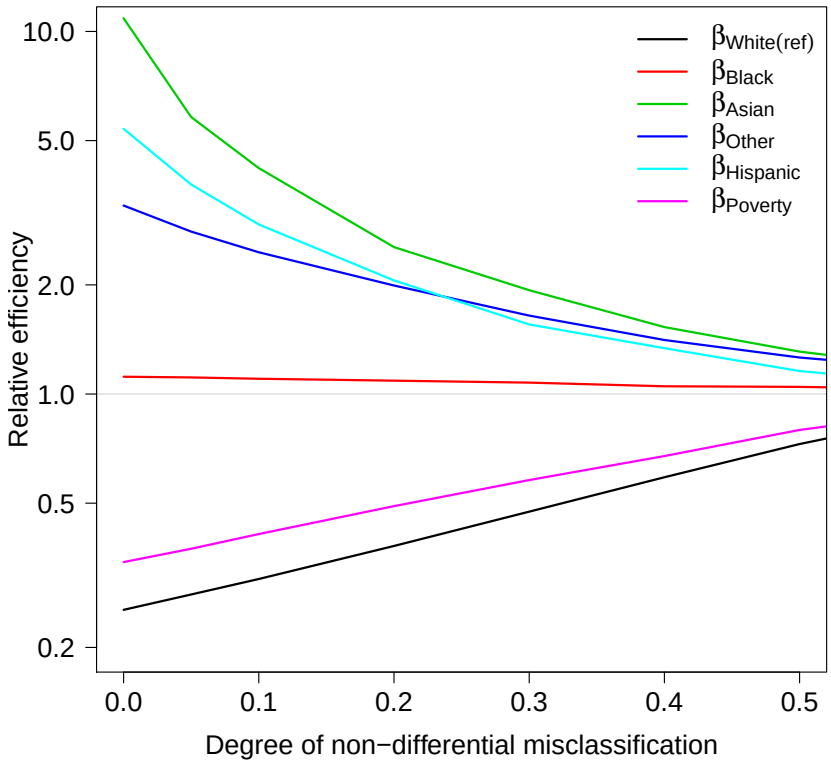


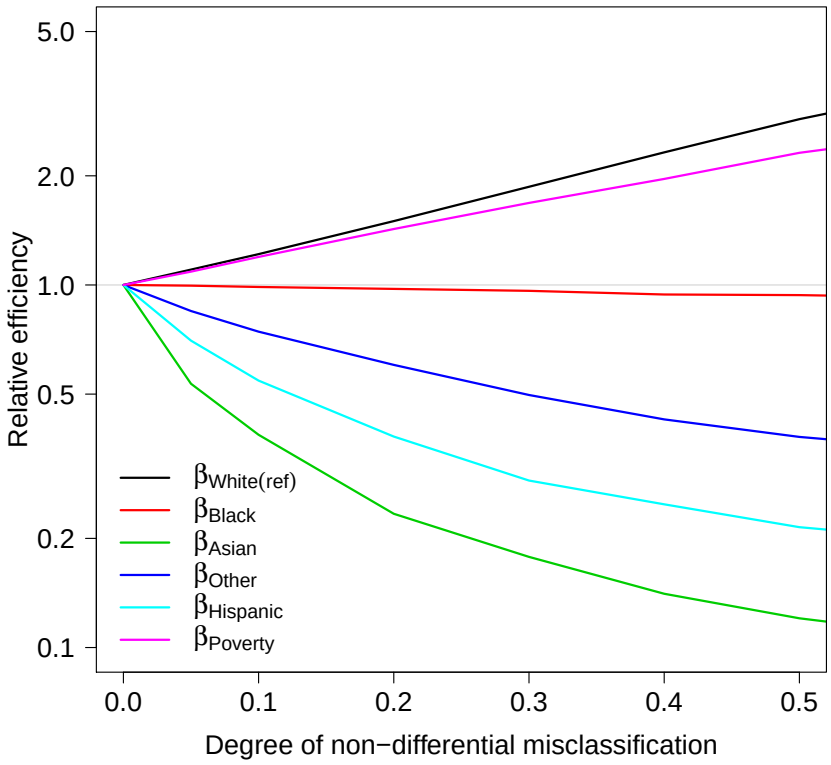
Figure 3.2: Relative efficiencies of design-based logistic regression parameter estimates under non-differential symmetric stratum misclassification comparing disproportionate and random sampling. Misclassification ranged from 0 (100% correctly classified, no misclassification) to 0.5 (50% correctly classified).

Figure 3.3 and Table 3.9 summarize the relative efficiencies of logistic regression parameter estimates under non-differential stratum misclassification comparing varying degrees of misclassification to no misclassification. The relative efficiency is estimated as 100 times the average variance under no misclassification divided by average variance under misclassification. For design effects that are associated with small subgroups (e.g., non-Hispanic Asian, Hispanic), efficiency is greatly reduced when design variables are misclassified - even by a "minimal" amount (5%, or 95% correctly classified). Conversely, efficiency gains are observed for covariate effects summarizing larger subgroups (e.g., non-Hispanic Whites, $RE > 2$; non-Hispanic Blacks, $RE > 1.25$) and those not explicitly included in the design of the study. As misclassification increased, so did the efficiency loss (gain) for small (large) subgroups.

Table 3.9: Relative efficiencies of logistic regression parameter estimates under non-differential stratum misclassification comparing varying degrees of misclassification to no misclassification. Estimates in brackets correspond to unweighted analyses; values are omitted for Models 1 and 3 since these models did not adjust for the design variables. Degree of symmetric non-differential misclassification ranged from 5% misclassified (MC_5; or 95% correctly classified) to 30% misclassified, and degree of non-symmetric non-differential misclassification included $MC_{0.5obs}$=the observed non-differential misclassification matrix where the off-diagonal elements are reduced by 0.5, and MC_{obs}=the observed misclassification matrix.

$\frac{Var(\mathbf{MC}_0)}{Var(\mathbf{MC}_x)}$	**Symmetric**			**Non-symmetric**	
	MC_5	MC_{20}	MC_{30}	$MC_{0.5obs}$	MC_{obs}
Model 1					
Intercept	1.10 []	1.46 []	1.78 []	1.34 []	1.75 []
Model 2					
Intercept (White)	1.10 [2.08]	1.50 [3.15]	1.87 [3.49]	1.37 [2.38]	1.83 [2.75]
Black	0.99 [1.37]	0.95 [1.36]	0.94 [1.25]	1.20 [1.52]	1.18 [1.58]
Asian	0.54 [0.80]	0.23 [0.33]	0.18 [0.24]	0.65 [0.85]	0.41 [0.59]
Other	0.84 [1.01]	0.58 [0.73]	0.48 [0.58]	0.65 [1.04]	0.51 [0.92]
Hispanic	0.69 [0.96]	0.37 [0.51]	0.28 [0.38]	0.28 [0.32]	0.17 [0.21]
Model 3					
Intercept	1.10 []	1.46 []	1.79 []	1.35 []	1.76 []
Poverty	1.09 []	1.41 []	1.66 []	1.32 []	1.66 []
Model 4					
Intercept (White)	1.10 [2.05]	1.50 [3.07]	1.87 [3.38]	1.36 [2.34]	1.82 [2.69]
Black	1.00 [1.35]	0.98 [1.33]	0.96 [1.22]	1.21 [1.49]	1.21 [1.55]
Asian	0.53 [0.80]	0.23 [0.33]	0.18 [0.24]	0.64 [0.85]	0.42 [0.59]
Other	0.85 [1.01]	0.60 [0.73]	0.50 [0.58]	0.67 [1.04]	0.53 [0.91]
Hispanic	0.70 [0.95]	0.38 [0.51]	0.29 [0.37]	0.28 [0.32]	0.17 [0.21]
Poverty	1.09 [0.89]	1.43 [0.78]	1.68 [0.74]	1.33 [0.84]	1.69 [0.81]

Figure 3.3: Relative efficiencies of logistic regression parameter estimates for Model 4 under non-differential stratum misclassification comparing varying degrees of misclassification to no misclassification. Misclassification ranged from 0 (100% correctly classified, no misclassification) to 0.5 (50% correctly classified).

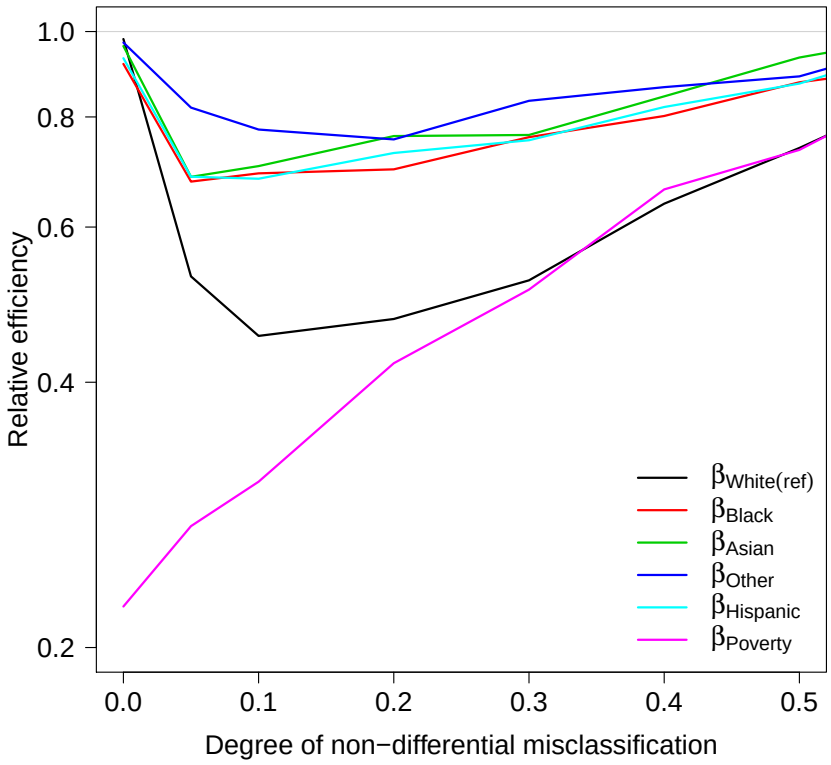


Relative efficiencies of logistic regression parameter estimates under non-differential stratum misclassification comparing design-based (weighted) to model-based (unweighted) analysis approaches are presented in Table 3.10 and Figure 3.4. Marginal models (1 and 3) are not presented because the model-based analytic approaches did not adequately account for the study design. Under no misclassification, the efficiency of both analytic approaches is comparable for the design variables. This implies that under a 100% response rate, weighting for the design and adjusting for the design variables simultaneously does not inflate the variance of model estimates. For non-design variables, design-based approaches are less efficient. Similarly, as misclassification increased, so did the design-based variance estimates of all model parameters due to the additional variation in the sampling weights. Therefore, the bias-variance tradeoff needs to be considered when choosing an approach to analyzing complex survey data.

Table 3.10: Relative efficiencies of logistic regression parameter estimates under non-differential stratum misclassification comparing design-based (weighted) to model-based (unweighted) analysis approaches. Models 1 and 3 are omitted since the unweighted models did not adjust for the design variables. Degree of symmetric non-differential misclassification ranged from 5% misclassified (MC_5; or 95% correctly classified) to 30% misclassified, and degree of non-symmetric non-differential misclassification included $MC_{0.5obs}$=the observed non-differential misclassification matrix where the off-diagonal elements are reduced by 0.5, and MC_{obs}=the observed misclassification matrix.

$\frac{Var(\textbf{unweighted})}{Var(\textbf{weighted})}$	**No Misclassification**	**Symmetric**			**Non-symmetric**	
	MC_0	MC_5	MC_{20}	MC_{30}	$MC_{0.5obs}$	MC_{obs}
Model 2						
Intercept (White)	1.00	0.53	0.48	0.54	0.57	0.67
Black	1.01	0.73	0.71	0.75	0.79	0.75
Asian	1.07	0.72	0.75	0.80	0.81	0.75
Other	1.02	0.85	0.82	0.84	0.64	0.56
Hispanic	1.04	0.75	0.75	0.76	0.90	0.81
Model 4						
Intercept (White)	0.97	0.52	0.47	0.54	0.56	0.66
Black	0.94	0.69	0.69	0.74	0.76	0.74
Asian	1.06	0.71	0.74	0.80	0.80	0.75
Other	0.97	0.82	0.80	0.83	0.63	0.56
Hispanic	0.99	0.73	0.74	0.76	0.88	0.80
Poverty	0.23	0.28	0.42	0.53	0.37	0.48

Figure 3.4: Relative efficiencies of logistic regression parameter estimates for Model 4 under non-differential stratum misclassification comparing design-based (weighted) to model-based (unweighted) analysis approaches. Misclassification ranged from 0 (100% correctly classified, no misclassification) to 0.5 (50% correctly classified).



3.4.2 Differential Misclassification

Table 3.11 displays the percent bias and coverage probabilities of the logistic regression parameter estimates for Models 1-4 (Equation 3.6) under random and stratified sampling. Parameter estimates of the full cohort are also provided to aid in the interpretation of the percent bias values. Valid estimates are obtained when data is collected under random sampling. When analyzing complex survey data in the presence of differential misclassification, only design-based approaches produced unbiased estimates with nominal coverage for all model estimates. Model-based analyses produced biased point estimates for small subgroups (e.g., non-Hispanic Asians, 30%).

Table 3.11: Bias and coverage probability estimates under differential misclassification as described in Table 3.4. Parameter estimates are provided for the full cohort, and percent biases [coverage probabilities] are reported by study design. Estimates are based on 2500 simulations when using both design- and model-based analytic methods.

	Full Cohort	**Random Sampling**	**Stratified Sampling**	
		Model-based	Design-based	Model-based
Model 1				
Intercept	0.86	0.3 [94.7]	0.3 [94.5]	4.6 [87.3]
Model 2				
Intercept (White)	0.85	0.4 [94.6]	0.4 [94.7]	0.9 [95.0]
Black	-0.06	4.5 [94.8]	4.5 [95.2]	-35.9 [95.6]
Asian	-0.43	0.7 [95.8]	0.7 [94.4]	32.4 [86.4]
Other	1.49	0.8 [95.4]	0.8 [94.0]	-20.6 [56.3]
Hispanic	-1.44	1.1 [95.7]	1.1 [95.5]	-2.2 [95.6]
Model 3				
Intercept	0.73	0.3 [94.4]	0.3 [94.2]	-2.7 [92.8]
Poverty	1.03	1.7 [95.6]	1.7 [94.6]	5.7 [93.2]
Model 4				
Intercept (White)	0.75	0.4 [94.4]	0.4 [94.3]	1.3 [95.0]
Black	-0.25	2.2 [95.0]	2.2 [95.8]	-9.9 [95.1]
Asian	-0.46	0.9 [95.3]	0.9 [94.3]	30.0 [87.0]
Other	1.25	0.9 [95.0]	0.9 [94.1]	-24.7 [58.0]
Hispanic	-1.53	1.2 [95.3]	1.2 [95.6]	-2.0 [95.7]
Poverty	1.00	2.1 [95.7]	2.1 [94.4]	-1.2 [94.6]

3.5 Example: CERC survey

To investigate the effects of stratum misclassification on survey weight analytic summaries in a real-world setting, we analyze a subset of the data from the eMERGE CERC survey. The primary goal of this survey is to understand the factors associated with an individual's willingness to participate in a biobank. Secondary objectives included the identification of factors associated with overall trust in the healthcare

system, trust in medical researchers, and concern about the privacy of health information. Details regarding the survey development, the sampling strategy, and the results have been published previously (Smith et al., 2016; Sanderson et al., 2017). For the present analysis, only data from Vanderbilt University Medical Center (VUMC) is explored, and two simplifications are made for ease of exposition. First, 69 of the 687 respondents are omitted because their EHR race or ethnicity is missing. In the original study, these data were imputed using geocoded-derived census data. Second, we define a five-level sampling frame using only race and ethnicity, whereas the original sampling frame consists of 288 levels using age, gender, race, ethnicity, education, and rural living. Due to the low response rate (16%), this simplification is necessary to perform adjusted regression analyses that account for the study design. Finally, our analysis focuses the demographic factors associated with an individual's trust in the healthcare system.

Table 3.12 summarizes respondent sample at VUMC that had complete EHR race and ethnicity information. We include unweighted observed percentages (counts) and the survey-adjusted estimates. Only 36% of the respondents in our sample self-identified as non-Hispanic White, and 17% provided information that indicated they lived below the poverty line (income, number of individuals living in the household), which indicates that the disproportionate sampling scheme based on race and ethnicity enriched the sample - and thus the respondent sample. The survey weighted estimates of these demographics estimate the population to which the respondent population generalizes.

Table 3.12: Demographics of the Vanderbilt University Medical Center CERC respondent population, and the associated survey-weighted population. Percentages (counts) are provided for each variable.

	Unweighted	**Weighted**
	N=618	N=326515
Gender		
Female	56 (336)	54 (173177)
Male	44 (268)	46 (150041)
Age group		
18-35	26 (156)	27 (86738)
36-50	21 (126)	14 (44399)
51-64	28 (169)	31 (99210)
65+	24 (142)	27 (87224)
Race/ethniciy		
White	36 (216)	83 (268372)
Black	18 (110)	9 (29734)
Asian	15 (91)	1 (2999)
Other	12 (74)	5 (15408)
Hispanic	18 (107)	2 (5561)
Poverty		
No	83 (456)	85 (246747)
Yes	17 (96)	15 (44033)

Table 3.13 presents regression results by analysis method (columns; design-based, and model-based) and by race and ethnicity definition (rows; EHR, and self-report). Three sets of comparisons are made using these results: 1) comparisons of analytic approaches under no stratum misclassification by using models that adjust for EHR defined race and ethnicity, 2) comparisons of analytic approaches under differential stratum misclassification by using models that adjust for self-reported race and ethnicity, and 3) evaluation of the effects of misclassification on design-based estimators by comparing models that use EHR data and those that use self-report data.

Under no stratum misclassification, valid estimates are obtained using both analytic methods. Efficiency losses typically associated with design-based approaches are not observed when weighting and adjusting for the design variables simultaneously. Discrepancies between design- and model-based estimates of the poverty effect in Model 4 (estimate [95% confidence interval], 1.15 [0.03, 2.27] vs 0.31[-0.19, 0.83]) indicate that this model does not adequately account for the study design and additional interactions between race/ethnicity and poverty are needed.

Under differential stratum misclassification, model-based approaches do not adequately account for the design since EHR data are not utilized. Valid estimates are only obtained by performing a design based analysis since the study design is acknowledged via weighting.

Design-based analytic methods result in valid parameter estimates using either EHR or self-reported race and ethnicity data. Observed differences in these estimates are due to either differences in variable coding or stratum misclassification. For example, the definitions of the "Other non-Hispanic" subgroup in the EHR database differed from that of the survey instrument, and thus is not explore further. The EHR-Hispanic effect is due to stratum misclassification since 21.8% of EHR-reported non-Hispanic Whites self-identified as Hispanic (Table 3.3). The EHR Hispanic effect is a weighted combination of the effects of Hispanics and non-Hispanic Whites resulting in an overall effect not significantly different than non-Hispanic Whites.

Based on the above observations, we interpret only the design-based results that adjust for self-reported race and ethnicity. The overall prevalence of healthcare system trust in the VUMC population is 71% (95% CI: 64-77%; Model 1). The odds of an individual trusting their healthcare system among those living below the poverty line is 3-times that of those that do not (OR=3.0, 1.0-9.1; Model 3). When compared, Hispanics are 88% less likely to trust their healthcare system compared to non-Hispanic Whites (OR: 0.2, 95% CI: 0.1-0.6; Model 2). There was not sufficient evidence to conclude that other racial and ethnic groups differed from the referent

group. Similar effects are observed while simultaneously accounting for race/ethnicity and poverty (Hispanics: 0.2, 0.1-0.6; poverty: 2.9, 0.9-9.1; Model 4).

Table 3.13: Design- and model-based logistic regression analyses in which trust in the healthcare system was regressed on: 1) intercept only, 2) race/ethnicity, 3) poverty, and 4) race/ethnicity and poverty. Design weights were defined using EHR race/ethnicity. Race/ethnicity was also collected at the time of the survey, self-report. Both definitions of race/ethnicity were used in Models 2 and 4. For each analysis approach, point estimates [standard errors] are presented.

	Design-based (weighted)				**Model-based (unweighted)**			
	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
EHR								
Intercept (White)	0.91[0.16]	0.94[0.18]	0.78[0.17]	0.82[0.19]	0.68[0.09]	0.94[0.18]	0.62[0.10]	0.90[0.18]
Black		-0.08[0.29]		-0.28[0.30]		-0.08[0.29]		-0.14[0.29]
Asian		-0.61[0.27]		-0.66[0.28]		-0.61[0.28]		-0.63[0.28]
Other		-0.40[0.27]		-0.36[0.28]		-0.40[0.29]		-0.39[0.29]
Hispanic		-0.29[0.24]		-0.37[0.24]		-0.29[0.24]		-0.32[0.24]
Poverty			1.11[0.56]	1.15[0.57]			0.32[0.26]	0.31[0.27]
Self report								
Intercept (White)	0.91[0.16]	0.89[0.18]	0.78[0.17]	0.79[0.19]	0.68[0.09]	0.87[0.15]	0.62[0.10]	0.83[0.15]
Black		0.14[0.30]		-0.03[0.32]		0.12[0.26]		0.05[0.26]
Asian		-0.48[0.30]		-0.49[0.31]		-0.31[0.27]		-0.33[0.27]
Other		1.57[0.55]		1.34[0.57]		-0.35[0.28]		-0.39[0.28]
Hispanic		-1.49[0.54]		-1.62[0.52]		-0.61[0.25]		-0.67[0.25]
Poverty			1.11[0.56]	1.07[0.58]			0.32[0.26]	0.37[0.27]

3.6 Discussion

This paper investigates the effect of utilizing an imperfect sampling frame on the planning and analysis of a complex study design. Motivated by the eMERGE CERC survey, which constructed a sampling frame using both EHR and census data, we explored the impact of stratum misclassification on the choice of study design, on the operating characteristics of descriptive and analytic summaries, and on the appropriateness of two common approaches to survey design analysis. Under the misclassification scenarios considered, disproportionate stratified sampling is recommended over random sampling if interest lies in making inferential statements regarding less prevent subgroups. The efficiency gains typically observed when using this design are typically dampened in the presence of misclassification, except for prevalent subgroups where efficiency gains are observed. If a complex study design is executed, then accounting for the design during the analysis phase is still required. For the design to be ignorable, a significant amount of misclassification must be observed, but rarely would one use such highly mis-measured variables to define the sampling frame.

Two common approaches to the analysis of complex survey data include design-based (weighting) and model-based (covariate adjustment) analyses. Rooted in the frequentist versus Bayesian controversy, the choice of which to use is not straightforward. Weighting approaches are typically less efficient, while adjustment methods may not adequately account for the design (e.g., lack of interactions) (Lin et al., 2014). In the presence of non-differential stratum misclassification, design-based analytic summaries tended to have desirable operating characteristics (e.g., unbiased, nominal coverage) for all covariate effects - even those not explicitly used in design of the study. Model-based analyses produced more precise point estimates than design-based methods. For marginal effects, these estimates were biased, but this bias did decrease as misclassification increased. Under differential misclassification, the design is informative, and must be accounted for valid inferences. In our simulation study, design-based methods were robust to the misclassification mechanism, whereas model-based methods could not adequately adjust for the design leading to significant biases.

There are several limitations of the current analysis that warrant consideration. We only consider the scenario when the design information is recollected at the time of the survey. If the sampling frame consists only of demographics, like the eMERGE survey, then including these items in the survey is not a significant burden, and may aid in better understanding the characteristics of the sample population. It is also

assumed that these self-reported data are true. The demographics considered are not sensitive in nature, thus this assumption seems reasonable. Finally, only two analysis methods of survey data are considered. Numerous extensions are available, including hierarchical Bayesian and calibrated Bayesian methods (Lin et al., 2014; Gelman, 2007). The methods investigated are the most common and are a justifiable starting point when investigating the effects of stratum misclassification on analytic summaries.

The eMERGE survey response rate at Vanderbilt University Medical Center was 16%. Due to the significant possibility of non-response bias, all analyses must be interpreted with caution. Since the survey was designed to enrich the sample population with under-studied subpopulations, the observed response rate may be an artifact of the design. If survey response is related to the demographics used to design the study, then non-response likely induces an informative design with a complex misclassification mechanism. Coding differences between the design and survey-instrument variables complicate the interpretation of the misclassification matrix, and all results based on this matrix (e.g., Simpon's Paradox). For example, the race item in the eMERGE survey included 'more than 1 race'. This option was not available in the EHR dataset and needed to be aggregated into the 'other' category to estimate misclassification probabilities. Misclassification adjustment methods that are based solely on the matrix are sensitive to these coding differences (Kuha and Skinner, 1997).

3.7 Appendix

The following derivation is used to derive the expectation and variance for the Horvitz-Thompson estimator of a total (t_h) under stratum misclassification (Equations 3.1 and 3.2).

$$
\begin{aligned}
E\left(t_h\right) &= E\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) = E\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) = \sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} E(I_{h^\star j}) I_{h|h^\star j} w_{h^\star} y_{h|h^\star j} \\
&= \sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} \frac{1}{w_{h^\star}} w_{h^\star} I_{h|h^\star j} y_{h|h^\star j} = \sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j} = \sum_{j=1}^{N_h} y_{hj} = t_h \\
Var\left(t_h\right) &= Var\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{n_{h^\star}} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) = Var\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) = Var\left(\sum_{h^\star=1}^{H^\star}\sum_{j=1}^{N_{h^\star}} I_{h^\star j} I_{h|h^\star j} w_{h^\star} y_{h|h^\star j}\right) \\
&= Var\left(\sum_{j=1}^{N_1} I_{1j} I_{h|1j} w_1 y_{h|1j} + \ldots + \sum_{j=1}^{N_{H^\star}} I_{H^\star j} I_{h|H^\star j} w_{H^\star} y_{h|H^\star j}\right) \\
&= Var\left(\sum_{j=1}^{N_1} I_{1j} I_{h|1j} w_1 y_{h|1j}\right) + \ldots + Var\left(\sum_{j=1}^{N_{H^\star}} I_{H^\star j} I_{h|H^\star j} w_{H^\star} y_{h|H^\star j}\right) \\
&= Cov\left(\sum_{j=1}^{N_1} I_{1j} I_{h|1j} w_1 y_{h|1j}, \sum_{k=1}^{N_1} I_{1k} I_{h|1k} w_1 y_{h|1k}\right) + \ldots + Cov\left(\sum_{j=1}^{N_{H^\star}} I_{H^\star j} I_{h|H^\star j} w_{H^\star} y_{h|H^\star j}, \sum_{k=1}^{N_{H^\star}} I_{H^\star k} I_{h|H^\star k} w_{H^\star} y_{h|H^\star k}\right) \\
&= \sum_{j=1}^{N_1}\sum_{k=1}^{N_1} Cov\left(I_{1j} I_{h|1j} w_1 y_{h|1j}, I_{1k} I_{h|1k} w_1 y_{h|1k}\right) + \ldots + \sum_{j=1}^{N_{H^\star}}\sum_{k=1}^{N_{H^\star}} Cov\left(I_{H^\star j} I_{h|H^\star j} w_{H^\star} y_{h|H^\star j}, I_{H^\star k} I_{h|H^\star k} w_{H^\star} y_{h|H^\star k}\right) \\
&= \sum_{j=1}^{N_1}\sum_{k=1}^{N_1} I_{h|1j} I_{h|1k} w_1^2 y_{h|1j} y_{h|1k} Cov\left(I_{1j}, I_{1k}\right) + \ldots + \sum_{j=1}^{N_{H^\star}}\sum_{k=1}^{N_{H^\star}} I_{h|H^\star j} I_{h|H^\star k} w_{H^\star}^2 y_{h|H^\star j} y_{h|H^\star k} Cov\left(I_{H^\star j}, I_{H^\star k}\right) \\
&= \sum_{j=1}^{N_1} I_{h|1j} w_1^2 y_{h|1j}^2 Var\left(I_{1j}\right) + \sum_{j=1}^{N_1}\sum_{k\neq j}^{N_1} I_{h|1j} I_{h|1k} w_1^2 y_{h|1j} y_{h|1k} Cov\left(I_{1j}, I_{1k}\right) + \ldots + \\
&\quad \sum_{j=1}^{N_{H^\star}} I_{h|H^\star j} w_{H^\star}^2 y_{h|H^\star j}^2 Var\left(I_{H^\star j}\right) + \sum_{j=1}^{N_{H^\star}}\sum_{k\neq j}^{N_{H^\star}} I_{h|H^\star j} I_{h|H^\star k} w_{H^\star}^2 y_{h|H^\star j} y_{h|H^\star k} Cov\left(I_{H^\star j}, I_{H^\star k}\right)
\end{aligned}
$$

Note:

For subject's j and k in stratum $h^\star \in (1, \ldots, H^\star)$, the $Var\left(I_{h^\star j}\right) = \frac{n_{h^\star}}{N_{h^\star}}\left(1 - \frac{n_{h^\star}}{N_{h^\star}}\right)$, $Cov\left(I_{h^\star j}, I_{h^\star k}\right) = -\frac{n_{h^\star}}{N_{h^\star}}\left(\frac{1}{N_{h^\star}-1}\right)\left(1 - \frac{n_{h^\star}}{N_{h^\star}}\right)$ and $w_{h^\star} = \frac{N_{h^\star}}{n_{h^\star}}$.

The $h^\star$ element of $Var\left(t_h\right)$ can be rewritten as [minus the $w_{h^\star}^2$ term]:

$$
\begin{aligned}
Var\left(t_h;h^\star\right) &= \sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2 Var\left(I_{h^\star j}\right) + \sum_{j=1}^{N_{h^\star}} \sum_{k\neq j}^{N_{h^\star}} I_{h|h^\star j} I_{h|h^\star k} y_{h|h^\star j} y_{h|h^\star k} Cov\left(I_{h^\star j}, I_{h^\star k}\right) \\
&= \sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2 \left[\frac{n_{h^\star}}{N_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\right] + \sum_{j=1}^{N_{h^\star}} \sum_{k\neq j}^{N_{h^\star}} I_{h|h^\star j} I_{h|h^\star k} y_{h|h^\star j} y_{h|h^\star k} \left[-\frac{n_{h^\star}}{N_h{}^\star}\left(\frac{1}{N_{h^\star}-1}\right)\left(1-\frac{n_{h^\star}}{N_{h^\star}}\right)\right] \\
&= \frac{n_{h^\star}}{N_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\left[\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2 - \left(\frac{1}{N_{h^\star}-1}\right) \sum_{j=1}^{N_{h^\star}} \sum_{k\neq j}^{N_{h^\star}} I_{h|h^\star j} I_{h|h^\star k} y_{h|h^\star j} y_{h|h^\star k}\right] \\
&= \frac{n_{h^\star}}{N_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\left[\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2 - \left(\frac{1}{N_{h^\star}-1}\right) \sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j} \sum_{k=1}^{N_{h^\star}} I_{h|h^\star k} y_{h|h^\star k} + \left(\frac{1}{N_{h^\star}-1}\right) \sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2\right] \\
&= \frac{n_{h^\star}}{N_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\left[\left(1+\frac{1}{N_{h^\star}-1}\right) \sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2 - \left(\frac{1}{N_{h^\star}-1}\right)\left(\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}\right)^2\right] \\
&= \frac{n_{h^\star}}{N_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\frac{N_h{}^\star}{N_h{}^\star-1}\left[\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}^2 - \frac{1}{N_{h^\star}}\left(\sum_{j=1}^{N_{h^\star}} I_{h|h^\star j} y_{h|h^\star j}\right)^2\right] \\
&= \frac{n_{h^\star}}{N_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\frac{N_h{}^\star}{N_h{}^\star-1}\left[N_{h^\star h} S_h^2 + \left(\frac{1}{N_{h^\star h}} - \frac{1}{N_{h^\star}}\right) N_{h^\star h}^2 \bar{y}_{ph}^2\right] \\
&= n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\frac{N_{h^\star h}}{N_h{}^\star-1} S_h^2 + n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\frac{N_{h^\star h}}{N_h{}^\star-1}\left(\frac{N_{h^\star}-N_{h^\star h}}{N_{h^\star}}\right)\bar{y}_{ph}^2 \\
&= n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\frac{\alpha_{h|h^\star h} N_{h^\star}}{N_h{}^\star-1} S_h^2 + n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\frac{\alpha_{h|h^\star} N_{h^\star}}{N_h{}^\star-1}\left(\frac{N_{h^\star}-\alpha_{h|h^\star} N_{h^\star}}{N_{h^\star}}\right)\bar{y}_{ph}^2 \\
&= n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\alpha_{h|h^\star} S_h^2 + n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\alpha_{h|h^\star}(1-\alpha_{h|h^\star})\bar{y}_{ph}^2 \\
&= n_{h^\star}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\alpha_{h|h^\star}\left[S_h^2 + (1-\alpha_{h|h^\star})\bar{y}_{ph}^2\right]
\end{aligned}
$$

Therefore, $Var\left(t_h\right) = \sum_{h^\star=1}^{H^\star} \frac{N_{h^\star}^2}{n_{h^\star}}\left(1-\frac{n_{h^\star}}{N_h{}^\star}\right)\alpha_{h|h^\star}\left[S_h^2 + (1-\alpha_{h|h^\star})\bar{y}_{ph}^2\right]$

The previous derivation is used to derive the expectation and variance for the ratio estimator of a mean under stratum misclassification (Equations 3.3 and 3.4). We first derive the relationship between the variance of the ratio estimator of $\bar{y}_{hr}$ and the Horvitz-Thompson estimator of $\sum_{j=1}^{n_h} w_h e_{hj}$.

Under no stratum misclassification, the mean squared error of $\bar{y}_{hr}$ is

$$
\begin{aligned}
E\left[\left(\bar{y}_{hr}-\bar{y}_{hp}\right)^2\right] &= E\left[\left(\frac{\bar{y}_h}{\bar{x}_h}\bar{x}_{hp} - B_h\bar{x}_{hp}\right)^2\right] = E\left[\left(\frac{\bar{x}_{hp}}{\bar{x}_h}(\bar{y}_h - B_h\bar{x}_h)\right)^2\right] \\
&= E\left[\left((\bar{y}_h - B_h\bar{x}_h)\left(1-\frac{\bar{x}_h-\bar{x}_{hp}}{\bar{x}_h}\right)\right)^2\right] \approx E\left[\left(\bar{y}_h - B_h\bar{x}_h\right)^2\right] \\
&\approx Var\left[\bar{y}_h - B_h\bar{x}_h\right] = Var\left[\bar{e}_h\right] = \frac{1}{N_h^2} Var\left[\frac{N_h}{n_h}\sum_{j=1}^{n_h} e_{jh}\right] = \frac{1}{N_h^2} Var\left[\sum_{j=1}^{n_h} w_h e_{jh}\right]
\end{aligned}
$$

In our setting, $E\left[\left(\bar{y}_{hr}-\bar{y}_{hp}\right)^2\right] = Var\left(\bar{y}_{hr}\right) = \frac{1}{N_h^2} Var\left[\sum_{j=1}^{n_h} w_h e_{jh}\right]$.

Under stratum misclassification,

$$Var\left(\bar{y}_{hr}\right) = \frac{1}{N_h^2} Var\left[\sum_{j=1}^{n_h} w_h e_{jh}\right] = \frac{1}{N_h^2} \sum_{h^\star=1}^{H^\star} \frac{N_{h^\star}^2}{n_{h^\star}^2} n_{h^\star} \left(1 - \frac{n_{h^\star}}{N_h{}^\star}\right) \frac{N_{h^\star}}{N_h{}^\star - 1} \alpha_{h|h^\star} \left[S_{he}^2 + (1 - \alpha_{h|h^\star}) \bar{e}_{ph}^2\right]$$

$$= \frac{1}{N_h^2} \sum_{h^\star=1}^{H^\star} \frac{N_{h^\star}^2}{n_{h^\star}^2} n_{h^\star} \left(1 - \frac{n_{h^\star}}{N_h{}^\star}\right) \frac{N_{h^\star}}{N_h{}^\star - 1} \alpha_{h|h^\star} \left[S_{he}^2\right]$$

since $\bar{e}_{ph} = 0$. N_h is unknown and is estimated as $\tilde{N}_h = \sum_{h^\star=1}^{H^\star} \sum_{j=1}^{N_{h^\star}} I_{h|h^\star} w_{h^\star}$.

CHAPTER 4

MARGINALIZED MODELS FOR LONGITUDINAL BINARY DATA: THE `MMLB` R PACKAGE

4.1 Abstract

The `MMLB` package is introduced and examples are provided to demonstrate how to estimate parameters from marginalized regression models for longitudinal binary data. Estimation of model parameters is described when data are collected prospectively under random sampling, and under a class of outcome dependent sampling (ODS) designs. Using data from the Madras Longitudinal Schizophrenia Study, we demonstrate how this package can be used to fit three types of marginalized regression models, including: the marginalized latent variable model, the marginalized transition model, and the marginalized latent variable and transition model. Examples are provided to show how `MMLB` functions may be used to generate longitudinal binary outcomes under a pre-specified marginal model, and to demonstrate how it is used to estimate marginalized model parameters under single- and two-stage ODS sampling designs.

4.2 Introduction

Longitudinal binary data are commonplace in the health sciences. Be it monitoring the presence of delirium in the ICU (Pandharipande et al., 2008) or describing the association between depression and asthma (Brunner et al., 2014), the desire to answer questions that quantify temporal changes between- or within-subjects, or aid in establishing causal relationships between a set of covariates and a binary outcome, is of interest. Unlike univariate analysis, where observations are assumed to be conditionally independent given observed covariates, the analysis of longitudinal data must acknowledge that observations are correlated within participants over time.

To address scientific questions while acknowledging within-subject dependence, two classes of models are often considered: conditional and marginal mean models (Zeger et al., 1988). Conditional mean models explain the dependence structure by explicitly incorporating additional subject-specific terms into the specification of the mean model, such as response history (transition models) or latent characteristics (latent variable models) (Diggle et al., 2002; Breslow and Clayton, 1993; Stiratelli et al., 1984). That is, all moments are captured in a single regression model. In

contrast, marginal mean models do not explicitly account for subject characteristics, but explain within-subject dependence either by jointly specifying a marginal mean and a marginal association measure (e.g., correlations or odds-ratios (Molenberghs and Verbeke, 2005)) or by separating the estimation of the mean from the higher order moments (Schildcrout and Heagerty, 2007). The target of inference should drive the model choice. The target of inference is the same for both mean models when using the identity link and a mean zero random effect, and model choice may be determined by other factors (e.g., modeling assumptions, computational availability). When the outcome is binary, inferential targets differ, due to Jensens' inequality, between these modeling approaches and thus the choice of model is not straightforward leading some to debate the appropriateness of each modeling approach (Zeger et al., 1988; Neuhaus et al., 1991; Lindsey and Lambert, 1998; Lee and Nelder, 2004)).

In this paper, we focus on marginalized regression models which separate the estimation of the mean from higher-order moments. This permits the estimation of population-level mean parameters, while allowing for a wide-range of dependence structures commonly encountered in the health sciences (Schildcrout and Heagerty, 2007). Additionally, we discuss how these models may be applied when analyzing data collected from a biased sampling design. A review of marginalized regression models and outcome-dependent sampling designs is presented in Section 2. In Section 3, estimation of marginalized model parameters is described. Syntax of key `MMLB` functions are introduced in Section 4. Examples using the Madras Longitudinal Schizophrenia Study and simulated data are demonstrated in Section 5. Finally, in Section 6 we describe future directions of the `MMLB` software.

4.3 Models

Marginalized regression models are defined by a pair of regression models that fully specify the multivariate distribution of binary outcome vector given the observed design matrix (Schildcrout and Heagerty, 2007; Heagerty, 1999, 2002). First, a marginal mean model is constructed to relate covariates to the logit-transformed probability of the binary outcome. To capture second and higher order moments, an association model (also called a conditional mean or dependence model) is defined to characterize serial and/or long-range dependence structures. The marginalized transition and latent-variable model (mTLV; (Schildcrout and Heagerty, 2007)) allows the specification of both types of dependence structures simultaneously. Special cases of the mTLV model include: the marginalized logistic normal model (mLV; (Heagerty,

1999)) which incorporates only a latent-term, such as a random intercept, and the a first-order marginalized transition model (mT; (Azzalini, 1994; Heagerty, 2002)) which acknowledges a single transition term in the specification of the dependence model.

To define the mTLV model, let Y_{ij} denote the binary outcome of subject i at observation j where $i = \{1, 2, \ldots, N\}$ and $j = \{1, 2, \ldots, n_i\}$. Let $\boldsymbol{X}_i$ denote a $n_i \times p$ design matrix, $\boldsymbol{X}_{ij}$ the corresponding p-dimensional design vector at time j and $\boldsymbol{\beta}^m$ the p-dimensional vector of parameters. Then, the marginal mean and dependence models are defined as:

$$\text{logit}\left(\mu_{ij}^m\right) = \boldsymbol{X}_{ij}\boldsymbol{\beta}^m \tag{4.1}$$

$$\text{logit}\left(\mu_{ij}^c\right) = \Delta_{ij} + \gamma(\boldsymbol{X}_i)Y_{ij-1} + b_i \quad \text{where } b_i \sim N\left(0, \sigma^2(\boldsymbol{X}_i)\right) \tag{4.2}$$

Δ_{ij} is the value that relates the marginal and conditional means via the convolution equation

$$\mu_{ij}^m = \int_{\boldsymbol{A}_{ij}} \mu_{ij}^c dF_{\boldsymbol{A}_{ij}} = \int_{\boldsymbol{A}_{ij}} \text{logit}^{-1}\left(\Delta_{ij} + \boldsymbol{A}_{ij}\boldsymbol{\alpha}\right) dF_{\boldsymbol{A}_{ij}} \tag{4.3}$$

where $\boldsymbol{A}_{ij}$ and $\boldsymbol{\alpha}$ denote the design matrix of the dependence model and the stacked parameter vector $(\gamma(\boldsymbol{X}_i), \sigma(\boldsymbol{X}_i))$, respectively (Schildcrout and Heagerty, 2007). For the remainder of this paper, we assume that the association model parameters are not modified by covariates. This implies that $\gamma(\boldsymbol{X}_i) = \gamma$ and $b_i \sim N(0, \sigma^2)$ where the latter can be rewritten as σZ_i where $Z_i \sim N(0, 1)$.

Estimates of the mTLV mean model parameters, $\boldsymbol{\beta}^m$, are interpreted as the difference in the log-odds of having the outcome associated with a unit change in the covariate of interest between two populations whose covariates are otherwise identical. By not explicitly accounting for subject-specific quantities, marginal mean models permit the straightforward interpretation of both time-varying and time-invariant covariates effects. Dependence model parameters characterize the magnitude of the variation in the log-odds between individuals within a group defined by their observed covariates (Heagerty, 1999).

Through the specification of both a marginal mean and dependence model, a full likelihood is defined. Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}^m, \boldsymbol{\alpha}\}$, then under random sampling, subject i's

contribution to the likelihood function is defined as:

$$L_i(\boldsymbol{\theta}|\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{X}_i = \boldsymbol{x}_i) = pr(\boldsymbol{y}_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \int_{z_i} \left[\prod_{j=1}^{n_i} {\mu_{ij}^c}^{y_{ij}} \left(1 - {\mu_{ij}^c}^{y_{ij}}\right)^{1-y_{ij}} \right] \phi(z_i) dz_i. \tag{4.4}$$

where ϕ denotes the standard normal distribution. This permits the likelihood-based estimation of model parameters and the application of frequentist, Bayesian or likelihoodist inferential paradigms, the comparison of non-nested models (model selection), the computation of individual-level predictions, and the reliance on the less stringent missing data assumptions as compared to semi-parametric approaches (e.g., GEE; MAR vs MCAR) (Heagerty, 1999; Laird, 1988). The effects of model misspecification on summaries of the mTLV model have been explored previously (Schildcrout and Heagerty, 2007). It was demonstrated that valid point estimates of $\boldsymbol{\beta}^m$ are obtained under either functional form or dependence model misspecifications, but their associated standard errors are sensitive to incorrect dependence model formulations, such as ignoring the modifying impact of a cluster-level covariate. Considering this, the `MMLB` software may be used to estimate marginalized model parameters, and robust standard errors, using weighted estimating equations (Robins et al., 1995).

4.3.1 Outcome-Dependent Sampling Designs

Outcome-dependent sampling (ODS) is a class of retrospective sampling schemes that are designed to enrich a sampled population with individuals that are the most informative (**?**Song et al., 2009; Zhou et al., 2013). The most notable ODS design for univariate binary response data is the case-control study whereby an individual's sampling probability is dependent on their response status (Anderson, 1972; Prentice and Pyke, 1979). Neuhaus and Jewell (1990) extended the case-control design to accommodate correlated binary response data by conditioning on the sum of an individual's response vector, and modeling the exposure-response relationship using a random-intercept logistic model.

We consider the class of ODS designs where each individual is stratified into one of three groups based on their response vector (Schildcrout and Heagerty, 2011). Let $V_i = g(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ denote the stratum membership for subject i. The three sampling strata are defined as: those that did not experience the event of interest (non-responders, $\sum_j Y_{ij} = 0$; $V_i = 0$), those that exhibited response variation (any-responders, $0 < \sum_j Y_{ij} < n_i$; $V_i = 1$), and those that only experienced the event

(all-responders, $\sum_j Y_{ij} = n_i$; $V_i = 2$). Since $S_i \perp\!\!\!\perp (\boldsymbol{Y}_i, \boldsymbol{X}_i)|V_i$, the corresponding sampling probabilities are defined as

$$pr(S_i = 1|\boldsymbol{y}_i, \boldsymbol{x}_i) = pr(S_i = 1|V_i = v_i) \equiv \pi(v_i) \tag{4.5}$$

An ODS design is defined as either a pre-specified triplet of sampling probabilities, $\pi_1(0)$,$\pi_1(1)$, and $\pi_2(2)$, or as the sampling probabilities that are induced from a pre-specified vector of expected stratum sample sizes n_{10}, n_{11} and n_{12}. For ease of interpretation, we define the ODS design as $D[n_{10}, n_{11}, n_{12}]$. Via Bayes' theorem, subject i's contribution to the conditional, or ascertainment-corrected, likelihood is

$$\begin{aligned} L_i^c(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i, S_i = 1) = pr(\boldsymbol{y}_i|\boldsymbol{x}_i, S_i = 1; \boldsymbol{\theta}) &= \frac{pr(S_i = 1|\boldsymbol{y}_i, \boldsymbol{x}_i) pr(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta})}{pr(S_i = 1|\boldsymbol{x}_i)} \\ &= \frac{\pi(v_i)}{pr(S_i = 1|\boldsymbol{x}_i)} \cdot pr(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}) \equiv \frac{\pi(v_i)}{AC_i} \cdot L_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i) \end{aligned} \tag{4.6}$$

where $S_i = 1$ denotes a sampling indicator, AC_i (ascertainment correction) $= \sum_{v_i=0}^{2} \pi(v_i) \cdot pr(V_i = v_i|\boldsymbol{x}_i)$, and $L_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i)$ is defined in Equation 4.4.

In Chapter 2 the standard, or single-stage, ODS design is extended to both fixed, and adaptive two-stage designs. These designs allow for the combined analysis of data collected in two waves (e.g., internal pilot and main study), as well as "conditionally optimal" designs that utilize stage one information to inform how to choose the optimal stage two design. For a fixed two-stage ODS design, the first stage sampling probabilities, and likelihood contributions (Equations 4.5 and 4.6), are identical to the single-stage quantities, but the second stage sampling probabilities, π_2, are defined as:

$$\begin{aligned} pr(S_{2i} = 1|\boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) &\equiv pr(S_{2i} = 1, S_{1i} = 0|\boldsymbol{y}_{2i}, \boldsymbol{y}_{2i}, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \\ &= pr(S_{2i} = 1|S_{1i} = 0, \boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \cdot pr(S_{1i} = 0|\boldsymbol{Y}_{2i}, \boldsymbol{x}_{2i}) \\ &= \pi_2(v_i; \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \cdot [1 - \pi_1(v_i)] \end{aligned} \tag{4.7}$$

where $S_{2i}, \boldsymbol{y}_{2i}$, and $\boldsymbol{x}_{2i}$ denote subject i's stage two sampling indictor, response vector, and observed design matrix. Let $\boldsymbol{S}_1$ denote the $N \times 1$ vector of stage one sampling indicators, and $\boldsymbol{y}_1$ and $\boldsymbol{x}_1$ represent the response vector, and design matrix for those sampled at stage one, respectively. Note, $\boldsymbol{x}_1$ contains both the original design matrix, as well as the newly collected exposure information. Even in a fixed two-stage ODS design, the second stage sampling probabilities are conditional on the data from the first stage, since the sampling is performed without replacement. The corresponding

contribution to the ascertainment-corrected likelihood is:

$$\begin{aligned} L^c_{2i}(\boldsymbol{\theta}|\boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) &= pr(\boldsymbol{y}_{2i}|\boldsymbol{x}_{2i}, S_{2i} = 1, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \\ &= \frac{\pi_2(v_i; \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1)[1 - \pi_1(v_i)]}{pr(S_{2i} = 1|\boldsymbol{x}_{2i}, \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1)} \cdot pr(\boldsymbol{y}_{2i}|\boldsymbol{x}_{2i}) \\ &\equiv \frac{\pi_2(v_i; \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1)[1 - \pi_1(v_i)]}{AC_{2i}} \cdot L_{2i}(\boldsymbol{\theta}|\boldsymbol{y}_{2i}, \boldsymbol{x}_{2i}) \end{aligned} \tag{4.8}$$

where $AC_{2i} = \sum_{v_i=0}^{2} \pi_2(v_i; \boldsymbol{y}_1, \boldsymbol{x}_1, \boldsymbol{S}_1 = 1) \cdot [1 - \pi_1(v_i)] \cdot pr(V_i = v_i|\boldsymbol{x}_{1i})$.

4.4 Estimation

4.4.1 Maximum Likelihood Estimation

Assuming that subjects are independent, the likelihood for $\boldsymbol{\theta}$ in the mTLV model given the data is the product of each individual's likelihood contribution:

$$L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=1}^{N} L_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i) = \prod_{i=1}^{N} \int_{z_i} \prod_{j=1}^{n_i} {\mu^c_{ij}}^{y_{ij}} (1 - \mu^c_{ij})^{1-y_{ij}} \phi(z_i) dz_i \equiv \prod_{i=1}^{N} \int_{z_i} L_{i,z_i} \phi(z_i) dz_i. \tag{4.9}$$

The corresponding score equation for parameter $\theta \in \boldsymbol{\theta}$ is:

$$\frac{\partial}{\partial \theta} \log L(\theta|\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{N} \frac{\int_{z_i} L_{i,z_i} \left[\sum_{j=1}^{n_i} \left(y_{ij} - \mu^c_{ij}\right) \frac{\partial}{\partial \theta} \left(\Delta_{ij} + \gamma y_{ij-1} + \sigma z_i\right)\right] \phi(z_i) dz_i}{\int_{z_i} L_{i,z_i} \phi(z_i) dz_i} \tag{4.10}$$

The likelihood for subject i requires the calculation of Δ_{ij} for all $j \in (1, n_i)$. Using Equation 4.3, it can be show that:

$$\begin{aligned} \mu^m_{ij} &= \int_{z_i} \mu^{pc,z_i}_{ij} \phi(z_i)\, dz_i \\ &\equiv \int_{z_i} \left[\text{logit}^{-1}(\Delta_{ij} + \sigma z_i) \left[1 - \mu^{pc,z_i}_{ij-1}\right] + \text{logit}^{-1}(\Delta_{ij} + \gamma + \sigma z_i)\, \mu^{pc,z_i}_{ij-1}\right] \phi(z_i)\, dz_i \end{aligned} \tag{4.11}$$

where μ^{pc,z_i}_{ij} denotes the partially conditioned mean for subject i at time j. The term partially conditioned is used to reflect that the integrand corresponds to the expectation of the conditional mean over the lagged response distribution.

Since Δ_{ij} is analytically intractable, it is estimated iteratively using the Newton-Raphson algorithm. Let $f(\Delta_{ij}) = \int_{z_i} \mu^{pc,z_i}_{ij}(\Delta_{ij}) \phi(z_i)\, dz_i - \mu^m_{ij} = 0$. Using a first-

order Taylor's series approximation of $f\left(\Delta_{ij}\right)$, the estimated value of Δ_{ij} at iteration $k+1$ is defined as $\Delta_{ij}^{(k+1)} = \Delta_{ij}^{(k)} - \frac{f\left(\Delta_{ij}^{(k)}\right)}{f'\left(\Delta_{ij}^{(k)}\right)}$ where

$$f'\left(\Delta_{ij}^{(k)}\right) = \int_{z_i} \left[\left[1 - \mu_{ij-1}^{pc,z_i}\right] \frac{\partial}{\partial \Delta_{ij}^{(k)}} \text{logit}^{-1}\left(\Delta_{ij}^{(k)} + \sigma z_i\right) + \mu_{ij-1}^{pc,z_i} \frac{\partial}{\partial \Delta_{ij}^{(k)}} \text{logit}^{-1}\left(\Delta_{ij}^{(k)} + \gamma + \sigma z_i\right) \right] \phi\left(z_i\right) dz_i \tag{4.12}$$

From Equation 4.11, it can be seen that estimating Δ_{ij} requires an estimate of μ_{ij-1}^{pc,z_i}. We sequentially estimate each element of $\boldsymbol{\Delta}_i = vec(\Delta_{i0}, \Delta_{i1}, \ldots, \Delta_{in_i})$ by first assuming that $\hat{\mu}_{i0}^{pc,z_i} = 0$ and calculating $\hat{\Delta}_{i1}$, and then use $\hat{\Delta}_{i1}$ to estimate $\hat{\mu}_{i1}^{pc,z_i}$, and so on.

Guassian-Hermite quadrature is used to approximate all integrals (e.g., Equation 4.9). This numerical integration technique is used to approximate integrals of the form $\int_{-\infty}^{\infty} e^{-x^2} f(x)\, dx$ with $\sum_{i=1}^{q} w_i f(x_i)$ where q, x_i and w_i denote the number of quadrature points, abscissa (or nodes) and weights, respectively. Let $h(z)$ denote an integrand of interest, say $\mu_{ij}^{pc,z}$, then $E_z\left[h(z)\right] = \int_{-\infty}^{\infty} h(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$. Let $x = \frac{z}{\sqrt{2}}$, then $z = \sqrt{2}x$ and $E_z\left[h(z)\right] = \int_{-\infty}^{\infty} h\left(\sqrt{2}x\right) \frac{1}{\sqrt{\pi}} e^{-x^2} dx \approx \sum_{i=1}^{q} \frac{1}{\sqrt{\pi}} w_i h\left(\sqrt{2}x_i\right) = \sum_{i=1}^{q} w_i' h\left(x_i'\right)$ where w_i' and x_i' denote scaled weights and abscissa.

Under (single stage) outcome-dependent sampling, the ascertainment-corrected likelihood is:

$$L^c(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=1}^{N^s} L_i^c(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i) = \prod_{i=1}^{N^s} \frac{\pi(v_i)}{AC_i} \cdot L_i(\boldsymbol{\theta}|\boldsymbol{y}_i, \boldsymbol{x}_i) \tag{4.13}$$

The corresponding score equation for parameter $\theta \in \boldsymbol{\theta}$ is:

$$\frac{\partial}{\partial \theta} \log L^c(\theta|\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{N^s} -\frac{1}{AC_i} \frac{\partial}{\partial \theta} AC_i + \frac{\partial}{\partial \theta} \log L_i(\theta|\boldsymbol{y}_i, \boldsymbol{x}_i) \tag{4.14}$$

where $\frac{\partial}{\partial \theta} AC_i = \frac{pr(V_i=0|\boldsymbol{x}_i)}{\partial \theta}[\pi(0) - \pi(1)] + \frac{pr(V_i=2|\boldsymbol{x}_i)}{\partial \theta}[\pi(2) - \pi(1)]$ and $\frac{\partial}{\partial \theta} \log L_i(\theta|\boldsymbol{y}_i, \boldsymbol{x}_i)$ is of the form defined in Equation 4.10.

Similarly, the combined two-stage ascertainment-corrected likelihood is:

$$L^c(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{S}=1) = pr(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{S}=1) = pr(\boldsymbol{y}_1|\boldsymbol{x}_1,\boldsymbol{S}_1=1)\cdot pr(\boldsymbol{y}_2|\boldsymbol{x}_2,\boldsymbol{S}_2=1,\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1=1)$$
$$= \left[\prod_{i=1}^{N_1^s}\frac{\pi_1(v_i)}{AC_{1i}}\cdot L_{1i}(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{x})\right]\cdot\left[\prod_{i=N_1^s+1}^{N_1^s+N_2^s}\frac{\pi_2(v_i;\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1=1)[1-\pi_1(v_i)]}{AC_{2i}}\cdot L_{2i}(\boldsymbol{\theta}|\boldsymbol{y}_{2i},\boldsymbol{x}_{2i})\right] \tag{4.15}$$

where N_1^s denotes the number of subjects sampled in stage one, and $N_2^s \equiv N_2^s(\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1 = 1)$ denotes the number of subjects sampled in stage two.
The corresponding score equation for parameter $\theta \in \boldsymbol{\theta}$ is:

$$\frac{\partial}{\partial\theta}\log L^c(\theta|\boldsymbol{y},\boldsymbol{x}) = \left[\sum_{i=1}^{N_1^s} -\frac{1}{AC_{1i}}\frac{\partial}{\partial\theta}AC_{1i} + \frac{\partial}{\partial\theta}\log L_{1i}(\theta|\boldsymbol{y}_{1i},\boldsymbol{x}_{1i})\right]$$
$$+\left[\sum_{i=N_1^s+1}^{N_1^s+N_2^s} -\frac{1}{AC_{2i}}\frac{\partial}{\partial\theta}AC_{2i} + \frac{\partial}{\partial\theta}\log L_{2i}(\theta|\boldsymbol{y}_{2i},\boldsymbol{x}_{2i})\right] \tag{4.16}$$

where $\frac{\partial}{\partial\theta}AC_{1i} = \frac{\partial}{\partial\theta}pr(V_i=0|\boldsymbol{x}_{1i})[\pi_1(0)-\pi_1(1)] + \frac{\partial}{\partial\theta}pr(V_i=2|\boldsymbol{x}_{1i})[\pi_1(2)-\pi_1(1)]$, $\frac{\partial}{\partial\theta}AC_{2i} = \frac{\partial}{\partial\theta}pr(V_i=0|\boldsymbol{x}_{2i})[\pi_2(0;\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1=1)[1-\pi_1(0)] - \pi_2(1;\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1=1)[1-\pi_1(1)]] + \frac{\partial}{\partial\theta}pr(V_i=2|\boldsymbol{x}_{2i})[\pi_2(2;\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1=1)[1-\pi_1(2)] - \pi_2(1;\boldsymbol{y}_1,\boldsymbol{x}_1,\boldsymbol{S}_1=1)[1-\pi_1(1)]]$ and $\frac{\partial}{\partial\theta}\log L_{ki}(\theta|\boldsymbol{y}_{ki},\boldsymbol{x}_{ki})$ is of the form defined in Equation 4.10 for $k = 1, 2$.

4.4.2 Weighted Estimating Equations

An alternative estimation approach to maximum likelihood is weighted estimating equations (WEE) (Robins et al., 1995; Cai et al., 2001). The WEE estimator of $\boldsymbol{\theta}$ is defined as the solution to the system of equations defined by $\sum_i^{N^s} U_i^w(\boldsymbol{\theta}) = 0$ where $U_i^w(\boldsymbol{\theta}) = \frac{1}{\pi(v_i)}\frac{\partial}{\partial\boldsymbol{\theta}}\log[L_i(\boldsymbol{\theta}|\boldsymbol{y}_i,\boldsymbol{x}_i)]$ and N^s denotes the total number of sampled subjects (e.g., $N^s \equiv N_1^s + N_2^s$ if performing a two-stage ODS design). The covariance matrix is of the form $A^{-1}B(A^{-1})'$ where $A = E\left[\frac{\partial}{\partial\boldsymbol{\theta}}U_i^w(\boldsymbol{\theta})\right]$ and $B = E\left[U_i^w(\boldsymbol{\theta})U_i^w(\boldsymbol{\theta})'\right]$.

4.5 `MMLB` Syntax

The `mm` function is used to fit marginalized models for longitudinal binary data. Table 4.1 summarizes each argument of `mm`. Regardless of the sampling design, or the estimation method, `mm` requires the specification of a marginal mean model (Equation 4.1, `mean.formula`), the right-hand side of the association model (Equation 4.2, `lv.formula` and/or `t.formula`), a cluster identifier, `id`, and a data object, `data`, which is assumed to be sorted by id and time.

Three notable arguments of the `mm` function include `cond.like`, `samp.probs`, and `samp.probi`. Under an ODS design, `cond.like` is set to `TRUE` to indicate that the ascertainment-corrected likelihood should be maximized. ODS sampling probabilities are supplied to this function using the `samp.probs` argument, and depending on the type of ODS design (single or two-stage), it is either vector of size 3, or an nrow(data) x 3 matrix. Model parameters may be estimated via weighted estimating equations by assigning a vector of subject-specific sampling probabilities to the `samp.probi` argument.

The remaining `mm` arguments pertain to estimation. Gauss-Hermite quadrature is used to approximate all integrals defined in Section 4.4. Weights and abscissa are calculated using the `gaussHermiteData` function from the `fastGHQuad` `R` package (Blocker, 2014). `MMLB`'s internal `get.GH` function is a wrapper for the `gaussHermiteData` and returns appropriately scaled abscissa and weights for a given `q` value. `R`'s `nlm` function is used to calculate the maximum likelihood estimates of $\boldsymbol{\theta}$. The `inits`, `step.max`, `step.tol`, `hess.eps`, and `iter.lim` arguments are all optional arguments for this function, and additional details are provided in the `nlm` help file.

Fitting a model using `mm` results in the creation of an object of class `MMLongit`. Basic summary and extraction functions for `MMLongit` objects, such as `print`, `summary`, `coef`, and `vcov`, have been incorporated into the `MMLB` package. Since marginalized models are defined by a pair of regression models, the `coef`, and `vcov` functions return a list estimates by model. Detailed descriptions of `mm`'s output are described in Table 4.2.

Table 4.1: MMLB's mm() arguments

mean.formula	mean model formula in which a binary variable is regressed on covariates
lv.formula	latent variable component of the dependence model (right hand side only)
t.formula	transition component of the association model (right hand side only)
id	a vector of cluster identifiers
data	a required data frame, ordered by id and time
cond.like	logical, if the ascertainment-corrected likelihood should be maximized
samp.probs	if analyzing data from an outcome-dependent sampling design, then this argument is either a vector of 3 values (single-stage) or a matrix (nrow(data) $\times$ 3) that denote the sampling probability of non-responders, any-responders, and all-responders.
samp.probi	a vector of sampling probabilities if using weighted estimating equations
inits	an optional list of length 3 containing initial values for marginal mean parameters and all dependence parameters. The format of the list should be: (1) estimates of the mean parameters, (2) estimates of the transition parameters (or NULL if only fitting a mLV model) and (3) estimates of the latent variable parameters (or NULL if only fitting a mT model). If NULL, initial values will be automatically generated.
offset	an optional offset term; typically used when maximizing a profile likelihood
q	a scalar to denote the number of quadrature points used for Gauss-Hermite numerical integration
step.max	a scalar
step.tol	a scalar
hess.eps	a scalar
iter.lim	a scalar to denote the maximum iteration limit
return_args	logical, if key arguements should be returned post model fit (only used when performing indirect imputation)
verbose	logical, if model output should be printed to the screen during fitting process

Table 4.2: MMLB's mm() output

call	mm function call
logLik	estimate of minimum log-likelihood
beta	a vector mean model parameter estimates; equivalent to coef(fit)$beta
alpha	a vector association model parameter estimates; equivalent to coef(fit)$alpha
mod.cov	estimated model-based covariance matrix; see also vcov(fit)$beta and vcov(fit)$alpha
rob.cov	estimated robust, or sandwich, covariance matrix
control	a vector of summaries used in summary.MMLongit()
info_stats	a vector of information criteria (AIC, BIC, logLik, Deviance); used in summary.MMLongit()
LogLikeSubj	a vector of subject-specific contributions to the log-likelihood
ObsInfoSubj	list of subject-specific contributions to the observed information matrix
ACSubj	a vector of subject-specific log-transformed ascertainment corrections

The `GenBinaryY` function is used to generate binary response data under a marginalized regression model. The outcome, Y_{ij}, for subject i at time j, is generated from a Bernoulli distribution where the probability of success is defined as the inverse-logit of the conditional mean, μ_{ij}^c. The calculation of μ_{ij}^c is not straightforward since it is a function of Δ_{ij}. The arguments of `GenBinaryY` are summarized in Table 4.3. This function returns the entire `data` object augmented with the new binary outcome labeled `Yname`.

Table 4.3: `MMLB`'s `GenBinaryY()` arguments

`mean.formula`	mean model formula (right side only)
`lv.formula`	latent variable component of the dependence model (right hand side only)
`t.formula`	transition component of the association model (right hand side only)
`beta`	a vector of values for `mean.formula`
`sigma`	a vector of values for the latent variable portion of the association model
`gamma`	a vector of values for the transition porition of the association model
`id`	a vector of cluster identifiers
`data`	a required data frame, ordered by id and time
`q`	a scalar to denote the number of quadrature points used for Gauss-Hermite numerical integration
`Yname`	a character string of the name of new binary variable

4.6 Examples

4.6.1 Madras Longitudinal Schizophrenia Study

The Madras Longitudinal Schizophrenia Study was performed to characterize the clinical course of schizophrenia among those with first-episode psychoses in developing countries (Thara et al., 1994). Between October 1981 and October 1982, a total of 90 subjects met eligibility criteria. Demographics and clinical data were collected including monthly measurements of six psychiatric symptoms (hallucinations, delusions, thought disorders, flat affect, apathy, and withdrawal). A subset of these data has been analyzed previously in which thought disorder trajectories were compared between genders and age of onset groupings (Diggle et al., 2002; Schildcrout and Heagerty, 2007).

The subset of the Madras data used in Diggle et al. (2002) and Schildcrout and Heagerty (2007) are provided in the the `MMLB` package. These data contain information on 86 patients including: patient identifier (`id`), an indicator of the presence of a thought disorder (`thought`), month since hospitalization (`month`), age at onset ($1 =< 20$, $0 =\geq 20$; `age`), and gender (1=female, 0=male; `gender`). Slight differences exist in the coding of the Madras variables when comparing these authors'

analyses and care is needed if attempting to replicate either analysis. To demonstrate functions found in the `MMLB` package, data subsetting (i.e., requiring measurements at baseline and month 1) and variable coding (e.g., female reference group) are applied to replicate the analysis of Schildcrout and Heagerty (2007).

The marginal mean model of interest is defined as:

$$\text{logit}(\mu_{ij}^m) = \beta_0 + \beta_1 \text{month}_{ij} + \beta_2 \text{age}_i + \beta_3 \text{gender}_i + \beta_4 \text{age}_i \cdot \text{month}_{ij} + \beta_5 \text{gender}_i \cdot \text{month}_{ij}$$

We explore three different dependence models:

$$\begin{aligned} \text{logit}(\mu_{ij}^c) &= \Delta_{ij} + \gamma Y_{ij-1} + \sigma z_i && \text{(mTLV)} \\ &= \Delta_{ij} + \gamma Y_{ij-1} && \text{(mT)} \\ &= \Delta_{ij} + \sigma z_i && \text{(mLV)} \end{aligned}$$

First, the `MMLB` package, which is currently maintained on GitHub, is installed using `devtools` library. The Madras data set is then loaded, and prepared for analysis.

```
#library(devtools)
#install_github('mercaldo/MMLB',force=TRUE)
library(MMLB)

data(madras)

# Prep data to match Schildcrout and Heagerty, 2007
madras2        <- madras[,c('id','thought','month')]
madras2$gender <- factor((madras$gender==0)*1,labels=c('female','male'))    # 1= Male
madras2$age    <- factor((madras$age==0)*1,labels=c('>=20','<20'))          # 1= <20
madras2$nvisit <- unlist( lapply( split(madras2$month, madras2$id), function(ZZ) rep(length(ZZ), length(ZZ)) ) )

head(madras2,n=5)

  id thought month gender  age nvisit
   1       1     0   male >=20     12
   1       1     1   male >=20     12
   1       1     2   male >=20     12
   1       1     3   male >=20     12
   1       1     4   male >=20     12

# Restrict analysis to those with at least 2 visits (id=82 is dropped)
madras2 <- madras2[ which(madras2$nvisit>1), ]
```

Next, the three marginalized models are declared by changing the `lv.formula` and `t.formula` arguments. By default, both are initially assigned `NULL`, and thus non-applicable formulas can be ignored if fitting either a mT or mLV model. If neither association models are specified, then an error is returned. Summaries may

be explored using either the `print` or `summary` functions, but for brevity, we report only the mean and association summary tables for the mTLV model.

```
mTLV <- mm(thought~month*gender+month*age, lv.formula=~1, t.formula=~1, data=madras2,id=id)
mT   <- mm(thought~month*gender+month*age, t.formula=~1,  data=madras2,id=id)
mLV  <- mm(thought~month*gender+month*age, lv.formula=~1, data=madras2,id=id)

#print(mTLV)
#summary(mTLV)
lapply( summary(mTLV)[c("mean.table","assoc.table")], round, 4)

$mean.table
                Estimate Model SE Chi Square Pr(>Chi)
(Intercept)       0.2657   0.3336     0.6343   0.4258
month            -0.3560   0.0716    24.7333   0.0000
gendermale        0.3224   0.4045     0.6350   0.4255
age<20            0.7633   0.4378     3.0399   0.0812
month:gendermale  0.1069   0.0783     1.8642   0.1721
month:age<20     -0.1211   0.0809     2.2385   0.1346

$assoc.table
                      Estimate Model SE Chi Square Pr(>Chi)
gamma:(Intercept)       2.5076   0.3039    68.0822   0.0000
log(sigma):(Intercept)  0.0786   0.2447     0.1033   0.7479
```

The primary question of interest is whether subjects with an older age-at-onset tend to recover more or less quickly than younger subjects, and whether female patients recover more or less quickly than males. Based on these data, and the assumed mTLV model, younger individuals (< 20) tended to recover more quickly than older subjects (-0.12, 95% CI: -0.28,0.04), and males more slowly than females (0.11, -0.04,0.26). A significant non-zero amount of serial and non-diminishing dependence is observed, and acknowledged by this model ($\hat{\gamma} = 2.5$, $\hat{\sigma} = 1.08$).

4.6.2 Simulation Example

To generate longitudinal binary response data, that is consistent with a marginalized regression model, both the functional forms, and parameter values of the marginal mean and dependence models need to be specified. Suppose it is of interest to generate response data on 2500 subjects according to the following mTLV model:

$$\begin{aligned} \text{logit}(\mu_{ij}^m) &= -1.5 + 0.25\ \text{time}_{ij} + 0.25\ \text{Xe}_i + 0.1\ \text{Xe}_i \cdot \text{time}_{ij} \\ \text{logit}(\mu_{ij}^c) &= \Delta_{ij} + Y_{ij-1} + z_i \end{aligned} \tag{4.17}$$

where $\text{time}_{ij} = 0, 1, \ldots n_i$ for $n_i \in \{5, 10\}$ and $\text{Xe}_i \sim \text{Bernoulli}(0.35)$ is a binary exposure.

```r
# Generate data
set.seed(1)
N      = 2500
nclust = sample( seq(5,10), N, replace=TRUE)
id     = rep(seq(N), nclust)
Xe     = rep(rbinom(N,size=1,prob=.35), nclust)
time   = unlist( sapply( as.list(nclust), function(ZZ) seq(ZZ)-1 ) )
data   = data.frame(id, time, Xe)
data   = data[order(data$id, data$time),]

# Generate response data, called Y
newdata = GenBinaryY(mean.formula=~time*Xe, lv.formula=~1, t.formula=~1,
                     beta=c(-1.5, .25, .25, .1), sigma=1, gamma=1, id=id, data=data, Yname = "Y")

head(newdata, n=5)

 id time Xe Y
  1    0  0 0
  1    1  0 0
  1    2  0 0
  1    3  0 0
  1    4  0 0

# Fit model; verify parameters and estimates
mod_mtlv = mm(Y~time*Xe,lv.formula=~1, t.formula=~1, data=newdata,id=id)
lapply( summary(mod_mtlv)[c("mean.table","assoc.table")], round, 4)

$mean.table
            Estimate Mod.SE Chi Square Pr(>Chi)
(Intercept)  -1.4939 0.0459  1059.6929   0.0000
time          0.2589 0.0091   814.2626   0.0000
Xe            0.1790 0.0760     5.5530   0.0184
time:Xe       0.1169 0.0162    52.2462   0.0000

$assoc.table
                      Estimate Mod.SE Chi Square Pr(>Chi)
gamma:(Intercept)       1.0543 0.0507   432.4830   0.0000
log(sigma):(Intercept) -0.0783 0.0423     3.4277   0.0641
```

4.6.3 Outcome-Dependent Sampling Examples

A hypothetical two-stage ODS design is performed on simulated data where the data are generated according to the mTLV model from Equation 4.17, except all subjects had 10 observations. To resemble a small internal pilot study of 100 subjects, and a follow-up study of 300 subjects, the first and second stage ODS designs are defined as $D[25, 50, 25]$ and $D[0, 300, 0]$, respectively. These data are available in the `MMLB` package and accessible via `data(odsdat)`. This dataset contains information on the 383 sampled subjects, and includes the following key variables: `id`, `time`, `Xe`, `Y`,

`sample`, `sp1`,`sp2`, and `sp3`. Note, since independent Bernoulli sampling is performed, it is not unusual to observe a sample size not equal to 400. The `sample` variable corresponds to the stage of the study (1=stage 1, and 2=stage 2), and `sp1-sp3` denote the appropriately calculated sampling weights for each of the sampling strata.

To demonstrate how the `mm` function is used when analyzing data from an ODS design, we restrict the `odsdat` to only those measurements from stage one. To estimate the mTLV model parameters that maximize the ascertainment-corrected likelihood (Equation 4.6), `cond.like` is assigned `TRUE` and `samp.probs` is assigned the numeric vector of sampling probabilities. Alternatively, model parameters associated with the unconditional likelihood (Equation 4.4) may be estimated using weighted estimating equations. This is accomplished by specifying `samp.probi` a vector of subject-specific sampling probabilities (i.e., if $v_1 = 1$, then `samp.probi[1]`$=\pi(1)$).

```
data(odsdat)

head(odsdat, n=5)

sample id time Xe Y nobs sumY ss       sp1        sp2       sp3
     1  3    0  0 0   10    0  1 0.1196172 0.02238138 0.4385965
     1  3    1  0 0   10    0  1 0.1196172 0.02238138 0.4385965
     1  3    2  0 0   10    0  1 0.1196172 0.02238138 0.4385965
     1  3    3  0 0   10    0  1 0.1196172 0.02238138 0.4385965
     1  3    4  0 0   10    0  1 0.1196172 0.02238138 0.4385965

# Restrict data to stage 1 data only
stage_1_ods <- odsdat[odsdat$sample==1,]
samp_probs  <- as.numeric(stage_1_ods[1, c('sp1','sp2','sp3')])
# or samp_probs=as.matrix( stage_1_ods[,c('sp1','sp2','sp3')] )

mtlv_ods   <- mm(Y~time*Xe, lv.formula=~1,   t.formula=~1,   data=stage_1_ods,
                 id=id, cond.like=TRUE, samp.probs=samp_probs)
lapply( summary( mtlv_ods )[c('mean.table','assoc.table')], round, 4)

$mean.table
            Estimate Model SE Chi Square Pr(>Chi)
(Intercept)  -1.9720   0.2345    70.7397   0.0000
time          0.2812   0.0428    43.2042   0.0000
Xe            0.3626   0.3040     1.4221   0.2331
time:Xe       0.0927   0.0618     2.2443   0.1341

$assoc.table
                     Estimate Model SE Chi Square Pr(>Chi)
gamma:(Intercept)      0.9438   0.2913    10.4997   0.0012
log(sigma):(Intercept) 0.1669   0.1448     1.3274   0.2493

# Via weighted estimating equations
mtlv_wee   <- mm(Y~time*Xe, lv.formula=~1,   t.formula=~1,   data=stage_1_ods,
                 id=id, samp.probi=as.numeric(stage_1_ods[1, c('sp1','sp2','sp3')])[stage_1_ods$ss])
lapply( summary( mtlv_wee )[c('mean.table','assoc.table')], round, 4)
```

```
$mean.table
            Estimate Robust SE Chi Square Pr(>Chi)
(Intercept)  -2.3293    0.3541    43.2767   0.0000
time          0.3253    0.0593    30.0684   0.0000
Xe            0.6774    0.4515     2.2517   0.1335
time:Xe       0.0174    0.0714     0.0591   0.8079

$assoc.table
                       Estimate Robust SE Chi Square Pr(>Chi)
gamma:(Intercept)        0.9514    0.2698    12.4308   0.0004
log(sigma):(Intercept)   0.1000    0.1536     0.4233   0.5153
```

Robust standard errors are displayed when summarizing a model whose parameters were estimated using weighted estimating equations. When comparing the estimation approaches, the point estimates are comparable (roughly within one standard error), and the robust standard errors are larger than the model-based standard errors due to variability in the sampling weights.

When analyzing the combined data from a two-stage ODS design, one only needs to supply a numeric matrix of sampling probabilities to the `samp.probs` argument.

```
# The sampling probabilites for each stage are:
unique(as.matrix( odsdat[,c('sp1','sp2','sp3')] ))

           sp1        sp2       sp3
597  0.1196172 0.02238138 0.4385965
1958 0.0000000 0.13453467 0.0000000

# To analyze data from a two-stage ODS design, the same function call is made.
mtlv_ods_2 <- mm(Y~time*Xe, lv.formula=~1,   t.formula=~1, data=odsdat,
                 id=id, cond.like=TRUE, samp.probs=as.matrix( odsdat[,c('sp1','sp2','sp3')] ))
lapply( summary( mtlv_ods_2 )[c('mean.table','assoc.table')], round, 4)

$mean.table
            Estimate Model SE Chi Square Pr(>Chi)
(Intercept)  -1.9921   0.1190   280.0728   0.0000
time          0.2707   0.0179   227.5144   0.0000
Xe            0.2972   0.1789     2.7596   0.0967
time:Xe       0.0800   0.0294     7.3884   0.0066

$assoc.table
                       Estimate Model SE Chi Square Pr(>Chi)
gamma:(Intercept)        0.8408   0.1152    53.3127   0.0000
log(sigma):(Intercept)   0.1389   0.0795     3.0514   0.0807
```

4.7 Conclusions and future developments

We have illustrated how the `MMLB` package can be used to analyze longitudinal binary data using marginalized regression models. Using the `mm` function, it is straight-

forward to fit models with different mean, and dependence model specifications. Data collected under random or outcome-dependent sampling may be analyzed with minimal modifications to the `mm` function call. We also demonstrated how longitudinal binary data can be generated under a marginally specified mean model.

Currently, the `MMLB` package only permits the estimation of model parameters, and their standard errors. We plan to incorporate the empirical Bayes modal prediction estimates of latent variables. This will allow for the estimation of individual-level predictions, as well as more efficiently approximating integrals by using adaptive Gauss-Hermite quadrature. Additionally, to evaluate the feasibility of an ODS design, or to perform an adaptive ODS two-stage design, it is essential to estimate exposure values for non-sampled subjects (Schildcrout and Heagerty, 2011; **?**). The `mm` function currently returns attributes that are needed to estimate the conditional exposure odds for the non-sampled subjects (e.g., subject-specific likelihood contributions and ascertainment-corrections), but flexible, user-friendly functions are still under development.

CHAPTER 5

ENRICHMENT SAMPLING FOR A MULTI-SITE PATIENT SURVEY USING ELECTRONIC HEALTH RECORDS AND CENSUS DATA

5.1 Abstract

We describe an enrichment-motivated stratified sampling design that combines electronic health records (EHR) and United States Census data to construct the sampling frame. The design was motivated by a multicenter survey that sought to examine patient concerns about and barriers to participating in research studies, especially among under-studied populations (e.g., minorities and those with low educational attainment). We defined sampling strata by the cross-tabulation of several key socio-demographic variables (age, gender, race, ethnicity, rural living, and education). We used individual-level data from the EHR when available, and when missing, we imputed aggregated census data. This sampling strategy led to a far more diverse sample than would have been expected under random sampling (e.g., 3-,8-,7- and 12-fold increase in African-Americans, Asians, Hispanics and those with less than a high-school degree, respectively). We observed that EHR data tend to misclassify minority races more often than majority races, and that non-majority races, Latino ethnicity, younger adult age, lower education, and urban/suburban living are associated with lower response rates to the mailed surveys.

5.2 Introduction

The United States health care system has become more reliant on health information technology and active data collection due in part to the Health Information Technology for Economic and Clinical Health Act of 2009 (HITECH). This Act provides financial incentives to institutions that are implementing and promoting the "meaningful use" of electronic health record (EHR) data. As the amount of EHR data proliferates, nationwide efforts (e.g., Project HealthDesign) have been initiated to generate novel secondary uses of EHR data to improve public health (Safran et al., 2007; Casey et al., 2016). These data are used to reevaluate prior research findings, to develop, assess and refine predictive models, to aid in the planning of epidemiological and survey studies, and combined with biorepositories to understand complex genotype and phenotype relationships (Roden et al., 2008).

To date, research derived from biorepositories is primarily based on individuals of northern European ancestry. To engage more diverse populations in genomic research, surveying under-studied populations is needed to better understand concerns about and barriers to participating in research studies. Such surveys are typically extremely resource intensive unless a well-defined sampling frame exists (Sudman et al., 1988). Defining a sampling frame from EHR demographic data is possible since recipients of HITECH funds are required to collect standardized demographic data that may be associated with health disparities (Douglas et al., 2015). The quality of the resulting sampling frame is dependent on the accuracy and completeness of each institution's EHR system and may not be sufficient for certain research questions (e.g., coarseness of racial/ethnic groups) (Douglas et al., 2015; Coorevits et al., 2013; Shivade et al., 2014; Holland and Palaniappan, 2012). In this paper, we describe a stratified sampling design that we used for the Electronic Medical Records and Genomic (eMERGE) Network's survey of perspectives on broad consent and data sharing in biomedical research (Smith et al., 2016). An aim of this multi-site survey was to ensure that under-studied populations were adequately represented (e.g., minorities and those from rural areas). We defined the sampling frame using EHR data, and when necessary using United States Census (USC) data. In addition to describing the design, we report response rates among the various subgroups and the extent to which EHR data used to define sampling strata agreed with those reported by survey respondents.

5.3 Methods

5.3.1 Population and data sources

The eMERGE Network was initiated by the National Human Genome Research Institute to "develop, disseminate, and apply approaches to research that combine DNA biorepositories with EHR systems for large-scale, high-throughput genetic research" (Gottesman et al., 2013). The Consent, Education, Regulation and Consultation (CERC) Working Group was commissioned to conduct a broad-based survey on the acceptability of and barriers to broad consent and data sharing for genomics research, especially among those with low socioeconomic status, low education, rural residence, younger adults, and ethnic and racial minorities (Garrison et al., 2016). Among the eMERGE Network's 11 US clinical centers, this survey was administered to seven sites that sampled from their adult patient population, three sites that sampled from their pediatric patient population only, and one site that sampled from both its adult

and pediatric populations. Patients that had an inpatient or outpatient encounter between October 1, 2013 and September 30, 2014 and were not known to be deceased, whose address was geocodable (see Linking EHR and USC Data), and whose age and gender could be identified in the EHR were eligible for sampling. Overall, the sampling frame consisted of approximately 2.4 million individuals. The completeness of the sociodemographic variables used to define sampling strata within each site's EHR varied greatly. When EHR data were not available, USC based estimates were used. The following subsections describe the EHR and USC data sources and the process of merging the datasets to create the variables needed to define the sampling strata.

5.3.2 EHR data

Table 5.1 summarizes the EHR data, including percentage of missing data, summarized by population (adult, pediatric) and by site. Within adult sites, the median patient age was 52 years. Fifty-eight percent were female, 87% were white and 4% were Hispanic/Latino. At pediatric sites, the median age was 8 years, and a majority was male (52%), white (66%) and not Hispanic/Latino (93%). Race and ethnicity was missing from 14% and 16% of adult EHR records, respectively, and from 13% and 12% of pediatric EHR records. We observed substantial site-to-site variability in the availability of race and ethnicity data with values ranging from 67 to 99%.

Table 5.1: Marginal distributions of age, gender, race, and ethnicity by population and by site. Age is summarized as the 5, 25, 50 (median), 75 and 95th percentiles while gender, race and ethnicity are summarized with percentages (percentage missing). Age and gender were complete by design.

		Age	**Gender**	**Race**							**Ethnicity**	
	N	5 25 50 75 95	Female	Missing	White	Black	Asian	AI/AN	HI/PI	Other	Missing	Hispanic
Population												
Adult	1787295	22-36-52-65-82	58.3	14.1	87.1	5.7	2.4	0.7	0.3	3.8	19.3	4.1
Pediatric	601867	1-4-8-13-17	47.6	13.1	66.0	19.3	2.9	0.2	0.1	11.6	11.8	6.7
Site												
Essentia Institute for Rural Health	243092	21-35-53-66-84	56.7	1.0	94.8	1.1	0.4	2.1	0.1	1.5	1.3	0.9
Group Health Cooperative	217959	22-35-51-63-78	58.3	30.1	78.3	5.5	9.7	2.0	1.3	3.1	30.0	5.3
Geisinger Health System	356488	22-36-52-66-82	58.0	3.9	96.3	2.6	0.6	0.1	0.3	< 0.1	7.3	2.9
Mayo Clinic	136391	23-41-57-69-83	53.2	3.4	93.5	2.0	1.9	0.4	0.1	2.1	10.1	2.1
Marshfield Clinic	134212	21-35-52-66-83	54.6	8.4	97.2	0.5	1.4	0.8	0.1	< 0.1	9.1	1.8
Mount Sinai School Medicine	162927	23-39-54-67-83	59.9	30.8	60.8	11.8	4.7	0.2	0.1	22.4	70.2	26.1
Northwestern University	206554	23-35-47-60-77	62.6	21.1	70.9	13.2	3.6	0.2	0.1	12.0	22.8	9.9
Vanderbilt University Medical Center	329672	21-36-52-66-81	59.7	18.3	86.6	10.7	1.4	0.2	0.1	0.9	19.0	2.4
Boston Children's Hospital	140304	1-4-8-13-17	47.4	21.5	67.5	10.0	4.2	0.2	0.1	17.9	19.9	6.8
Cincinnati Children's Medical Center	143994	1-3-8-12-16	48.4	11.3	75.8	18.5	1.7	0.1	0.1	3.9	6.2	4.5
Children's Hospital of Philadelphia	209755	1-4-8-13-17	47.6	1.0	55.3	24.7	3.1	0.1	0.1	16.7	3.0	7.0
Vanderbilt Children's Hospital	107814	1-3-8-13-16	46.6	28.1	76.0	19.1	2.5	0.3	0.1	1.9	25.6	9.6

AI/AN = American Indian / Alaska Native, NH/PI = Native Hawaiian / Pacific Islander. The population file used for sampling at Mt. Sinai contained 70.2% missing ethnicity values (the overall adult population had 19.3% missing ethnicity). Sampling was performed using the erroneous data while the corrected results are reported in this table.

5.3.3 USC data

Populations with low educational attainment and with rural residences have been understudied in prior research, and data fields that capture these characteristics were not available in any of the EHR systems. For these two fields, and for missing values in EHR records, we exploited US Census Bureau data to provide proxy values. For instance, rural residence as determined by the 2010 Census urban areas criteria is likely to be fully accurate to the extent that patients' addresses in the EHR are accurate. The US Census Bureau administers several surveys each year, in addition to the Decennial Census. This includes the American Community Survey (ACS), an ongoing nationwide program that collects sociodemographic and economic information about the US population (US Census Bureau, n.d.*a*,n). Table 5.2 describes the USC sources, variable definitions and transformations used to complete race, ethnicity, education and rural living when needed.

Table 5.2: US Census variables, sources, definitions and transformations used for imputing missing stratification information.

Variable	US Census Variable	Source	Description	Variables Transformation
Race/Ethnicity	Hispanic Or Latino Origin By Race	B03002 (001-021)	Number overall and of each race (White alone, Black or African American alone, American Indian / Alaska Native alone, Asian alone, Native Hawaiian / Other Pacific Islander, Some other race alone, Two or more races, White alone not hispanic or latino, hispanic or latino Two races including some other race, two races excluding some other race / three or more races) by ethnicity (Hispanic, not hispanic).	Marginal distributions of race were defined as white (003,013), Black or African American (004,014), Asian (006,016), American Indian/Alaska Native (005,015), Native Hawaiian/Pacific Islander (007,017), Other (008,009,018,019). Marginal distributions of ethnicity were defined as: Not Hispanic/Latino (002), Hispanic/Latino (012)
Education	Sex By Educational Attainment For The Population 25 Years And Over	B15002 (001-035)	Number of each educational attainment group (no schooling, nursery to fourth grade, 5^{th} and 6^{th}, 7^{th}-8^{th}, 9^{th}, 10^{th}, 11^{th}, 12^{th} with no diploma, HS grad/GED/Alternative, some college less than 1 year, some college one or more years and no degree, associates degree, bachelor's degree, masters degree, professional school degree, doctorate) by gender for those who are 25 or older.	Marginal distributions of education were defined as: < 12 (003-010,020-027), $12- < 16$ (011-014,028-031), ≥ 16 (015-018,032-035).
Rurality	LSAD10	2010 Census urban area criteria	75=urbanized area (50000 or more), 76=urban cluster (2500 to 50000), missing=rural.	75 or 76 (Suburban/Urban), Missing (Rural)

Age and gender were complete by design presented. ACS 2008-2012 5-Year summary files were used to define census block group values of race, ethnicity and education while the 2010 Census was used to assign urban and rural classifications.

5.3.4 Linking EHR and USC data

Estimating EHR data from USC data requires linking home addresses to USC geographical identifiers. Address processing involved cleaning address fields, such as the primary street address, city, state and zip code, and applying quality control

checks. Processed addresses were then converted to latitude and longitude values, or geocoded, using specialized software, such as ArcGIS and R (ESRI, 2011; R Core Team, 2017)]. Each site performed its own address processing and quality assessment. Seven sites geocoded their addresses, and the coordinating center (CC, Vanderbilt University) geocoded the rest. Geocoded addresses were then linked to USC block group geographical identifiers, which are the most granular identifiers found in USC datasets, using specialized state-specific files and software. The CC managed and curated data obtained from the 2008-2012 ACS summary tables and the 2010 urban areas database and then distributed the data to all sites for merging with the site-specific EHR data.

5.3.5 Imputing missing EHR data with USC data

To identify the sampling frame, we "filled-in" (i.e., imputed) missing demographic variables using the most-frequent (mode) value from the patient's census block group. For example, if race was missing for a patient and the most common race in the patientâĂŹs census block group was African American, we imputed "African American." We conducted single imputation where necessary to define the sampling strata, while recognizing that the sampling frame is measured with some error.

5.3.6 Sampling scheme

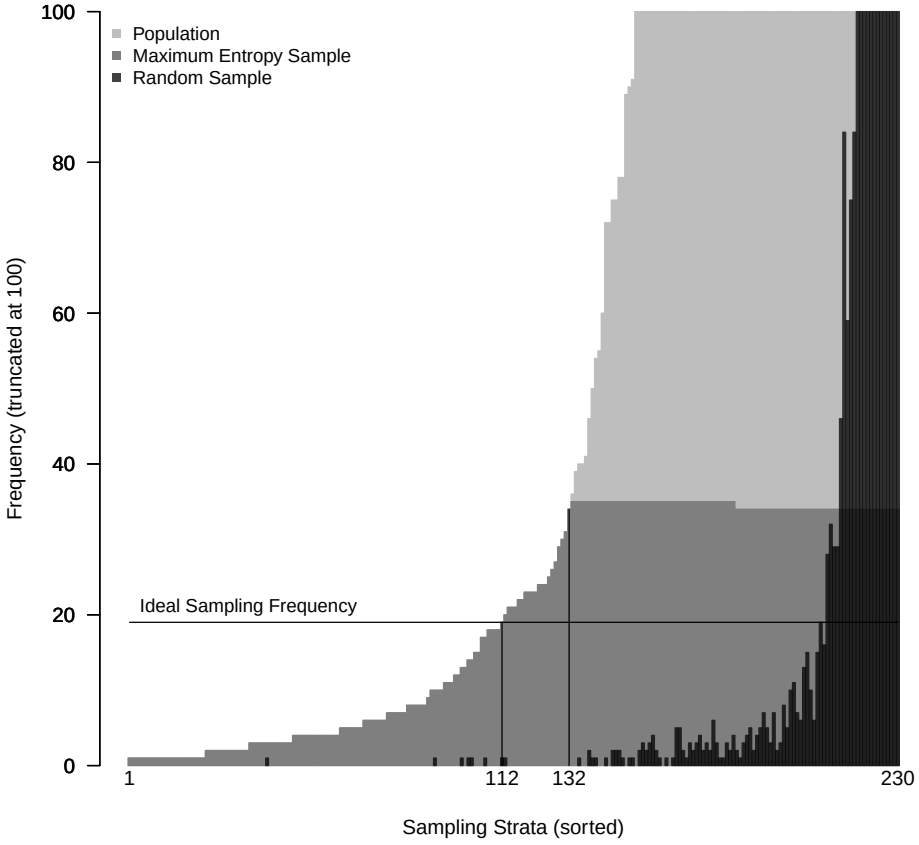
We conducted a disproportionate stratified sampling scheme to identify the sample. Using the combined EHR and USC data, we defined sampling strata at the adult sites based on the cross-classification age (< 35 and ≥ 35 years), gender, race (White, Black or African-American, Asian, Native American/Alaska Native, Hawaiian/Pacific Islander, Other), ethnicity (Hispanic or not), educational attainment (less than high school, high school degree or some college, and at least a bachelor's degree), and rural living (suburban/urban, rural). Patient data from pediatric sites were surrogates for their parents. That is, we sampled the parents from strata defined by the demographics of the child. We defined sampling strata similarly at pediatric sites, except the age variable was categorized as < 12 and ≥ 12 years. These categories were determined using results from an extensive literature review conducted by our team that showed the extent to which some subpopulations are under-represented in biorepository-derived research and based on the scientific questions of interest (Garrison et al., 2016). The cross-classification of the six sampling variables resulted in 288 possible strata although not all were observed at all sites.

5.3.7 Maximum entropy sampling algorithm

We conducted the sampling design to increase the diversity of those observed compared to the population at each site. Shannon's entropy, which corresponds to the uncertainty of predicting an individual's sampling stratum, was used to quantify diversity (Shannon, 2001). It is defined as $H = -p_i \log_2(p_i)$ where p_i denotes the probability of randomly selecting sampling stratum i. For s possible sampling strata, entropy values range from 0 to $\log_2(s)$ and correspond to the extreme scenarios where all individuals belong to the same stratum ($H = 0$) or where individuals are divided equally across all strata (i.e., assuming equal numbers of subjects were available from each stratum; $H = \log_2(s)$). To enrich our final sample with individuals from strata that tended to have small counts, we implemented a maximum entropy sampling (MES) algorithm. The MES algorithm iteratively determines the number of subjects to sample from each stratum so the desired sample size is obtained and the overall entropy is maximized. That is, MES seeks to sample as evenly as possible across strata under the constraints of overall desired sample size and the individual stratum sizes. Once the desired MES stratum counts were calculated, we implemented the sampling procedure with sampling probabilities defined as the ratio of the MES determined sample size for the stratum divided by the stratum size. Within each stratum sampling preference was given to those with complete (not imputed) stratification information.

Figure 5.1 describes the MES algorithm at Vanderbilt University Medical Center (VUMC) where 4,500 adults were sampled from a population of 329,672. Among the 288 total possible sampling strata, 230 were populated with at least one patient. The per stratum frequency in the population is denoted by the light gray shaded region (note the severe truncation at the top of the figure). When sampling 4,500 patients from 230 strata, Shannon entropy is maximized if 19 (4,500/230) were sampled from each stratum (see ideal sampling frequency line). However, only 118 strata contained more than 19 patients. To maximize entropy under stratum size constraints, all patients were sampled from the smallest 132 strata and 34 or 35 patients were sampled from the 98 strata with at least 35 patients. To contrast with MES, the black shaded region shows the numbers sampled from each stratum in one realization of a random sampling (RS) design. As expected, RS results in a far more skewed distribution (i.e., with lower Shannon entropy) and those from small strata are unlikely to be included in the sample. In this case, only 41% of the strata would be represented in the sample under this RS design.

Figure 5.1: Truncated histograms of the sorted sampling strata for the entire Vanderbilt-Adult population and for samples of size 4,500 using random and maximum entropy sampling (MES). The ideal sampling frequency is 19 per stratum with the remaining the 130 individuals being randomly selected from available strata. The MES sample is enriched compared to the random sample, especially with individuals from strata with sparse counts (strata 1-112); all individuals belonging to strata 1-132 were included in the final sample.



An R package was written to estimate MES counts for a given vector of stratum counts and an overall sample size. Code, installation instructions, and an example are publicly available at https://github.com/mercaldo/mes.

5.4 Results

5.4.1 Enrichment among those sampled

Table 5.3 summarizes the marginal distributions of the six stratification variables using the EHR data only and the combined EHR and USC data for the entire eMERGE network. Due to inclusion criteria, age and gender were available on all patients and so EHR and combined EHR and USC values are identical. At adult sites, the marginal distributions of race and ethnicity remained relatively unchanged after incorporating the USC data, likely due to only sampling 4.9% and 8.9% of participants with imputed race and ethnicity values, respectively. Most individuals lived in census block groups where the mode of the adult educational attainment distribution was between high school and some college (77%) followed by at least a bachelor's degree (22%). A total of 48% of the population resided in rural areas. Similar patterns were observed at pediatric sites, though fewer participants (29%) lived in rural areas.

Table 5.3: Marginal distributions of stratification variables when using only EHR data and when using both EHR and USC data. Percentages of non-missing values are reported by population (pediatric, adult) and for the sample of 90,000 households using maximum entropy sampling (MES).

	Adult			**Pediatric**		
	EHR Data Only	EHR/USC Data	MES	EHR Data Only	EHR/USC Data	MES
Age						
Low age group	22.9	22.9	43.8	68.7	68.7	56.7
Gender						
Female	58.3	58.3	52.9	47.6	47.6	49.7
Race						
White	87.1	87.6	34.3	66.0	69.3	33.0
Black	5.7	5.6	18.3	19.3	17.7	22.5
Asian	2.4	2.3	16.1	2.9	2.6	14.7
AI/AN	0.7	0.6	7.1	0.2	0.1	2.5
NH/PI	0.3	0.2	4.9	0.1	0.1	1.8
Other	3.8	3.6	19.2	11.6	10.2	25.5
Missing	14.1			13.1		
Ethnicity						
Hispanic/Latino	4.1	4.4	30.7	6.7	6.1	30.5
Missing	19.3			11.8		
Education						
<HS		1.0	11.9		1.2	13.6
HS+some college		76.8	54.9		72.4	48.9
≥Bachelor's		22.2	33.2		26.4	37.6
Rurality						
Rural		48.0	37.5		29.3	36.1

Low age group (< 12 in pediatric sites, < 35 in adult sites), AI/AN = American Indian / Alaska Native, NH/PI = Native Hawaiian / Pacific Islander.

As can be seen from the MES columns in Table 5.3, the sample identified by MES was enriched with target subpopulations as compared to the original population (EHR+USC Data). For example, the sample was enriched with all minority races; it

was enriched three-fold for African-Americans (18% vs 6%), 8-fold for Asians (16% vs 2%) and more than four-fold (19% vs 4%) for other races. The sample was also enriched more than 7-fold among those of Hispanic ethnicity (31% vs 4%), and 12-fold among those without a high school or equivalent degree (12% vs 1%). However, the survey was administered only in English and written at an 8th-grade literacy level thus possibly reducing the enrichment of the returned sample.

To further characterize enrichment due to MES sampling, entropy values by population are shown in Table 5.4. At adult sites, 262 of the 288 possible strata were observed corresponding to a maximum possible entropy of 8.03. The entropy values under random and maximum entropy sampling were 4.39 and 7.35, respectively. Overall, 81% and 72% of the maximum entropy was obtained by conducting the MES strategy compared to random sampling in the adult and pediatric sites, respectively. Site-specific results are provided in online appendices.

Table 5.4: Sampling frequencies and entropy estimates by sampling method and by population. Observed strata frequencies (nstrata) are provided along with maximum entropy (H_{max}), entropy under RS (H_{rs}) and MES (H_{mes}) and the percentage of maximum entropy accounted for by the MES sample above and beyond that of random sampling.

	MES Sample	n_{strata}	H_{max}	H_{rs}	H_{mes}	$\frac{H_{mes}-H_{rs}}{H_{max}-H_{rs}}$
Population						
Adult	58500	262	8.03	4.39	7.35	0.81
Pediatric	31500	251	7.97	5.16	7.18	0.72

5.4.2 Survey response rate and EHR accuracy

The CERC survey sampled 90,000 individuals and 7,761 were excluded due to invalid addresses (n=7,504), death/incapacity (n=168), or previous involvement in the pilot (n=89). A total of 13,000 surveys were returned resulting in an overall response rate of 16.7% at adult sites and 13.9% at pediatric sites (Table 5.5, see online appendix for pediatric and site-specific summaries). Response rates were also calculated for each stratification variable. Among adult sites, participants were less likely to respond if they were young (10.3% if $<$ 35 years and 21.6% if $\geq$ 35 years), male (16.0% if male and 17.4% if female), non-white (e.g., 13.2% if African American and 20.1% if white), Hispanic or Latino (14.2% if Hispanic and 17.9% if not), reside in low-education census blocks groups (13.6% if education model was <HS and 18.9% if education model was $\geq$ Bachelor's degree) or residing in non-rural areas (15.5% if

urban or suburban and 17.9% if rural).

Table 5.5: Overall and marginal response frequencies, rates and accuracy measures between combined EHR and USC sampling values and self-reported survey response values at adult sites.

	N	Response Rate	p	Se	Sp	PPV	NPV
Age Group							
< 35	8901	10.3	24.8	98.6	96.9	91.2	99.5
≥ 35		21.6	75.2	96.9	98.6	99.5	91.2
Gender							
Female	9011	17.4	56.0	97.3	98.3	98.7	96.7
Male		16.0	44.0	98.3	97.3	96.7	98.7
Race							
White	8941	20.1	48.7	77.3	91.1	89.1	80.8
Black or African American		13.2	11.2	93.2	96.3	76.0	99.1
Asian		17.1	16.2	84.9	96.4	82.1	97.1
American Indian or Alaska Native		17.4	2.7	81.8	94.7	29.9	99.5
Native Hawaiian or Pacific Islander		13.3	1.2	70.9	96.9	20.9	99.7
Other		14.2	20.0	33.3	88.4	41.8	84.1
Ethnicity							
Not Hispanic or Latino	8870	17.9	81.1	88.7	88.5	97.1	64.6
Hispanic or Latino		14.2	18.9	88.5	88.7	64.6	97.1
Education Group							
<HS	8769	13.6	7.6	20.7	91.4	16.7	93.3
HS+Some College		16.1	38.9	66.0	56.7	49.3	72.3
≥Bachelors		18.9	53.4	52.4	77.8	73.0	58.8
Rurality							
Rural	9185	18.7	42.7				
Suburban/Urban		15.5	57.3				

N= frequency p = prevalence, Se = sensitivity, Sp = specificity, PPV = positive predicted values, NPV = negative predictive value. Response rates equal the number of responses divided by the number in sample that satisfied integrity checks (previous inclusion in the pilot study, bad addresses, death, opt-out requests and blank responses were excluded).

Sensitivity (Se), specificity (Sp), positive and negative predictive values (PPV, NPV) were used to quantify the accuracy of EHR + USC data using survey response values as the gold standard. These are summarized in Table 5.5 for the adult sites only because at pediatric sites, the EHR data reflected characteristics of the child while survey responses reflected those of the parent or guardian. We therefore would not expect high accuracy. Accuracy estimates were not calculated for rural living since the true value is based on the address and not on a participant response. EHR age

and gender were at least 97% sensitive and specific for the 'true' value based on the survey response, and PPV and NPV were also reasonably high even though overall PPV for age < 35 years was only 91%. EHR and USC data showed variable sensitivity for race, ranging from 33% for other race to 93% for African American race, and the PPV for the smaller minority races was alarmingly low ($\sim$20-40%). Even though EHR and USC data were reasonably sensitive for Hispanic ethnicity (89%), the PPV was only 65%. Finally, utilizing only USC data to determine an individual's educational attainment resulted in low discriminative and predictive values (e.g., <HS: Se=21%, PPV=17%).

Overall, we observed that using EHR and possibly USC data to identify demographic subgroups may be a reasonable approach for common subgroups (African-American race, gender, non-Hispanic ethnicity); however, the smaller subgroups with very low prevalences (American Indian / Alaska Native race, Hispanic ethnicity) are often misclassified, and caution should be taken when using EHR data for their identification.

5.5 Discussion

This paper outlines a complex survey design that utilized both EHR and USC data for sample frame construction and introduced an algorithm that sought to enrich the final sample with individuals from rare subpopulations. In our sample, we observed substantial enrichment from subpopulations that would not have been observed had a standard random sampling scheme been used. There were several challenges with implementing such a design in this setting that include: incomplete and inaccurate EHR data, misclassification due to imputing missing EHR data with USC data, the targeting sampling to sparse sampling strata and ultimately, induced complexities associated design-based analyses.

The drawbacks of using EHR data for secondary research have been well documented (Weiskopf and Weng, 2013; Menachemi and Collum, 2011). Since these data are not primarily collected for research purposes their content and quality may vary by institution. The lack of universally accepted EHR criteria, except for the minimal criteria set by HITECH, may result in these data being insufficient to address certain research questions. In the primary results paper for our study, EHR data were used to define the sampling frame but were not used for primary study analyses (Sanderson et al., 2017). Further research is needed to quantify effects of measurement error

(or misclassification) on stratification variables, especially since EHR data are seldom complete and are often mismeasured (e.g., EHR race, Table 5.5) (Klinger et al., 2015; Fiscella and Fremont, 2006; Grundmeier et al., 2015).

The overall response rate in this study may have been influenced by the sampling of subgroups that are less inclined to participate in biomedical research. If sampling strata frequencies are related to willingness to respond, then this enrichment approach may result in a lower than expected response rate (e.g., $\sim$17% in the eMERGE survey). An alternative study design would decrease the number of subjects sampled while increasing resources towards ensuring that those who were sampled, answered the survey. However, at the onset of the study, we are determined that such a design was determined to be impractical across the 11 participating institutions.

In summary, we have outlined an approach that increases the diversity of a sample by oversampling those subjects that belong to rarer sampling strata. The magnitude of sample enrichment depends on the accuracy of the data used to define the sampling frame as well as the overall response rate. Thus, additional resources may be required to ensure that the frame is correctly enumerated and that sampled subjects complete the questionnaire. This approach may be especially well suited for health disparities research or other endeavors where it is of interest to elicit information from vulnerable or understudied populations.

CHAPTER 6

CONCLUSION

This dissertation has examined several topics related to the design and analysis of complex longitudinal and survey sampling studies.

In Chapter two, we extend outcome-dependent sampling (ODS) designs for longitudinal binary data to permit data collection in two stages. We consider two subclasses of designs: fixed designs where the designs at each stage are pre-specified, and adaptive designs that utilize stage one data to improve design choice at stage two. We demonstrate that data from both stages can be aggregated to generate valid parameter estimates using ascertainment-corrected maximum likelihood methods. Efficiency gains are observed compared to random sampling, and in certain situations, single-stage ODS sampling designs.

In Chapter three, we investigate the effects of utilizing an imperfect sampling frame on the design, and analysis of complex survey data. We explore the impact of stratum misclassification on the choice of study design, on the operating characteristics of survey estimators, and on the appropriateness of two common approaches to survey design analysis. Stratified sampling is recommended over random sampling if interest lies in making inferential statements regarding rare subgroups. In the presence of misclassification, the relative efficiency depends on the subgroup prevalence, and analytic methods that account for the design are still required for valid inferences.

In Chapter four, we introduce the MMLB R package which is used to estimate parameters from marginalized regression models for longitudinal binary data. These models are described, and estimation procedures outlined under random, and ODS schemes. We provide examples to demonstrate how to fit these models, and how data may be generated under a pre-specified marginal mean model.

We hope these chapters provide specific and general insights that will improve our ability to conduct efficient research studies under resource constraints.

REFERENCES

Anderson, J. A. (1972), Separate sample logistic discrimination, *Biometrika* **59**(1), 19–35.

Anthonisen, N. R. (2004), Lessons from the Lung Health Study., *Proceedings of the American Thoracic Society* **1**(2), 143–145.

Azzalini, A. (1994), Logistic regression for autocorrelated data with application to repeated measures, *Biometrika* **81**(4), 767–775.

Binder, D. A. (1983), On the variances of asymptotically normal estimators from complex surveys, *Statistical Review/Revue Internationale de Statistique* 279–292.

Blocker, A. W. (2014), *fastGHQuad: Fast Rcpp implementation of Gauss-Hermite quadrature*.

Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American statistical Association* **88**(421), 9–25.

Brunner, W. M., Schreiner, P. J., Sood, A. and Jacobs Jr., D. R. (2014), Depression and Risk of Incident Asthma in Adults. The CARDIA Study, *American Journal of Respiratory and Critical Care Medicine* **189**(9), 1044–1051.

Cai, J., Qaqish, B. and Zhou, H. (2001), Marginal analysis for cluster-based case-control studies, *Sankhyā: The Indian Journal of Statistics, Series B* 326–337.

Casey, J. A., Schwartz, B. S., Stewart, W. F. and Adler, N. E. (2016), Using Electronic Health Records for Population Health Research: A Review of Methods and Applications, *Annual Review of Public Health* **37**(1), 61–81.

Cochran, W. G. (1977), *Sampling techniques*, John Wiley & Sons.

Connett, J. E., Kusek, J. W., Bailey, W. C., O'Hara, P. and Wu, M. (1993), Design of the Lung Health Study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease., *Controlled clinical trials* **14**(2 Suppl), 3S–19S.

Coorevits, P., Sundgren, M., Klein, G. O., Bahr, A., Claerhout, B., Daniel, C., Dugas, M., Dupont, D., Schmidt, A., Singleton, P., De Moor, G. and Kalra, D. (2013), Electronic health records: new opportunities for clinical research, *Journal of Internal Medicine* **274**(6), 547–560.

Courbois, J.-Y. P. and Urquhart, N. S. (2004), Comparison of survey estimates of the finite population variance, *Journal of Agricultural, Biological, and Environmental Statistics* **9**(2), 236–251.

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002), *Analysis of longitudinal data*, Oxford University Press.

Douglas, M. D., Dawes, D. E., Holden, K. B. and Mack, D. (2015), Missed policy opportunities to advance health equity by recording demographic data in electronic health records., *American journal of public health* **105 Suppl 3**(S3), S380–8.

ESRI (2011), ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Fiscella, K. and Fremont, A. M. (2006), Use of geocoding and surname analysis to estimate race and ethnicity., *Health Services Research* **41**(4 Pt 1), 1482–1500.

Flegal, K. M., Keyl, P. M. and Nieto, F. J. (1991), Differential misclassification arising from nondifferential errors in exposure measurement., *American journal of epidemiology* **134**(10), 1233–1244.

Garrison, N. A., Sathe, N. A., Antommaria, A. H. M., Holm, I. A., Sanderson, S. C., Smith, M. E., McPheeters, M. L. and Clayton, E. W. (2016), A systematic literature review of individuals' perspectives on broad consent and data sharing in the United States., *Genetics in medicine : official journal of the American College of Medical Genetics* **18**(7), 663–671.

Gelman, A. (2007), Struggles with Survey Weighting and Regression Modeling, *Statistical Science* **22**(2), 153–164.

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., Brilliant, M., Carey, D. J., Chisholm, R. L., Chute, C. G., Connolly, J. J., Crosslin, D., Denny, J. C., Gallego, C. J., Haines, J. L., Hakonarson, H., Harley, J., Jarvik, G. P., Kohane, I., Kullo, I. J., Larson, E. B., McCarty, C., Ritchie, M. D., Roden, D. M., Smith, M. E., Böttinger, E. P. and Williams, M. S. (2013), The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future, *Genetics in medicine : official journal of the American College of Medical Genetics* **15**(10), 761–771.

Graubard, B. I. and Korn, E. L. (1996), Survey inference for subpopulations., *American journal of epidemiology* **144**(1), 102–106.

Greenland, S. (1988), Variance estimation for epidemiologic effect estimates under misclassification., *Statistics in medicine* **7**(7), 745–757.

Grundmeier, R. W., Song, L., Ramos, M. J., Fiks, A. G., Elliott, M. N., Fremont, A., Pace, W., Wasserman, R. C. and Localio, R. (2015), Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data, *Health Services Research* **50**(4), 946–960.

Haneuse, S., Schildcrout, J. and Gillen, D. (2012), A two-stage strategy to accommodate general patterns of confounding in the design of observational studies, *Biostatistics* **13**(2), 274–288.

Heagerty, P. J. (1999), Marginally specified logistic-normal models for longitudinal binary data., *Biometrics* **55**(3), 688–698.

Heagerty, P. J. (2002), Marginalized transition models and likelihood inference for longitudinal categorical data, *Biometrics* **58**(2), 342–351.

Holland, A. T. and Palaniappan, L. P. (2012), Problems With the Collection and Interpretation of Asian-American Health Data: Omission, Aggregation, and Extrapolation, *Annals of Epidemiology* **22**(6), 397–405.

Kalton, G. (2009), Methods for oversampling rare subpopulations in social surveys, *Survey methodology* **35**(2), 125–141.

Kim, J. K. and Skinner, C. J. (2013), Weighting in survey analysis under informative sampling, *Biometrika* **100**(2), 385–398.

Klinger, E. V., Carlini, S. V., Gonzalez, I., Hubert, S. S., Linder, J. A., Rigotti, N. A., Kontos, E. Z., Park, E. R., Marinacci, L. X. and Haas, J. S. (2015), Accuracy of race, ethnicity, and language preference in an electronic health record., *Journal of general internal medicine* **30**(6), 719–723.

Kuha, J. and Skinner, C. (1997), *Categorical Data Analysis and Misclassification*, Vol. 32, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Laird, N. M. (1988), Missing data in longitudinal studies., *Statistics in medicine* **7**(1-2), 305–315.

Lee, Y. and Nelder, J. A. (2004), Conditional and Marginal Models: Another View, *Statistical Science* **19**(2), 219–238.

Lin, X., Genest, C., Banks, D. L., Molenberghs, G. and Scott, D. W. (2014), *Past, present, and future of statistical science*, CRC Press.

Lindsey, J. K. and Lambert, P. (1998), On the appropriateness of marginal models for repeated measurements in clinical trials., *Statistics in medicine* **17**(4), 447–469.

Lohr, S. L. (2009), *Sampling: design and analysis. 2nd*, Cengage Learning.

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

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J. and Sherry, S. T. (2007), The NCBI dbGaP database of genotypes and phenotypes., *Nature genetics* **39**(10), 1181–1186.

Menachemi, N. and Collum, T. H. (2011), Benefits and drawbacks of electronic health record systems., *Risk management and healthcare policy* **4**, 47–55.

Molenberghs, G. and Verbeke, G. (2005), *Models for discrete longitudinal data*, Springer.

Neuhaus, J. M. and Jewell, N. P. (1990), The effect of retrospective sampling on binary regression models for clustered data, *Biometrics* **46**(4), 977.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991), A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data, *International Statistical Review/Revue Internationale de Statistique* **59**(1), 25–35.

Pandharipande, P., Cotton, B. A., Shintani, A., Thompson, J., Pun, B. T., Morris Jr, J. A., Dittus, R. and Ely, E. W. (2008), Prevalence and Risk Factors for Development of Delirium in Surgical and Trauma Intensive Care Unit Patients, *The Journal of Trauma: Injury, Infection, and Critical Care* **65**(1), 34–41.

Prentice, R. L. and Pyke, R. (1979), Logistic disease incidence models and case-control studies, *Biometrika* **66**(3), 403–411.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995), Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the american statistical association* **90**(429), 106–121.

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R. and Masys, D. R. (2008), Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine, *Clinical Pharmacology & Therapeutics* **84**(3), 362–369.

Rothman, K. J., Greenland, S. and Lash, T. L. (2008), *Modern epidemiology. 3rd Edition*, Lippincott Williams & Wilkins.

Rubin, D. B. (1976), Inference and missing data, *Biometrika* **63**(3), 581–592.

Rust, K. F. and Rao, J. N. (1996), Variance estimation for complex surveys using replication techniques., *Statistical methods in medical research* **5**(3), 283–310.

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., Detmer, D. E. and Panel, 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.

Sanderson, S. C., Brothers, K. B., Mercaldo, N. D., Clayton, E. W., Antommaria, A. H. M., Aufox, S. A., Brilliant, M. H., Campos, D., Carrell, D. S., Connolly, J., Conway, P., Fullerton, S. M., Garrison, N. A., Horowitz, C. R., Jarvik, G. P., Kaufman, D., Kitchner, T. E., Li, R., Ludman, E. J., McCarty, C. A., McCormick, J. B., McManus, V. D., Myers, M. F., Scrol, A., Williams, J. L., Shrubsole, M. J., Schildcrout, J. S., Smith, M. E. and Holm, I. A. (2017), Public Attitudes toward Consent and Data Sharing in Biobank Research: A Large Multi-site Experimental Survey in the US, *The American Journal of Human Genetics* 1–14.

Schildcrout, J. S., Garbett, S. P. and Heagerty, P. J. (2013), Outcome Vector Dependent Sampling with Longitudinal Continuous Response Data: Stratified Sampling Based on Summary Statistics, *Biometrics* **69**(2), 405–416.

Schildcrout, J. S. and Heagerty, P. J. (2007), Marginalized models for moderate to long series of longitudinal binary response data., *Biometrics* **63**(2), 322–331.

Schildcrout, J. S. and Heagerty, P. J. (2008), On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates, *Biostatistics* **9**(4), 735–749.

Schildcrout, J. S. and Heagerty, P. J. (2011), Outcome-Dependent Sampling from Existing Cohorts with Longitudinal Binary Response Data: Study Planning and Analysis, *Biometrics* **67**(4), 1583–1593.

Schildcrout, J. S., Rathouz, P. J., Zelnick, L. R., Garbett, S. P. and Heagerty, P. J. (2015), Biased Sampling Designs to Improve Research Efficiency: Factors Influencing Pulmonary Function Over Time in Children with Asthma, *The annals of applied statistics* **9**(2), 731–753.

Shannon, C. E. (2001), A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review* **5**(1), 3–55.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B. and Lai, A. M. (2014), A review of approaches to identifying patient phenotype cohorts using electronic health records, *Journal of the American Medical Informatics Association* **21**(2), 221–230.

Smith, M. E., Sanderson, S. C., Brothers, K. B., Myers, M. F., McCormick, J., Aufox, S., Shrubsole, M. J., Garrison, N. A., Mercaldo, N. D., Schildcrout, J. S., Clayton, E. W., Antommaria, A. H. M., Basford, M., Brilliant, M., Connolly, J. J., Fullerton, S. M., Horowitz, C. R., Jarvik, G. P., Kaufman, D., Kitchner, T., Li, R., Ludman, E. J., McCarty, C., McManus, V., Stallings, S., Williams, J. L. and Holm, I. A. (2016), Conducting a large, multi-site survey about patients' views on broad consent: challenges and solutions, *BMC Medical Research Methodology* **16**(1), 1–11.

Song, R., Zhou, H. and Kosorok, M. R. (2009), A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome, *Biometrika* **96**(1), 221–228.

Stiratelli, R., Laird, N. and Ware, J. H. (1984), Random-effects models for serial observations with binary response., *Biometrics* **40**(4), 961–971.

Sudman, S., Sirken, M. G. and Cowan, C. D. (1988), Sampling rare and elusive populations., *Science* **240**(4855), 991–996.

Sugden, R. A. and Smith, T. (1984), Ignorable and informative designs in survey sampling inference, *Biometrika* **71**(3), 495–506.

Thara, R., Henrietta, M., Joseph, A., Rajkumar, S. and Eaton, W. W. (1994), Ten-year course of schizophrenia–the Madras longitudinal study., *Acta psychiatrica Scandinavica* **90**(5), 329–336.

US Census Bureau (n.d.*a*), 2008-2012 American Community Survey 5-year estimates. http://www.census.gov/programs-surveys/acs/data/summary-file.html. Accessed June 2014.

US Census Bureau (n.d.*b*), 2010 Urban and Rural Classification and Urban Area Criteria. https://www.census.gov/geo/reference/ua/urban-rural-2010.html. Accessed June 2014. .

Weiskopf, N. G. and Weng, C. (2013), Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research., *Journal of the American Medical Informatics Association : JAMIA* **20**(1), 144–151.

Xu, W. and Zhou, H. (2012), Mixed effect regression analysis for a cluster-based two-stage outcome-auxiliary-dependent sampling design with a continuous outcome, *Biostatistics* **13**(4), 650–664.

Yuan, R., Hogg, J. C., Pare, P. D., Sin, D. D., Wong, J. C., Nakano, Y., McWilliams, A. M., Lam, S. and Coxson, H. O. (2009), Prediction of the rate of decline in FEV1 in smokers using quantitative computed tomography, *Thorax* **64**(11), 944–949.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988), Models for longitudinal data: a generalized estimating equation approach, *Biometrics* **44**(4), 1049.

Zhao, L. P. and Lipsitz, S. (1992), Designs and analysis of two-stage studies, *Statistics in medicine* **11**(6), 769–782.

Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T. and Longnecker, M. P. (2007), Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome., *Epidemiology (Cambridge, Mass.)* **18**(4), 461–468.

Zhou, H., Song, R., Wu, Y. and Qin, J. (2010), Statistical Inference for a Two-Stage Outcome-Dependent Sampling Design with a Continuous Outcome, *Biometrics* **67**(1), 194–202.

Zhou, H., Xu, W., Zeng, D. and Cai, J. (2013), Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling

scheme, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 197–215.