Title

Statistical Challenges in Incidence Estimation using Cross-Sectional Data and Multi-Biomarker Assay Algorithms

Author

Morrison, Douglas Ezra

Publication Date

2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Statistical Challenges in Incidence Estimation using Cross-Sectional Data and Multi-Biomarker Assay Algorithms

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Douglas Ezra Morrison

2021

ABSTRACT OF THE DISSERTATION

Statistical Challenges in Incidence Estimation using Cross-Sectional Data and Multi-Biomarker Assay Algorithms

by

Douglas Ezra Morrison

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2021

Professor Ron Brookmeyer, Chair

Accurate estimates of incidence rates of infectious diseases are important for monitoring trends and for designing and evaluating disease prevention and control programs. Traditionally, incidence has been estimated using cohort studies, which are costly, slow, and vulnerable to selection biases in both recruitment and attrition. Cross-sectional incidence estimation is an alternative approach that can avoid some of these problems. This approach involves collecting blood samples from a single representative cross-sectional survey of a target population, and analyzing the samples using multi-biomarker assay algorithms (MAAs) to detect recent infections. Under some assumptions about the dynamics of the epidemic, incidence is estimated from the prevalence of MAA-positive individuals, where MAA positive refers to a state defined by levels of biomarkers that is associated with recent acquisition of infection. A training data set is required to define and evaluate characteristics

of the MAA positive state. In order to achieve accurate estimates, cross-sectional incidence estimation analyses should be tailored to the population of scientific interest and to the data-generating process. This dissertation develops approaches for three challenges encountered in cross-sectional incidence estimation: analyzing incomplete or missing biomarker data; calibrating the cross-sectional estimation procedure for a specific target population; and accounting for interval-censored infection dates in longitudinal biomarker data.

The training data sets are used to operationally define the MAA positive state and to estimate the probabilities of being in the MAA positive state as a function of duration of time since acquisition of infection. The training data sets include longitudinal biomarker measurements on a sample of individuals. We first consider the challenge of missing biomarker data in the training data sets. We examine two naïve approaches, one using all samples that can be classified by the MAA and another using all samples with complete biomarker data, and we show that each of these approaches can lead to biased estimators of the mean window period. We propose a conditional approach for handling the missing data. We show that this method performs well in simulation studies. We then consider missing data in the context of cross-sectional surveys of biomarker prevalence. Again, we show that naïve approaches produce biased estimates, and we propose a conditional approach that performs well in simulation studies. We apply these methods to a training data set of biomarkers in HIV Subtype C infections collected from over two thousand individuals from multiple countries.

The target population refers to the population in which we wish to estimate incidence of infection. In order for a training data set to be useful for model calibration, any systematic

differences between the training data set and the target population must be addressed. We consider a scenario in which there is one covariate whose distribution differs between the training data set and the target population, and we propose a range of methods for correcting such a difference. Using simulation studies, we examine the performance characteristics of these methods under a range of analysis conditions and determine their sensitivity to model misspecifications.

Since infection status is usually only tested periodically in longitudinal studies, infection dates and durations of infection are typically interval-censored in MAA calibration data sets. We present a joint model of infection dates and subsequent biomarker values and an estimation procedure for this model, and we compare this approach with naïve methods assuming a uniform or symmetric distribution over the censoring intervals for the infection dates. We show that the joint modelling approach performs well in many situations compared to midpoint imputation and uniform imputation.

The methods presented in this dissertation were developed for the purpose of calibrating and performing HIV incidence estimation using cross-sectional surveys of biomarker prevalence. However, the cross-sectional survey-based approach to incidence estimation has applicability to infectious diseases other than HIV. This approach may be especially useful when it is crucial to rapidly detect changes in infection incidence to inform public health policies including epidemic control programs. We hope that the methods presented in this dissertation will encourage the use of the cross-sectional approach to incidence estimation in a variety of contexts and will help address the inevitable real-world complications in the data collection process.

The dissertation of Douglas Ezra Morrison is approved.

Onyebuchi A. Arah

Sudipto Banerjee

Tom Belin

Ron Brookmeyer, Committee Chair

University of California, Los Angeles

2021

To my grandparents, whose memory continues to encourage me every day.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply grateful for the support and guidance of my advisor, Dean Ron Brookmeyer. Under your mentorship, I have accomplished more than I thought I was capable of when I started this program. Thank you for steadily and patiently guiding me through this work; it has been a pleasure and an honor to learn from you.

I also want to thank Drs. Onyi Arah, Sudipto Banerjee, and Tom Belin for serving as members of my committee and for their invaluable contributions to my education and the content of this dissertation. In particular, my understanding of causal inference and missing data analysis has been greatly enriched by the courses taught by Drs. Belin and Arah, and I especially owe my confidence in wielding Bayes' Theorem and linear algebra to Dr. Banerjee. Thanks also to Dr. Vivek Shetty for including me in his fascinating work in mHealth and for providing a fantastic home base on campus (and a steady supply of espresso!).

I am grateful to my fellow students and research collaborators for their camaraderie throughout my coursework and research. Thanks also to the many UCLA faculty members who have taught, advised, and encouraged me throughout my time here, and to the Biostatistics and FSPH staff, especially Roxy Naranjo who expertly guided me through funding and award applications and patiently helped me navigate all of the logistical issues I encountered or created during the doctoral program.

I thank my friends, family, and mentors for their steady support and encouragement. In particular, thanks to my parents, Susan and Elliot, and my siblings, Nathaniel and Julia, for their love and inspiration. Thanks to Vangelis Hytopoulos for many years of friendship and mentorship. Thanks to Trevor Shaddox for many years of friendship, for countless adventures and misadventures, and for being the ideal college roommate. Lastly, thanks to Maura O'Leary for her love, understanding, and constant companionship. I'm so grateful that we found each other, and I'm looking forward to continuing our adventures together!

A version of the material presented in Chapter 3 has been published in *Statistical Communications in Infectious Diseases* with contributions from co-authors, Drs. Oliver Laeyendecker, Jacob Konikoff, and Ron Brookmeyer. (D. Morrison et al. 2018) A version of the material presented in Chapter 4 has been published in *Statistics in Medicine* with

contributions from co-authors, Drs. Oliver Laeyendecker and Ron Brookmeyer. (D. Morrison et al. 2019) A version of the material presented in Chapter 5 has been published in *Biometrics* with contributions from co-authors, Drs. Oliver Laeyendecker and Ron Brookmeyer. (D. Morrison et al. 2021) This work was supported in part by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH), R01-AI095068, by the UCLA Graduate Research Mentorship Program, and by the Celia and Joseph Blann Fellowship.

VITA

2009	B.S. (Symbolic Systems), Stanford University, Stanford, CA
2010	M.S. (Statistics), Stanford University, Stanford, CA
2010-2011	Biostatistician/Bioinformatician, Aviir, Inc., Palo Alto, CA
2012-2015	Statistician, Surgery Department, Stanford University School of Medicine, Stanford, CA

HONORS

2020	Celia and Joseph Blann Fellowship for academic excellence and commitment to public health, UCLA Fielding School of Public Health
2021	Carolbeth Korn Prize for most outstanding graduating student, UCLA Fielding School of Public Health

SELECTED PUBLICATIONS

A. Peer-reviewed journals

Morrison D, Laeyendecker O, Brookmeyer R. Regression with Interval-Censored Covariates: Application to Cross-Sectional Incidence Estimation. Biometrics 2021. doi:10.1111/biom.13472

Elmore JG, Wang PC, Kerr KF, Schriger DL, Morrison DE, Brookmeyer R, Pfeffer MA, Payne TH, Currier JS. Excess Patient Visits for Cough and Pulmonary Disease at a Large US Health System in the Months Prior to the COVID-19 Pandemic: Time-Series Analysis. J Med Internet Res 2020;22(9): doi:10.2196/21562

Shetty V, Morrison D, Belin T, Hnat T, Kumar S. Development and feasibility testing of a scalable system for passively monitoring oral health behaviors in the home setting. JMIR Mhealth Uhealth. 2020;8(6): doi:10.2196/17347

Morrison D, Laeyendecker O, Brookmeyer R. Cross-sectional HIV incidence estimation in an evolving epidemic. Stat Med. 2019;38(19):3614–27. doi:10.1002/sim.8196

Morrison D, Laeyendecker O, Konikoff J, Brookmeyer R. Cross-Sectional HIV Incidence Estimation with Missing Biomarkers. Stat Commun Infect Dis. 2018;10(1). doi:10.1515/scid-2017-0003

Laeyendecker O, Konikoff J, Morrison D, Brookmeyer R, Wang J, Celum C, et al. Identification and validation of a multi-assay algorithm for cross-sectional HIV incidence estimation in populations with subtype C infection. J Int AIDS Soc. 2018;21(2): doi:10.1002/jia2.25082

B. Under review

Morrison D. E., Nianogo R., Manuel V., Arah O. A., Anderson N., Kuo T., & Inkelas M. Modeling infection dynamics and mitigation strategies to support K-6 in-person instruction during the COVID-19 pandemic. Preprinted on medRxiv, 2021; doi:10.1101/2021.02.27.21252535

C. Software

Morrison D, Brookmeyer R. rwicc ("Regression with Interval-Censored Covariates"). R package: https://d-morrison.github.io/rwicc/

CHAPTER 1
Introduction

1.1 Background

Disease and epidemic surveillance provides important information to help reduce the spread of infectious diseases. (Gordis 2014) Accurate measures of disease incidence rates are crucial to this objective: incidence estimates help public health workers to efficiently allocate resources to populations experiencing high rates of transmission and to measure the effects of interventions. (Mastro 2013)

Traditionally, incidence has been estimated using cohort studies, in which a sample is recruited from the disease-free, at-risk portion of a population of interest; members of the cohort are then monitored periodically for signs of infection, and the incidence rate is estimated as the ratio of the number of participants who became infected to the amount of time the participants spent at-risk and under observation.

Cohort studies have several limitations. They are expensive to conduct, because of the need to follow large numbers of subjects for an extended time, and follow-up rates may be low especially in marginalized populations. They can take years to complete, and thus may not provide timely information for assessing the current growth rate of an epidemic. They are also vulnerable to selection biases in both recruitment and attrition. Selective attrition would arise, for example, if participants who are more likely to become infected over the course of follow-up are also more likely to be lost to follow-up.

Cross-sectional estimation is an alternative approach that can avoid some of the problems of cohort studies. This approach is based on a single representative cross-sectional

survey of the target population whose incidence rate we aim to measure. (Busch et al. 2010) Incidence is inferred from the prevalence of biomarker values associated with recent infection. This inference is based on assumptions about the dynamics of the epidemic under study, as described in Section 2.1. Cross-sectional studies require only a single in-person interaction with study staff per participant, which makes these studies faster to complete, avoids drop-out bias completely, and can reduce recruitment disparities, especially in marginalized subpopulations whose members might not want to be tracked longitudinally. Cross-sectional incidence estimation has been applied successfully in numerous settings, especially for measuring HIV incidence. (Brookmeyer and Quinn 1995; Brookmeyer, Laeyendecker, et al. 2013; Brookmeyer, Konikoff, et al. 2013; Laeyendecker et al. 2012; Konikoff et al. 2013; Laeyendecker et al. 2018)

In order to achieve accurate estimates, cross-sectional incidence estimation must be implemented carefully. Several common statistical challenges should be considered, including model specification, calibration to specific populations and time periods, and adjustments for nonuniform sampling, non-ignorable missingness, censoring, and measurement error. In short, cross-sectional incidence estimation analyses should be tailored to the population of scientific interest and to the data-generating process.

1.2 Overview of Dissertation

This dissertation examines three statistical complications that have arisen in a recent series of cross-sectional studies and presents methods to address these issues. Chapter 2 reviews the established cross-sectional estimation framework. Chapter 3 presents methods for handling incomplete data. (D. Morrison et al. 2018) Chapter 4 presents methods for

transporting results from training data sets to target populations. (D. Morrison et al. 2019) Chapter 5 presents methods for handling interval-censored event times. (D. Morrison et al. 2021) Chapter 6 summarizes the results and discusses themes connecting the work throughout this dissertation as well as future extensions. A consolidated list of notation is included before the references.

CHAPTER 2
The Cross-Sectional Incidence Estimation Framework

The cross-sectional incidence estimation framework, in simplest form, has the following structure. First, a "target" population is identified for which disease incidence will be estimated. A single representative cross-sectional survey of N participants is recruited from this target population. The survey participants provide biological specimens which are tested for infection status; specimens that test positive for infection are additionally assayed for a prespecified series of biomarkers that have some ability to help distinguish persons who were recently infected from those who have been infected for a long time. The cross-sectional incidence rate estimator is:

$$\hat{\iota} \stackrel{\text{def}}{=} \frac{V}{\mu \cdot N_u} \tag{2.1}$$

where V is the number of infected individuals in the survey whose biomarker measurements indicate the infection occurred recently, N_u is the number of individuals in the survey who are uninfected, and μ, the "mean window period," denotes the expected duration of time during which an infected person's biomarkers would indicate a recent infection. (Brookmeyer and Quinn 1995) The mean window period μ depends on the specific set of biomarkers assayed, as well as the operational definition for classifying an individual as a "recent infection" based on the biomarker values.

2.1 Approximate consistency of the cross-sectional incidence estimator

In this section, we show that $\hat{\iota}$ is an approximately consistent estimator of the target population's incidence rate, given some assumptions about the underlying data-generating distributions.

For an individual in the cross-sectional survey, let S denote the calendar time of infection. We define the incidence rate as the hazard function of S:

$$h(s) \stackrel{\text{def}}{=} p(S = s | S \geq s) = \frac{p(S = s)}{P(S \geq s)} \tag{2.2}$$

Here, we use lowercase $p(\cdot)$ to denote probability density functions, and uppercase $P(\cdot)$ to denote probability mass functions. Let t_0 denote the calendar time of the cross-sectional survey; then $h(t_0)$, the hazard rate in the target population at the time of the cross-sectional survey, is the estimand of primary interest. We assume that S has a continuous distribution; therefore, $P(S \geq s) = P(S > s)$ and thus:

$$h(s) = \frac{p(S = s)}{P(S > s)} \tag{2.3}$$

Now, let $T \stackrel{\text{def}}{=} t_0 - S$; i.e., $T = t$ means that the individual was infected t time units prior to the cross-sectional survey. Note that T is positive if the individual was infected before the survey, and T is negative if the individual was not infected until after the survey. Let $g(t)$ denote the probability density of T; i.e., $g(t) \stackrel{\text{def}}{=} p(T = t)$. Note that $p(S = s) = p(T = t_0 - s) = g(t_0 - s)$; we can thus rewrite Eq. 2.3 as:

$$h(s) = \frac{g(t_0 - s)}{P(S > s)} \tag{2.4}$$

Specifically, for $s = t_0$, we have:

$$h(t_0) = \frac{g(0)}{P(S > t_0)} \tag{2.5}$$

Next, let Y denote a binary classification of the individual's biomarker assay values at the time of the cross-sectional survey, where $Y = 1$ denotes a "positive" classification, associated

with recent infection, and $Y = 0$ denotes a "negative" classification, indicating a longstanding infection. The specifics of these classifications will be discussed in Section 2.2.2. Let $\phi(t) \stackrel{\text{def}}{=} P(Y = 1|T = t)$ be the conditional probability of a positive classification, given infection t units prior to t_0. Then by the law of total probability, $P(Y = 1) = \int_{t=0}^{\infty} P(Y = 1|T = t)p(T = t)dt = \int_{t=0}^{\infty} \phi(t)g(t)dt$. We assume that there is some point $t_{\max}$ beyond which $\phi(t) = 0$; then $P(Y = 1) = \int_{t=0}^{t_{\max}} \phi(t)g(t)dt$.

2.1.1 Scenario 1: $g(t)$ approximately constant

We might further assume that $g(t)$ is approximately a constant g for $t \in [0, t_{\max}]$; this assumption could be reasonable if the biomarker classification, Y, is defined such that $t_{\max}$ is short. Then we have:

$$P(Y = 1) \approx \int_{t=0}^{t_{\max}} \phi(t)g dt = g \int_{t=0}^{t_{\max}} \phi(t)dt \tag{2.6}$$

The quantity $\int_{t=0}^{\infty} \phi(t)dt = \int_{t=0}^{t_{\max}} \phi(t)dt$ is the mean duration of time during which a person has a positive biomarker classification. We call this duration the "mean window period" and denote it by μ; i.e.,

$$\mu \stackrel{\text{def}}{=} \int_{t=0}^{\infty} \phi(t)dt = \int_{t=0}^{t_{\max}} \phi(t)dt \tag{2.7}$$

Thus from Eq. 2.6, we have $P(Y = 1) \approx g \cdot \mu$, and therefore:

$$g(0) \approx g \approx P(Y = 1) \cdot \mu^{-1} \tag{2.8}$$

Substituting 2.8 into 2.5, we find that:

$$h(t_0) \approx \frac{P(Y = 1)}{P(S > t_0)} \cdot \mu^{-1} \tag{2.9}$$

In other words, if $g(t)$ is approximately constant for $t \in [0, t_{\max}]$, then the incidence rate at the time of the cross-sectional study is approximately equal to the probability of having a positive biomarker classification divided by the probability of being uninfected at the time of the survey, divided by the mean window period. We will denote this population-level quantity by ι:

$$\iota \stackrel{\text{def}}{=} \frac{P(Y=1)}{P(S > t_0)} \cdot \mu^{-1} \tag{2.10}$$

By the law of large numbers and the continuous mapping theorem, we can consistently estimate ι by plugging in the sample analog estimates $\hat{P}(Y=1) = V/N$ and $\hat{P}(S > t_0) = N_u/N$, where N is the number of survey participants, N_u is the number of uninfected survey participants, and $V = \sum_{i=1}^{N} Y_i$ is the number of biomarker-positive survey participants. Then we have:

$$\hat{\iota} \stackrel{\text{def}}{=} \frac{\hat{P}(Y=1)}{\hat{P}(S > t_0)} \cdot \mu^{-1} = \frac{V/N}{N_u/N} \cdot \mu^{-1} = \frac{V}{\mu N_u}$$

which is Eq. 2.1. Thus, if $g(t)$ is approximately constant over $t \in [0, t_{\max}]$, then $\hat{\iota}$ is an approximately consistent estimator for $h(t_0)$, the incidence rate at the time of the cross-sectional survey; i.e., $\hat{\iota} \to_P \iota \approx h(t_0)$, where "$\to_P$" denotes convergence in probability.

2.1.2 Scenario 2: $g(t)$ approximately linear

Now suppose that $g(t)$ is not constant over $t \in [0, t_{\max}]$ but is approximately linear in t; i.e., $g(t) \approx g(0) + \beta t$ for some β. Since $g(t) = p(T=t) = p(S = t_0 - t) = h(t_0 - t) \cdot P(S \geq t_0 - t)$, this assumption is valid if the hazard function $h(s)$ is approximately linear in s for $s \in (t_0 - t_{\max}, t_0)$ and $P(S \geq s) = \exp\{-\int_{u=-\infty}^{s} h(u)du\}$ is approximately constant in s for $s \in (t_0 - t_{\max}, t_0)$. Under this assumption, (2.6), (2.8), and (2.9) no longer hold; instead, we

have:

$$\begin{aligned} P(Y=1) &= \int_{t=0}^{t_{\max}} \phi(t)g(t)dt \\ &\approx \int_{t=0}^{t_{\max}} \phi(t)[g(0)+\beta t]dt \\ &= g(0)\int_{t=0}^{t_{\max}} \phi(t)dt + \beta\int_{t=0}^{t_{max}} t\cdot\phi(t)dt \\ &= g(0)\mu + \beta\int_{t=0}^{t_{\max}} t\cdot\phi(t)dt \end{aligned}$$

Let $\psi \stackrel{\text{def}}{=} \mu^{-1}\int_{t=0}^{t_{\max}} t\cdot\phi(t)dt$, so that $\mu\psi = \int_{t=0}^{t_{\max}} t\cdot\phi(t)dt$. Then:

$$\begin{aligned} P(Y=1) &\approx g(0)\mu + \beta\mu\psi \\ &= [g(0)+\beta\psi]\cdot\mu \\ &= g(\psi)\cdot\mu \end{aligned} \tag{2.11}$$

Dividing both sides of Eq. 2.11 by μ, we have:

$$g(\psi) \approx P(Y=1)\cdot\mu^{-1} \tag{2.12}$$

Further, suppose that the hazard rate $h(t)$ is not very large, such that $P(S > t_0 - \psi) \approx P(S > t_0)$; then by evaluating Eq. 2.4 at $s = t_0 - \psi$ and applying Eq. 2.12, we have:

$$h(t_0-\psi) = \frac{g(\psi)}{P(S>t_0-\psi)} \approx \frac{P(Y=1)}{P(S>t_0)}\cdot\mu^{-1} = \iota \tag{2.13}$$

Thus, if $g(t)$ is approximately linear in t over $t \in [0, t_{\max}]$, then $\hat{\iota}$ is an approximately consistent estimator for $h(t_0-\psi)$, the incidence rate ψ time units prior to the date of the cross-sectional survey; i.e.,

$$\hat{\iota} \to_P \iota \approx h(t_0-\psi) \tag{2.14}$$

We refer to ψ as the "shadow" of $\hat{\iota}$. If ψ is not too large and $h(s)$ does not change too quickly, then $h(t_0 - \psi) \approx h(t_0)$, and $\hat{\iota}$ is still approximately consistent for $h(t_0)$, as in the case where $g(t)$ was approximately constant. Note that ψ, like μ, is a function of $\phi(t)$ and thus depends on the biomarker classification rules used to define Y.

More generally, if $g(t)$ is a nonlinear function of t, then by a Taylor series approximation it can be shown that:

$$\iota - h(t_0 - \psi) \approx \frac{1}{P(S > t_0)} \cdot \frac{g''(\psi)}{2} \cdot \sigma^2 \tag{2.15}$$

where $\sigma^2 = \mu^{-1} \int_{t=0}^{t_{\max}} (t - \psi)^2 \phi(t) dt$. (E. H. Kaplan and Brookmeyer 1999; Konikoff 2015) Note that if $g(t)$ is approximately linear in t, i.e., if $g''(\psi) \approx 0$, then Eq. 2.15 entails that $\iota \approx h(t_0 - \psi)$ as in Eq. 2.13. Alternatively, if σ^2 is sufficiently small relative to $g''(\psi)$, we again have $\iota \approx h(t_0 - \psi)$.

Additionally, consider the special case when the epidemic is not advanced, and specifically we mean the case when most individuals in the population are uninfected; i.e., $P(S > t_0) \approx 1$. Then:

$$h(t_0 - \psi) \approx \iota = \frac{P(Y = 1)}{P(S > t_0)} \cdot \mu^{-1} \approx P(Y = 1) \cdot \mu^{-1} \tag{2.16}$$

This expression can be recognized as an algebraic rearrangement of the relationship "Prevalence = Incidence × Mean Duration of Disease", where the disease in this case is the condition of a positive biomarker classification; i.e., $Y = 1$. This relationship has been previously derived in the epidemiological literature under steady state conditions. (Freeman and Hutchison 1980)

2.1.3 Effects of migration and mortality on cross-sectional incidence estimation

The participants in the cross-sectional survey may not have always been in the target population for which we are trying to estimate inference; this possibility affects the interpretation of cross-sectional incidence estimates. For example, consider the first time when someone with coronavirus disease 2019 (COVID-19) entered the United States; at that moment, the prevalence of COVID-19 in the United States was a consequence of past incidence in the population where that person contracted COVID-19; the past incidence of COVID-19 in the U.S. population was precisely 0. Similarly, if the target population is defined partly based on age, some of the individuals who were in the target population at the time of the cross-sectional survey might not yet have been in the target population when they became infected. Thus, if the shadow parameter is substantially larger than 0, then $\hat{\iota}$ may be estimating a hazard function representing a mixture distribution across multiple populations.

To model this possibility, let us assume that every participant in the cross-sectional survey has been a member of a particular population at each calendar time s prior to t_0. Let $W(s)$ be a categorical stochastic process representing a given participant's population affiliation status at calendar time s. Let $\mathcal{W}$ denote the support of $W(s)$; i.e., the set of populations of which the survey participants could have been members. $\mathcal{W}$ can include populations in other geographic areas, younger age groups or other subpopulations in the same geographic area from which individuals can enter the target population, and a default category representing individuals who were not alive yet at time s. Let τ be the element of $\mathcal{W}$ denoting the target population, in which the cross-sectional survey is performed.

At the start of this chapter, we defined the variables S, T, and Y as characteristics of participants in the cross-sectional survey of the target population at calendar time t_0. Thus, all of the preceding probability expressions have been implicitly conditional on $W(t_0) = \tau$; for example, letting $h(s|A) = p(S = s|S \geq s, A)$ where A is any stochastic event, we could have more explicitly written $h(s) \stackrel{\text{def}}{=} p(S = s|S \geq s)$ as $h\big(s \,\big|\, W(t_0) = \tau\big) \stackrel{\text{def}}{=} p(S = s|S \geq s, W(t_0) = \tau)$. In the rest of this chapter, we will continue to use the less-explicit notation, in order to make the expressions more concise and readable, but the condition $W(t_0) = \tau$ should always be understood to be present.

In the consistency result $\hat{\iota} \rightarrow_P \iota \approx h(t_0 - \psi)$ derived in the previous subsection (Eq. 2.14), $h(t_0 - \psi)$ is not necessarily equal to the hazard experienced at $t_0 - \psi$ by the individuals who were in τ at both t_0 and $t_0 - \psi$; i.e., it is not necessarily true that $h(t_0 - \psi) = h(t_0 - \psi|W(t_0 - \psi) = \tau)$. Instead, for calendar time $s < t_0$, $h(s)$ is a mixture of the incidence rates that the target population's eventual members experienced at s, in the populations which they were members of at time s:

$$\begin{aligned} h(s) &= p(S = s|S \geq s) \\ &= \sum_{w \in \mathcal{W}} p(S = s|S \geq s, W(s) = w) P(W(s) = w|S \geq s) \\ &= \sum_{w \in \mathcal{W}} h(s|W(s) = w) P(W(s) = w|S \geq s) \end{aligned} \tag{2.17}$$

If we assume that nearly all of the members of the target population at t_0 who had not yet been infected at $t_0 - \psi$ were already in the target population at $t_0 - \psi$, i.e., if we assume that $P(W(t_0 - \psi) = \tau|S \geq t_0 - \psi) \approx 1$, then from Eq. 2.17, we have:

$$h(t_0 - \psi) \approx h(t_0 - \psi|W(t_0 - \psi) = \tau)$$

Then from Eq. 2.14 we have:

$$\hat{\iota} \rightarrow_P \iota \approx h(t_0 - \psi | W(t_0 - \psi) = \tau)$$

The assumption that $P(W(t_0 - \psi) = \tau | S \geq t_0 - \psi) \approx 1$ is most plausible if ψ is short, limiting the plausibility that a substantial portion of the target population immigrated or aged into the target population during the interval $(t_0 - \psi, t_0)$.

Likewise, recall that the estimand of primary interest is the current incidence rate in the target population, $h(t_0)$, rather than the past incidence rate $h(t_0 - \psi | W(t_0 - \psi) = \tau)$. It is more plausible that these rates are similar if ψ is short. Additionally, in order to derive the approximation $\iota \approx h(t_0 - \psi)$ in Eq. 2.14, we assumed that $g(t) = h(t_0 - t) \cdot P(S \geq t_0 - t)$ is approximately linear for $t \in (0, t_{\max})$. Eq. 2.17 shows that for $s < t_0, h(s)$ is a mixture of hazards in different populations with mixing weights $P(W(s) = w | S \geq s)$ that can vary with s. Even if the component hazard functions $h(s|W(s) = w)$ are changing linearly and $P(S \geq s)$ is approximately constant, nonlinearity in the mixing weights $P(W(s) = w | S \geq s)$ could induce nonlinearity in $h(s)$. Thus, it is highly desirable that the biomarker classification Y be defined so as to minimize ψ.

Note that death, emigration, and other forms of exit from the target population prior to the cross-sectional survey have not played a role in these calculations, because all of the expressions are conditional on being alive and in the target population at the time of the cross-sectional survey.

2.1.4 Summary of preceding results

In summary, the chain of approximations underlying the cross-sectional incidence estimation approach is:

$$\hat{\iota} \rightarrow_P \iota \approx h(t_0 - \psi) \approx h(t_0 - \psi | W(t_0 - \psi) = \tau) \approx h(t_0)$$

We have justified these approximations using the following four assumptions:

1. The probability of a positive biomarker classification is approximately 0 for infections lasting longer than $t_{\max}$ time units: $P(Y = 1|T = t) \approx 0, t > t_{\max}$.

2. The density of infection duration for individuals in the target population at the time of the cross-sectional survey is approximately linear for $t_{\max}$ units prior to the survey: $g(t) \approx \alpha + \beta t, t \in [0, t]$.

3. Nearly all of the current members of the target population who had not yet been infected ψ time units prior to the survey were living in the target population at that time: $P(W(t_0 - \psi) = \tau | S \geq t_0 - \psi) \approx 1$.

4. The current incidence rate in the target population is approximately equal to the incidence rate for individuals who were in the target population both currently and ψ time units ago: $h(t_0) \approx h(t_0 - \psi | W(t_0 - \psi) = \tau)$.

2.2 Calibration of cross-sectional incidence estimators

A requirement of the cross-sectional method before it can be applied in practice is: (1) a set of biomarkers to assay has been identified; (2) it has been established how to combine those biomarkers to create an operational definition of "recent infection"; and (3) an estimate of μ based on that biomarker definition of a recent infection is available. These three critical pieces of information are typically determined from an initial "calibration" data set of infected individuals. This data set must include biomarker measurements on biological specimens from the participants and, in contrast with the cross-sectional survey, must

include the duration of infection for each specimen, at least in interval-censored form. Information on infection duration is typically available when the data come from longitudinal studies of initially-uninfected participants who are tested for infection periodically. Between study visits, some of the participants will become infected; these participants then contribute additional blood samples at various time points after infection, which can be used to model the evolution of biomarker distributions as a function of infection duration.

The dates of the last negative and first positive diagnostic tests can be interpreted as endpoints of a censoring interval $[L, R]$, referred to as the "seroconversion window", and the exact seroconversion date (i.e., the moment at which the individual would first be infection-positive if tested) can be imputed within this interval (details in Section 2.2.3 and Chapter 5). Given an imputed seroconversion date, the imputed duration of infection for each of the participant's subsequent biomarker measurements is calculated as the difference between the imputed seroconversion date and the biomarker measurement date. We define duration of infection starting from the seroconversion date, rather than the date of exposure, because without detailed contact tracing of each participant's points of possible exposure, it can be difficult to identify a lower bound for the exact date of exposure; for some infections, including HIV, there can be a substantial delay between exposure and the development of detectable levels of infection biomarkers such as antigens or antibodies. A participant may have been infected even before the last visit in which they appeared to be uninfected.

The need for a longitudinally collected calibration data set does not invalidate the use of cross-sectional incidence estimation approach to avoid the problems of cohort-based

incidence estimation. A single longitudinal data set can be used to calibrate multiple subsequent cross-sectional studies, thus providing efficiency gains compared to conducting multiple cohort studies. Furthermore, the requirements for validity of the calibration data set are less stringent than the requirements for a cohort study for incidence estimation; since only the data of infected participants will be used, selective attrition may be less of a concern, if attrition depends on factors that affect infection probability but not subsequent biomarker distributions.

If a calibration data set is assumed to have been sampled from a population with the same biomarker distributions (conditional on time since infection) as the target population, then the calibration procedure can proceed as follows.

2.2.1 Direction of biomarker association

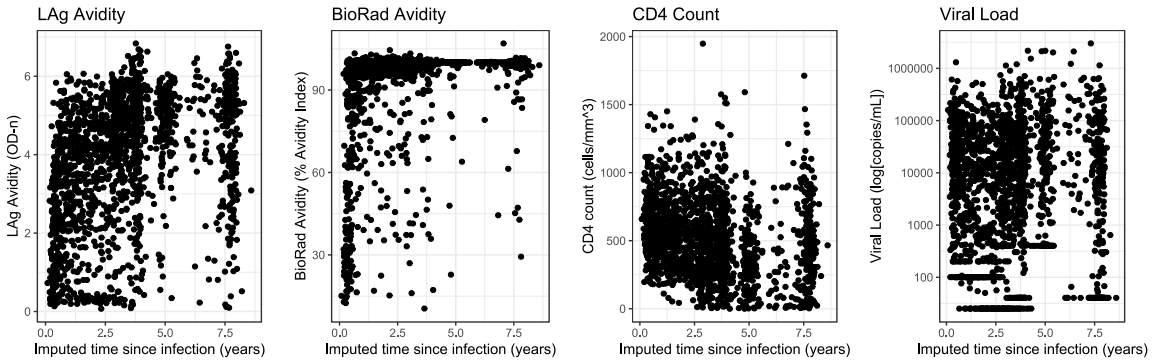
First, the direction of association between each biomarker and time must be determined. This can be done based on domain knowledge about the underlying biological processes or by graphing or regressing each biomarker against infection duration, as imputed (for example) by the midpoint of the censoring interval (Figure 2.1). Biomarkers useful for cross-sectional incidence estimation will have a monotonic relationship (in expectation) with infection duration, either increasing or decreasing.

2.2.2 Construction of multi-assay algorithms for recency classification

For biomarkers that appear to have such a relationship, a grid of possible dichotomization cutoff values is selected. Multi-assay algorithms (MAAs) for classifying samples as "recent" (MAA-positive) or "non-recent" are then defined by selecting one dichotomization value for each biomarker. (Brookmeyer, Konikoff, et al. 2013) For example, suppose there are two

continuous-valued biomarkers, B_1 and B_2, with cutoff options $\{a, b, c\}$ and $\{d, e\}$ respectively; further suppose that $E[B_1|T = t]$ increases with infection duration $T = t$, whereas $E[B_2|T = t]$ decreases with t. Then a possible MAA would be $Y = 1_{\{B_1<a \,\&\, B_2>d\}}$, where $\mathbf{B} = (B_1, B_2)'$ is the vector of biomarker values; that is, classify a serum sample as MAA-positive ($Y = 1$) if $B_1 < a$ and $B_2 > d$, and MAA-negative ($Y = 0$) otherwise. Another MAA would be $Y = 1_{\{B_1<c \,\&\, B_2>e\}}$. A biomarker increasing with time can be effectively ignored by setting its cutoff to $+\infty$, and a biomarker decreasing with time can be ignored by setting its cutoff to $-\infty$. Note that MAAs are defined by intersections of acceptance regions, and thus can be evaluated sequentially for efficiency; the more expensive, labor-intensive, or time-consuming biomarkers only need to be assayed for a particular specimen if the other biomarkers allow the possibility of a "recent" classification.

Figure 2.1: Measured values of the LAg Avidity, BioRad Avidity, CD4 cell count, and viral load biomarkers versus midpoint-imputed infection duration, in a data set of Clade B HIV infections. (Brookmeyer, Konikoff, et al. 2013)



2.2.3 Estimation of the mean window period

For each MAA, we can estimate the mean window period μ, using the fact that $\mu = \int_0^{t_{\max}} \phi(t)dt$ where $\phi(t) \stackrel{\text{def}}{=} P(Y = 1|T = t)$ (Eq. 2.7). We can estimate $\phi(t)$ from the calibration data set by regression modeling. (Brookmeyer, Laeyendecker, et al. 2013) Then

$\hat{\mu} = \int_{t=0}^{t_{\max}} \hat{\phi}(t)dt$. To account for uncertainty in the imputation of t, multiple imputation can be performed by repeated random draws from the uniform distribution on the censoring interval $[L, R]$. (Alternative procedures will be considered in Chapter 5). For each randomly imputed data set, a regression model for $P(Y = 1|T = t)$ can be estimated, and the coefficients of this model can be averaged across the imputed data sets to generate a consensus model $\hat{\phi}(t)$. (Brookmeyer, Konikoff, et al. 2013)

The regression model used is typically logistic regression; that is, a generalized linear model with a Bernoulli distribution and logit (log-odds) link function. (Dobson and Barnett 2008) The linear predictor's functional form is determined based on regression diagnostics, and can include flexible forms such as polynomial splines. (Konikoff 2015)

Bootstrapping can be used to quantify the uncertainty in the estimate $\hat{\mu}$. (Efron 1979) Considering the observed sample distribution as a nonparametric estimate of the population distribution, the calibration data set can be resampled to create bootstrapped data sets, and the entire μ estimation procedure is repeated for each bootstrapped data set. The $\alpha/2$ and $1 - \alpha/2$ quantiles of the resulting distribution of $\hat{\mu}$ values are then used as the limits of a $(1 - \alpha) \times 100\%$ confidence interval.

2.2.4 Optimal MAA selection

Next, an optimal MAA (i.e., an optimal set of cutoff values) is selected to be used for cross-sectional estimation. As discussed in Section 2.1, it is helpful for ψ to be small in order to make the approximation $\iota \approx h(t_0)$ more plausible. Then since $\hat{\iota}$ is consistent for ι, it follows that $\mathrm{E}[\hat{\iota}] \to \iota \approx h(t_0)$; that is, $\hat{\iota}$ is asymptotically approximately unbiased for $h(t_0)$, as long as ψ is sufficiently small. Hence, our main objectives in selecting an optimal MAA are to

minimize $\mathrm{Var}(\hat{\iota})$ and ψ.

By the Conditionality Principle (Birnbaum 1962; Royall 1986) we consider the conditional variance:

$$\mathrm{Var}(\hat{\iota}|N_u) = \mathrm{Var}\left(\frac{V}{\mu N_u}\middle|N_u\right) = \frac{1}{(N_u\mu)^2}\mathrm{Var}(V|N_u) \tag{2.18}$$

Typically, V is much smaller than N_u; then $\mathrm{Var}(V|N_u) \approx \mathrm{Var}(V|N_u + V)$. Let X denote HIV seroconversion status (1 = seropositive, 0 = seronegative); then $V|N_u + V \sim \mathrm{Binom}(N_u + V, \pi)$, where $\pi \stackrel{\text{def}}{=} P\big(Y = 1 \,\big|\, Y = 1 \cup X = 0\big)$. Since π is typically small:

$$\mathrm{Var}(V|N_u + V) = (N_u + V)(\pi - \pi^2) \approx (N_u + V)\pi \approx N_u\pi$$

Further, using ∪ to denote union and ∩ to denote intersection, we have:

$$\begin{aligned}
\pi &\stackrel{\text{def}}{=} P(Y = 1|Y = 1 \cup X = 0) \\
&= \frac{P(Y = 1)}{P(Y = 1 \cup X = 0)} \\
&\approx P(Y = 1) \text{ [assuming } P(Y = 1 \cup X = 0) \approx 1] \\
&\approx g(\psi)\,\mu \text{ [by Eq. 2.11]}
\end{aligned}$$

Returning to Eq. 2.18, we now have:

$$\mathrm{Var}(\hat{\iota}|N_u) \approx \frac{N_u\, g(\psi)\,\mu}{(N_u\,\mu)^2} = \frac{g(\psi)}{N_u\,\mu} \tag{2.19}$$

Thus, the approximate conditional variance of $\hat{\iota}$ is inversely related to μ. Assuming that $g(\psi)$ does not change too substantially with ψ, we would minimize $\mathrm{Var}(\hat{\iota})$ by choosing an MAA that maximizes μ.

Unfortunately, μ and ψ are positively correlated; as the mean window period becomes longer, so too does the shadow. Thus, we are faced with a tradeoff between precision and lag

time, analogous to a classical bias-variance tradeoff. Typically, the optimal MAA is defined as the one that maximizes μ, subject to the constraint that the upper 95% bootstrap confidence interval for ψ is less than 365 days. (Brookmeyer, Konikoff, et al. 2013) This constraint increases the likelihood that the resulting cross-sectional incidence estimate will correspond to recent incidence in the population of interest.

Note that selecting an MAA to maximize $\hat{\mu}$ will induce some bias, analogous to model selection bias and the phenomenon of regression to the mean; thus, the $\hat{\mu}$ estimate for the chosen MAA generated from the data set that was used to select that MAA can no longer be assumed to be a consistent estimate of μ. However, the induced bias might not be very large. A simulation study estimated a bias of approximately 3 days when selecting an optimal MAA out of a set of 31,680 MAAs involving up to four biomarkers. (Konikoff 2015)

2.3 Uncertainty quantification for cross-sectional incidence estimation

Confidence intervals for cross-sectional incidence estimates should account for the uncertainty both in estimating the prevalence of MAA-positive infection and in relating that prevalence to incidence via the μ parameter. Parametric methods have been developed for this purpose. (Brookmeyer 1997; Cole et al. 2007)

2.4 HIV Biomarker Data Sets

The calibration data set that we will consider in this dissertation consists of 2,442 samples of HIV Clade C infections from 278 participants with interval-censored durations of infection (approximately 0.1 to 9.9 years after seroconversion; see Table 2.1). These samples were obtained from three cohort studies which recruited individuals who had acquired HIV infections while enrolled in clinical trials evaluating interventions for HIV prevention.

2.4.1 The CAPRISA 004 and 002 studies

The CAPRISA 004 study, conducted by the Centre for the AIDS Programme of Research in South Africa (CAPRISA), was a randomized controlled trial of a vaginal antiretroviral microbicide to prevent HIV infection. (Abdool Karim et al. 2010) 1085 initially HIV-negative women living in KwaZulu-Natal, South Africa, were recruited between 2007 and 2009 and randomly assigned into treatment and placebo study arms. Each participant was scheduled to return monthly for HIV testing for 30 months. 98 participants acquired HIV during the trial; these participants were offered enrollment into the CAPRISA 002 Acute Infection cohort study, which included blood sample collection biweekly for the first three months after enrollment, then monthly until 12 months, then every three months or as medically indicated, until antiretroviral therapy initiation or at least two years from seroconversion. (Garrett et al. 2015). 518 of these blood samples from 90 participants, collected between one month and four years after detection of seroconversion, were subsequently assayed for several biomarkers of recent infection. (Laeyendecker et al. 2018)

2.4.2 The FHI 360 HC-HIV and GS studies

The Hormonal Contraception and Risk of HIV Acquisition (HC-HIV) study, conducted by Family Health International (FHI) 360, evaluated hormonal contraception and HIV infection. (C. S. Morrison et al. 2007) 4439 initially HIV-uninfected women seeking healthcare services from family planning clinics in Uganda and Zimbabwe were recruited for the study between November 1999 and January 2004. Follow-up HIV tests were conducted every 12 weeks for 15-24 months; 213 of those 4439 participants tested positive at some point after enrollment. 188 of these 213 participants were then enrolled in the Hormonal Contraception and HIV Genital Shedding and Disease Progression (GS) Study, which included blood sample

collection at 4, 8, and 12 weeks following enrollment in the GS study and then at 12-week intervals for up to 9.3 years (C. S. Morrison et al. 2011) 1,839 blood samples from 162 participants were subsequently assayed for several biomarkers of recent infection; all but three of these samples came from the participants in Zimbabwe. (Laeyendecker et al. 2018)

2.4.3 The HPTN 039 and 039-01 studies

The HIV Prevention Trials Network (HPTN) 039 study evaluated the effects of herpes simplex virus type 2 treatment on HIV acquisition risk. (Reid et al. 2010) 602 participants were recruited from a study site in Lusaka, Zambia between October 2003 and November 2007. HIV testing was performed quarterly, for up to 18 months. Participants who acquired HIV infections during the course of the study were invited to join the HPTN-039-01-Ancillary study. (Celum and Wald 2004) These participants had blood samples collected at enrollment in the ancillary study and at 1, 5, and 6 months after enrollment. 85 of these blood samples, from 25 individuals, were subsequently assayed for several biomarkers of recent infection. (Laeyendecker et al. 2018)

2.4.4 The HPTN-068 study

An additional sample set was obtained from an independent, longitudinal cohort study that evaluated the impact of conditional cash transfer on HIV acquisition by young women in South Africa (HPTN 068). The study was conducted from 2012 to 2015. Samples collected in 2014 were used for cross-sectional incidence estimation; these results can be compared to the observed longitudinal incidence in the cohort. This analysis included 1,360 participants (1,269 HIV-uninfected and 91 HIV-infected participants; 61 participants were infected in 2013 or earlier). The observed longitudinal incidence in HPTN 068 in the 2014 survey was

1.9% (95% CI: 1.3, 2.7).

Table 2.1: Demographics of several study cohorts of HIV Subtype C infected individuals

	Cohort		
Characteristic	CAPRISA 002	FHI-360 GS	HPTN 039-01
Country of origin	South Africa	Zimbabwe*	Zambia
Number of samples	518	1,839	85
Number of unique subjects	90	162	25
Range of duration of infection in years	0.06 to 3.7	0.04 to 9.9	0.15 to 0.8
Mean samples per subject (range)	6 (1-7)	12 (1-20)	4 (1-4)
Female sex, % of subjects	100%	100%	100%
Number samples from subjects on ART (%)	12 (2.2%)	220 (11.3%)	0 (0%)
Duration of infection in years			
0.0 to 0.5	159	306	42
0.5 to 1.0	173	262	43
1.0 to 2.0	88	448	0
2.0 to 3.0	76	105	0
3.0 to 5.0	22	347	0
≥ 5.0	0	371	0
CD4 cell count			
>500	228	685	54
500-200	271	822	26
<200	14	104	0
missing	5	228	5
Viral load (copies/mL)			
>10,000	260	560	37
10,000 to 1,000	161	278	26
<1,000	92	227	19
missing	5	774	3

**All participants from South Africa, Zimbabwe and Zambia were assumed to have subtype C infection based on the prevalence of subtype C in those countries. The FHI-360 cohort included one individual from Uganda with three samples. That individual was infected with HIV subtype C based on subtype assessment of the pol region.*

CHAPTER 3
Missing Biomarker Data

One key advantage of MAAs is that they can significantly reduce assay costs compared to testing all biological samples with all biomarkers. The cost savings results because only samples provisionally classified as MAA positive need to be tested in the next step of the algorithm with another biomarker in order to determine the MAA classification. Furthermore, the order of the steps of the algorithm can be arranged to help minimize costs: the less expensive or less labor-intensive assays can be performed in the initial steps, and the most expensive assays performed in the final step. By so doing, the number of biological samples that need to be tested in the final step is relatively small. For example, in HIV incidence estimation, viral load testing is often placed at the final step of MAAs.

However, exploiting this opportunity for cost-savings leads to incomplete data sets, which can become a problem when attempting to apply MAAs different from those for which the data set was originally intended. Biomarker measurements could also be missing from a data set because there was insufficient biological sample material to adequately perform some of the assays, because biological samples have been lost, or because of other logistical and administrative reasons.

The objective of this chapter is to consider some of the statistical challenges for addressing missing biomarker data for cross-sectional incidence estimation. In Section 3.1 we consider several methods for handling missing biomarkers when estimating the mean window period of an MAA based on biological samples with approximately known durations of infection. In Section 3.1.3 we evaluate the performance of these various methods by simulation. We examine two naïve approaches, one using all samples that can be classified

by the MAA and another using all samples with complete biomarker data, and we show that each of these approaches can lead to biased estimators of the mean window period. We propose a conditional approach for handling the missing data. The main idea of the conditional approach with two biomarkers is to decompose the likelihood into a product of two factors. The first factor corresponds to the first biomarker result and the second factor corresponds to the second biomarker result conditional on the first biomarker result. This approach accounts for the possibility that the missingness mechanism for a given biomarker may depend on the values of the other biomarkers in the MAA. We show that this method performs well. In Section 3.1.5, we apply these methods in practice to calibration data set of biomarker data collected longitudinally from individuals with HIV Subtype C infections.

Once an MAA has been developed and its mean window period has been estimated, it is ready to be used in a cross-sectional survey to estimate incidence. However, missing biomarker data can also occur in the survey for any of the reasons mentioned previously. In Section 3.2 we consider approaches for handling missing biomarkers in a cross-sectional survey for estimating incidence, and in Section 3.2.1 we evaluate these approaches by simulation. We again show that several naïve estimators can lead to biased results, whereas a conditional approach performs well. The results are discussed in Section 3.3.

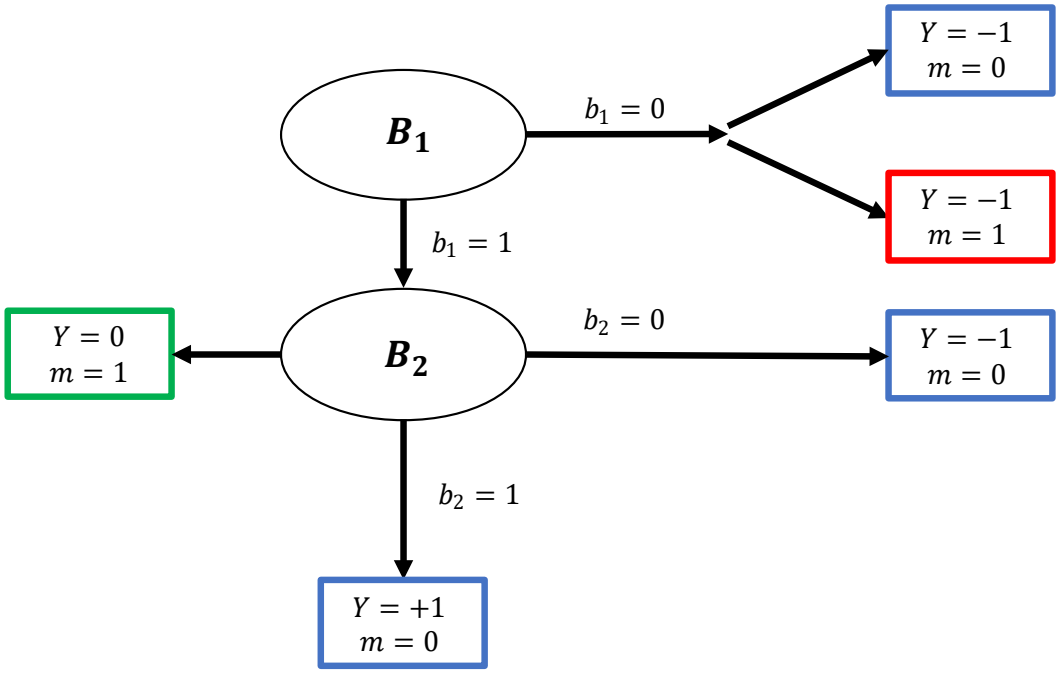
3.1 Estimating the mean window period with incomplete calibration data

In this section we consider approaches for handling missing biomarker data when estimating the mean window period of an MAA from a calibration data set. Here we consider algorithms consisting of two biomarkers assayed sequentially, where missing biomarker data is only possible for the second biomarker in the algorithm. One example of a two

biomarker MAA for HIV incidence estimation is a widely used algorithm based on LAg-Avidity assay and viral load, for which MAA positivity requires LAg-Avidity < 1.5 normalized optical density units (OD-n) and viral load > 1000 copies/mL. (Rehle et al. 2015) We label the two biomarkers B_1 and B_2 where biomarker B_2 is the one with potentially missing values. We use the notation m_i to indicate whether the value of the biomarker B_2 is missing for the i^{th} serum sample, that is, $m_i = 1$ if the B_2 measurement is missing and $m_i = 0$ if the measurement is known. We define the indicator random variables b_{1i} and b_{2i} to indicate whether the biomarkers meet the criteria for MAA positivity for biomarkers B_1 and B_2 respectively. The random variable b_{2i} is only observed if $m_i = 0$. For example, in the two-assay MAA mentioned above (LAg and viral load), b_{1i} is set to 1 if LAg-Avidity < 1.5 and 0 otherwise, and b_{2i} is set to 1 if viral load > 1000, to 0 if viral load ≤ 1000 and is not observed if $m_i = 1$. Let Y_i indicate the MAA classification for the i^{th} sample based on the observed data, where $Y_i = +1$ if the MAA is positive (which occurs if $b_{1i} = b_{2i} = 1$), $Y_i = -1$ if the MAA is negative (which occurs if either $b_{1i} = 0$ or $b_{2i} = 0$), and $Y_i = 0$ if the MAA classification is indeterminate (which occurs if $b_{1i} = 1$ and $m_i = 1$).

Figure 3.1 shows the classification of biological samples according to MAA status (MAA positive, negative and indeterminate) and biomarker B_2 missing status ($m = 0$ or 1). As shown in the figure, some biological samples (the red squares) can be classified as MAA negative ($Y = -1$) even though biomarker B_2 is missing. The blue squares correspond to samples with $m_i = 0$, that is, samples for which both biomarkers have been measured. The green square corresponds to samples where the MAA status could not be determined (that is, where $b_1 = 1$ and $m = 1$).

Figure 3.1: Flow-chart of the sample classification process for a two-biomarker MAA with missing data on the second biomarker.



MAA positive, negative, and indeterminate are indicated by $Y = +1, -1, 0$ respectively. Methods WP1 and CS1 use all samples where Y is determined (red and blue squares); methods WP2 and CS2 use samples only if B_2 is observed (blue squares); methods WP3 and CS3 use all samples (red; blue; and green squares where MAA is indeterminate).

As discussed in Chapter 2, the calibration data used for window period estimation typically consist of biological samples from seropositive individuals along with interval-censored durations of infection. Thus, let us denote the data from the i^{th} biological sample as $(L_i, R_i, m_i, b_{1i}, b_{2i})$, where L_i and R_i are the left and right bounds respectively for the duration of infection at the time of biological sample collection. Our analysis task is to accurately estimate a model for $\phi(t) = P(b_{1i} = b_{2i} = 1 | T_i = t)$, from which we will derive an estimate of μ.

3.1.1 Assumptions

We will evaluate several methods for estimating the mean window period under a class of missingness mechanisms applicable to MAAs. We suppose that the probability that

biomarker B_2 is missing is independent of b_{2i}, given b_{1i}; i.e., $P(m_i = 1|b_{1i} = 0, b_{2i}) = P(m_i = 1|b_{1i} = 0) = \lambda_1$, and similarly $P(m_i = 1|b_{1i} = 1, b_{2i}) = P(m_i = 1|b_{1i} = 1) = \lambda_2$. In the practical application of MAAs we would typically expect $\lambda_1 > \lambda_2$, because biological samples that meet the MAA criterion for the first biomarker are more likely to be assayed for the second biomarker. If the biomarker tests were run strictly sequentially, then $\lambda_1 = 1$. However, in practice some biological specimens may have been evaluated for biomarker B_2 even if $b_{1i} = 1$, because of other research or clinical requirements. In that case λ_1 may be less than 1. In the following analyses, while we are considering primarily sequentially-run evaluations of the biomarkers, we also allow for the possibility that some measurements may be available on B_2 even if $b_{1i} = 0$.

It is important to note that even if the probability that B_2 is missing does not depend on B_1 (i.e., $\lambda_1 = \lambda_2 = \lambda$), the probability that the MAA classification is indeterminate still depends on the values of the biomarkers. To see why, note that $P(Y_i = 0|b_{1i} = 1) = P(m_i = 1|b_{1i} = 1) = \lambda_2$, but on the other hand, we have $P(Y_i = 0|b_{1i} = 0) = 0$. It is also worth noting that although $Y_i = -1$ whenever $b_{1i} = 0$, regardless of whether b_{2i} is observed, the value of $\lambda_1 = P(m_i = 1|b_{1i} = 0)$ still affects the simulation results below, because the second approach considered removes all observations with b_{2i} missing, even when the MAA classification is already determined by $b_{1i} = 0$; we will find that this approach leads to biased results specifically when $\lambda_1 \neq \lambda_2$.

3.1.2 Analysis approaches

The first approach that we consider for window period estimation (Method WP1) uses all samples whose MAA classification has been determined (i.e., where Y_i takes the values of +1

or -1) and ignores those samples whose MAA status is undetermined. We thus refer to this approach as the "MAA determined analysis". The data that would go into the analysis are the red and blue squares in Figure 3.1. We could fit a flexible logistic regression model for $P(Y_i = 1)$, such as a cubic spline as a function of the time since seroconversion to the data (Y_i, T_i) where Y_i is either -1 or +1. The resulting fitted predicted probability curve, denoted $\hat{\phi}(t)$, is integrated from 0 to ∞ to estimate the mean window period; i.e., $\hat{\mu} = \int_{t=0}^{\infty} \hat{\phi}(t)dt$. Method WP1 has the appeal of seeming to be simple and straightforward, because it uses all the data where the MAA classification can be determined, but it is a naïve analysis that can lead to unwanted selection effects and biased estimates if $\lambda_2 > 0$. We quantify this bias by simulation in Section 3.1.3.

The second approach we consider (Method WP2) uses all samples where both biomarkers have been measured; that is, sample i will be excluded from the analysis whenever $m_i = 1$, even if $b_{1i} = 0$, in which case the sample can be classified as MAA-negative. Thus, the data that would go into this analysis are only the blue squares in Figure 3.1. We refer to this approach as the "complete biomarker analysis" as it uses only samples where both biomarkers have been measured. The intuition for this approach is that we should remove all samples with incomplete data, rather than only samples with B_2 missing and $b_1 = 1$, because the subset of samples with $b_1 = 1$ has a higher proportion of MAA-positives than the overall population; therefore, removing incomplete observations from only this subset induces bias. In contrast, if the probability of missingness is not associated with the underlying biomarker values (implying $\lambda_1 = \lambda_2$), then removing all incomplete samples leaves us with an unbiased sample of biomarker values. However, method WP2 is

still biased whenever $\lambda_1 \neq \lambda_2$. In particular, as we show in Section 3.1.3, even if $\lambda_2 = 0$, the method will lead to biased results when $\lambda_1 > 0$.

The third approach we consider (Method WP3) uses all the biological samples illustrated in Figure 3.1 (blue, red, and green squares). We call this approach a conditional likelihood analysis. The factor that is contributed to the likelihood function for the i^{th} individual is:

$$\mathcal{L}_i = \begin{cases} P(b_{1i}=1|T_i=t_i)\,\lambda_2 & \text{if } b_{1i}=1, m_i=1 \\ P(b_{1i}=1|T_i=t_i)(1-\lambda_2)P(b_{2i}=1|b_{1i}=1,T_i=t_i) & \text{if } b_{1i}=1, m_i=0, b_{2i}=1 \\ P(b_{1i}=1|T_i=t_i)(1-\lambda_2)[1-P(b_{2i}=1|b_{1i}=1,T_i=t_i)] & \text{if } b_{1i}=1, m_i=0, b_{2i}=0 \\ [1-P(b_{1i}=1|T_i=t_i)]\,\lambda_1 & \text{if } b_{1i}=0, m_i=1 \\ [1-P(b_{1i}=1|T_i=t_i)]\,(1-\lambda_1)\,P(b_{2i}=1|b_{1i}=0,T_i=t_i) & \text{if } b_{1i}=0, m_i=0, b_{2i}=1 \\ [1-P(b_{1i}=1|T_i=t_i)]\,(1-\lambda_1)\,[1-P(b_{2i}=1|b_{1i}=0,T_i=t_i)] & \text{if } b_{1i}=0, m_i=0, b_{2i}=0 \end{cases} \quad (3.1)$$

We use two different logistic regression models $\phi_1(t_i|\boldsymbol{\alpha})$ and $\phi_2(t_i|\boldsymbol{\beta})$ for $P(b_{1i}=1|T_i=t_i)$ and $P(b_{2i}=1|b_{1i}=1,T_i=t_i)$ respectively, where $\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4)$ and $\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)$, are the regression parameters:

$$\begin{aligned}\phi_1(t_i|\alpha) &= P(b_{1i}=1|T_i=t_i) \\ &= logit^{-1}\left(\alpha_0+\alpha_1 t_i+\alpha_2 t_i^2+\alpha_3 t_i^3+\alpha_4\left(I_{t_i>2}(t_i-2)^3\right)\right)\end{aligned}$$

$$\begin{aligned}\phi_2(t_i|\beta) &= P(b_{2i}=1|b_{1i}=1,T_i=t_i) \\ &= logit^{-1}\left(\beta_0+\beta_1 t_i+\beta_2 t_i^2+\beta_3 t_i^3+\beta_4\left(I_{t_i>2}(t_i-2)^3\right)\right)\end{aligned}$$

We further assume that $P(b_{2i}=1|b_{1i}=0,T_i=t_i)$, λ_1, and λ_2 do not involve the parameters α or β. Then it follows, after rearranging and collecting terms, that the likelihood contribution, up to a proportionality constant $K(\lambda_1,\lambda_2,m_i,b_{1i},b_{2i})$, is the product of two components:

$$\mathcal{L}_i \propto \left[\phi_1(t_i|\alpha)^{b_{1i}}\left(1-\phi_1(t_i|\alpha)\right)^{1-b_{1i}}\right]\cdot\left[\phi_2(t_i|\beta)^{b_{2i}}\left(1-\phi_2(t_i|\beta)\right)^{1-b_{2i}}\right]^{(1-m_i)b_{1i}} \quad (3.2)$$

Thus, the first component of the likelihood corresponds to the likelihood for the binary

outcome b_{1i}. The second component of the likelihood corresponds to the binary outcome b_{2i} conditional on $b_{1i} = 1$ among biological samples where b_{2i} is not missing, that is where $m_i = 0$. Thus, in practice, to find the maximum likelihood estimates of α and β we (1) fit a logistic regression model for the outcome b_1 using all the data, and (2) fit a separate logistic regression model for the outcome b_2 conditional on $b_1 = 1$, using the data where b_2 is observed and $b_1 = 1$. Then, we estimate $P(b_1 = b_2 = 1|T = t)$ by $\hat{\phi}(t) = \hat{\phi}_1(t|\alpha)\hat{\phi}_2(t|\beta)$. Finally, the mean window period is estimated by $\hat{\mu} = \int_{t=0}^{\infty} \hat{\phi}(t)dt$ provided $\hat{\phi}(t)$ converges.

The intuition for this approach is that we have a complete dataset for estimating $\phi_1(t|\alpha)$, since this model does not involve b_2. We can also estimate $\phi_2(t|\beta)$ without bias, using the samples where b_2 is observed (and $b_1 = 1$), assuming that conditional on $b_1 = 1$, the probability that b_2 is missing is a constant, i.e. $P(m_i = 1|b_{1i} = 1) = \lambda_2$. Therefore, the resulting estimator $\hat{\mu}$ will be consistent in scenarios where WP1 and WP2 produce biased estimates. Approach WP3 does not require the estimation of parameters for λ_1 or λ_2; it only requires the assumption that these parameters are constant with respect to b_2.

3.1.3 Simulation study of window period estimation with missing biomarkers

We assessed the abilities of window period methods WP1-WP3 to estimate μ using the following simulation framework. We first selected a dataset which had complete biomarker data for LAg-Avidity and viral load (Konikoff et al. 2013). This dataset contained 1780 observations from 709 participants of 3 longitudinal cohort studies in the US; participants contributed between 1 and 16 observations, at estimated times since seroconversion ranging between 1 month and >8 years. We then generated 1000 simulated datasets, sampling observations with replacement from the original dataset, clustering by subject ID

and stratifying by study cohort, so that each simulated dataset had the same number of sampled subjects in each cohort as the original dataset had. We assigned subjects a new ID number each time they were sampled. We then imputed the infection duration T_i for each ID number by a draw from a uniform distribution over the interval $[L_i, R_i]$, where L_i and R_i are the left and right bounds for the duration of infection corresponding to that biological sample (Brookmeyer, Konikoff, et al. 2013).

We then performed the following analysis on each simulated dataset. First, we produced recency classifications ($Y_i = +1$ or $Y_i = -1$) using the MAA consisting of LAg-Avidity < 1.5 and viral load > 1000. We used these recency classifications, along with the imputed times since seroconversion, to produce estimates of the mean window period ($\hat{\mu}_{j,T}$), using the procedure for completely observed data described in Section 2.2.3. Thus, in the simulation study, the estimated mean window period of each simulated dataset is used as the "gold standard." We considered these estimates to be the "gold standards" for each simulated dataset because they were the estimates which would be achieved with completely observed data, and which methods WP1-3 should attempt to reproduce.

We randomly assigned the biomarker B_2 to be missing by simulating Bernoulli trials for each serum sample in the dataset. The probability that B_2 was assigned to be missing depended solely on whether B_1's value met its criterion for MAA positivity, i.e., whether $b_1 = 1$. We set $P(m_i = 1|b_{1i} = 0) = \lambda_1$ and $P(m_i = 1|b_{1i} = 1) = \lambda_2$, and varied the values of λ_1 and λ_2 across six different simulation scenarios (Table 3.1).

After determining which observations were missing, we recalculated the MAA classifications Y_i, updating the value to $Y_i = 0$ when $b_{1i} = 1$ and $m_i = 1$, and then applied

methods WP1, WP2, and WP3 to the resulting incomplete datasets, producing estimates $\hat{\mu}_{j,k}$ for the j^{th} simulated dataset using the k^{th} method (where $j = 1, \ldots, 1000$ and $k = 1, 2, 3$). We computed error scores $e_{j,k} = \hat{\mu}_{j,k} - \hat{\mu}_{j,T}$ comparing each gold standard estimate with the corresponding incomplete-data estimates. We defined the bias of estimator $\hat{\mu}_k$ as the expectation of the error score for the k^{th} method, $\mathrm{E}[e_{j,k}]$, and we estimated this bias using the sample mean $\bar{e}_k = \frac{1}{n}\sum_{j=1}^{n} e_{j,k}$. Similarly, we defined the root mean squared error (RMSE) as $\sqrt{E\left[\left(e_{j,k}\right)^2\right]}$, which we estimated by $\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(e_{j,k}\right)^2}$. We also calculated sample standard deviations of the error scores, $s_k = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{n}\left(e_{j,k} - \bar{e}_k\right)^2}$, as well as the sample standard error of each estimator $\hat{\mu}_k$, $SE(\hat{\mu}_k) = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{n}\left(\hat{\mu}_{j,k} - \frac{1}{n}\sum_{j=1}^{n}\hat{\mu}_{j,k}\right)^2}$.

3.1.4 Simulation results

The simulation results are shown in Table 3.1. As λ_2 increased across the simulation scenarios, Method WP1 produced increasingly negatively biased estimates of μ, and correspondingly positively biased incidence estimates. The value of $\lambda_1 = P(m_i = 1|b_{1i} = 0)$ did not affect this method's results, because all of the samples with $b_{1i} = 0$ could be classified as $W_i = -1$, regardless of whether B_2 was missing; hence, rows 3 and 9 of Table 3.1 are identical. (In contrast, note that rows 15 and 18 are only identical due to rounding; there were differences between these two simulations for WP3 at the 0.1 level of precision.)

Method WP2 had substantial positive bias for μ in scenarios with $\lambda_2 < \lambda_1$. In scenarios where λ_1 approximately equaled λ_2, this method had less bias. Method WP3 had relatively small biases across all the scenarios studied, but did have a consistent, modestly negative

bias for μ.

Relative to the naïve methods, the conditional likelihood approach reduced bias in the mean window period estimates by over 80% in several scenarios considered; for example, in the scenario with $\lambda_1 = 0.5$ and $\lambda_2 = 0$, method WP3 had 89% smaller magnitude of bias than WP2. The standard errors were similar between the methods in most scenarios; in the cases with $\lambda_1 = 0.5$, WP2 had larger standard errors than WP1 and WP3. RMSE was comparable among the methods in some scenarios, but substantially larger for WP1 and WP2 than for WP3 in the scenarios in which these methods had large biases.

Table 3.1: Performance of 3 methods for handling missing biomarkers for estimation of the Mean Window Period

λ_1	λ_2	Method	Estimated bias (days)	$SD(error)$ (days)	$SE(\hat{\mu}_k)$ (days)	RMSE($\hat{\mu}_k$) (days)
0.25	0.25	MAA determined (WP1)	-24	8	17	25
		Complete biomarker (WP2)	0	9	20	9
		Conditional likelihood (WP3)	-8	8	18	12
0.50	0.50	MAA determined (WP1)	-53	12	16	54
		Complete biomarker (WP2)	-1	17	25	17
		Conditional likelihood (WP3)	-9	13	20	15
0.50	0.25	MAA determined (WP1)	-24	8	17	25
		Complete biomarker (WP2)	40	15	27	43
		Conditional likelihood (WP3)	-8	8	18	12
0.50	0	MAA determined (WP1)	0	0	18	0
		Complete biomarker (WP2)	74	13	28	75
		Conditional likelihood (WP3)	-8	5	17	9
0.20	0.10	MAA determined (WP1)	-9	5	17	10
		Complete biomarker (WP2)	11	6	20	13
		Conditional likelihood (WP3)	-8	6	17	10
0.10	0.05	MAA determined (WP1)	-4	3	18	6
		Complete biomarker (WP2)	5	4	19	6
		Conditional likelihood (WP3)	-8	6	17	10

The simulation results indicate that Method WP1 produced biased estimates of μ in the scenarios in which a substantial fraction of the samples could not be classified, even in the

idealized scenarios in which the probability that B_2 was missing did not depend on B_1 (i.e., $\lambda_1 = \lambda_2$). Method WP2 was approximately unbiased only if $\lambda_1 \approx \lambda_2$. The bias was small with Method WP3 for all scenarios considered.

3.1.5 Application to HIV Clade C dataset

In this section, we apply each of our calibration analyses (WP1, WP2, WP3) to the MAA "LAg avidity < 1.5, Viral load > 1000" using the Clade C calibration data set described in Section 2.4. In this data set, all 2442 samples were assayed for the LAg avidity assay but only 1657 were assayed for viral load; in particular, 97% of samples with LAg avidity < 1.5 OD-n were assayed for viral load, but only 62% of samples with LAg avidity > 1.5 OD-n were assayed for viral load (Table 3.2). We estimated the mean window period with each missing-data calibration analysis approach using this data set, and we then applied the resulting $\hat{\mu}$ estimates to the cross-sectional data from the HPTN-068 study described in Chapter 2.4.4. All of the observations in this data set were classifiable for this MAA; there were 12 MAA-positive observations ($Y_i = 1$) and 79 MAA-negatives ($Y_i = 0$). The resulting incidence estimates were compared with the longitudinal incidence estimate from the HPTN-068 study (1.9%) to compute relative error, treating the longitudinal estimate as a gold standard.

Table 3.2: Number of samples assayed for viral load in Clade C data set, by LAg assay result

	Number of samples with viral load assayed	Number of samples with viral load not assayed (missing)
LAg ≥ 1.5	1271 (62%)	775 (38%)
LAg < 1.5	386 (97%)	10 (3%)
Total	1657 (68%)	785 (32%)

In this example analysis, methods WP1 and WP3 produced similar results, while WP2 produced substantially different results (Table 3.3). Very few of the observations in the

calibration data set were unclassifiable for the MAA being considered (i.e., only 10 observations had LAg < 1.5 and viral load missing); hence WP1 and WP3 produced similar results, as in the corresponding simulation scenarios above. There was a substantial discrepancy in viral load missingness rates between LAg ≥ 1.5 and LAg < 1.5, and WP2 produced a substantially different $\hat{\mu}$ from WP1 and WP3, as in the simulations. Interestingly, WP2 produced the estimate with the smallest discrepancy from the longitudinal estimate (1.9%); however, the confidence intervals for the cross-sectional estimates all cover the longitudinal point estimate, and likewise the longitudinal confidence interval (1.3%, 2.7%) covers all of the cross-sectional point estimates.

The preceding analysis does not allow us to assess accuracy, for two reasons. First, there is no guarantee that the longitudinal estimate represents the true incidence rate, since it is derived from a cohort which was sampled stochastically from a larger population. Second, this comparison constitutes a single observation from the distribution of errors for each method; we cannot draw reliable inferences about that distribution of errors from a single observation.

Table 3.3: Mean window period and incidence estimates using the MAA "LAg avidity < 1.5, Viral load > 1000", calibrated using the HC-HIV data set and estimated using the HPTN-068 data set, by calibration set missing data analysis method

Analysis Method	Estimated mean window period $\hat{\mu}$ (95% CI)	Cross-sectionally estimated incidence in HPTN-068	Relative error vs. longitudinal estimate
WP1 (all classifiable samples)	132 (112, 155)	2.6% (1.3, 4.7)	38%
WP2 (all completely assayed samples)	179 (144, 219)	1.9% (1.0, 3.5)	2%
WP3 (two-step procedure)	134 (113, 160)	2.6% (1.3, 4.6)	35%

3.2 Cross-sectional incidence estimation with missing biomarkers

The previous section demonstrated that if missing biomarkers are not handled

appropriately, estimates of the mean window period can be biased. In this section, we consider incidence estimation using an MAA applied to a cross-sectional sample with missing biomarker values. In this section, it is assumed that the mean window period has already been estimated accurately.

As in Section 3.1, we consider the situation in which an MAA uses two biomarkers, B_1 and B_2, and only B_2 has missing values. If both biomarkers were observed for all persons, the incidence estimate would be $\hat{\iota} = V/(\mu N_u)$, where $V = \sum_{i=1}^{N_x} 1_{(Y_i=1)}$ is the number of persons in the cross-sectional survey who are classified as MAA positive, $1_{(Y_i=1)}$ is an indicator function with value 1 if $Y_i = 1$ and 0 otherwise, N_u is the number of seronegative persons in the survey, N_x is the number of seropositive persons in the survey, and μ is the (previously estimated) mean window period. We evaluate several methods for estimating the incidence rate under the same class of missingness mechanisms as in Section 3.1.

The first cross-sectional incidence approach (Method CS1), which we call the "MAA determined analysis", uses all samples whose MAA status can be determined (i.e. for which $Y_i = \pm 1$, the red and blue squares in Figure 3.1) and ignores those samples whose MAA status is indeterminate ($Y_i = 0$, green square in Figure 3.1). Method CS1 uses this subset to estimate the proportion of seropositive samples that would have been classified as MAA positive if there were no missing values, using the estimator $\hat{p}_1 = V/n_1$ where $n_1 = \sum_{i=1}^{N_x} |Y_i|$ is the number of seropositive samples determined as MAA positive or MAA negative. Then the incidence is estimated by:

$$\hat{\iota}_1 = \frac{\hat{p}_1 N_x}{\mu N_u} = \left(\frac{V}{\mu N_u}\right)\left(\frac{N_x}{\sum_{i=1}^{N_x} |Y_i|}\right) \tag{3.3}$$

As before, Method CS1 appears straightforward, because it uses all the samples for which the MAA classification can be determined. However, it turns out to be a naïve analysis that can lead to severely biased estimates if $\lambda_2 > 0$, and we quantify the bias by simulation in Section 3.2.1.

The second approach (Method CS2) which we call the complete biomarker analysis, uses all samples where both biomarkers have been measured (blue squares in Figure 3.1). Method CS2 estimates $P(b_{1i} = b_{2i} = 1)$ by $\hat{p}_2 = V/n_2$, where $n_2 = \sum_{i=1}^{N_x}(1 - m_i)$ is the number of samples with no missing biomarker values. Then, the incidence rate estimate is:

$$\hat{\iota}_2 = \frac{\hat{p}_2 N_x}{\mu N_u} = \left(\frac{V}{\mu N_u}\right)\left(\frac{N_x}{\sum_{i=1}^{N_x}(1 - m_i)}\right) \tag{3.4}$$

This method is also biased whenever $\lambda_1 \neq \lambda_2$, as we will show in Section 3.2.1.

The third approach (Method CS3), which we call the conditional likelihood analysis, uses all of the biological samples in Figure 3.1 (blue, red, and green squares). The idea of the approach is to estimate $P(b_{1i} = b_{2i} = 1)$ as a product of $P(b_{1i} = 1)$ and the conditional probability

$P(b_{2i} = 1 \mid b_{1i} = 1)$. First, we estimate $P(b_{1i} = 1)$ by $\hat{p}_{3a} = V_1/N_x$, where $V_1 = \sum_{i=1}^{N_x} b_{1i}$ is the number of samples with biomarker 1 indicating recent infection. Second, we estimate $P(b_{2i} = 1 \mid b_{1i} = 1)$ by $\hat{p}_{3b} = V/n_3$, where $n_3 = \sum_{i=1}^{N_x} b_{1i}(1 - m_i)$ is the number of seropositive samples where $b_1 = 1$ and the second biomarker B_2 is observed. Then the estimator of $P(b_{1i} = b_{2i} = 1)$ is $\hat{p}_3 = \hat{p}_{3a}\hat{p}_{3b}$. It follows that the estimator of incidence is:

$$\hat{\iota}_3 = \frac{\hat{p}_3 N_x}{\mu N_u} = \left(\frac{V}{\mu N_u}\right)\left(\frac{\sum_{i=1}^{N_x} b_{1i}}{\sum_{i=1}^{N_x} b_{1i}(1 - m_i)}\right) \tag{3.5}$$

3.2.1 Simulation study

We assessed the abilities of the three methods to estimate incidence using the following simulation framework. We used the same source dataset of N_x seropositive samples as in the previous simulation. We then generated 1000 simulated datasets, by sampling N_x observations from this original dataset, with replacement for each simulation.

We then performed the following analysis on each simulated dataset j. First, we produced recency classifications ($Y_i = +1$ or $Y_i = -1$) using the same MAA as in the previous section. We determined the (true) count of MAA-positive samples, denoted $V_{j,T} = \sum_{i=1}^{N_x} b_{1i} b_{2i}$.

As above, we introduced missing values in B_2 by generating a Bernoulli trial for each serum sample in the dataset. We considered the same six missingness scenarios as in Table 1.

After determining which observations were missing, we recalculated the MAA classifications Y_i, updating the value to $Y_i = 0$ when $b_{1i} = 1$ and $m_i = 1$, and then we applied the three methods to the resulting incomplete dataset to produce estimates $\hat{p}_{j,k}$ for the j^{th} simulated dataset using the k^{th} method (where $j = 1, \ldots, 1000$ and $k = 1, 2, 3$).

We computed relative error scores $r_{j,k} = (\hat{\iota}_{j,k} - \hat{\iota}_{j,T})/\hat{\iota}_{j,T}$ for the j^{th} simulation using the k^{th} method using the fact that $\hat{\iota}_{j,k} = \hat{p}_{j,k} N_x/(\mu N_u)$ and $\hat{\iota}_{j,T} = V_{j,T}/(\mu N_u)$ share the term $1/(\mu N_u)$; thus the relative error reduces to $r_{j,k} = \left(\hat{p}_{j,k} N_x - V_{j,T}\right)/V_{j,T}$. We defined the relative bias of estimator ι_k as the expectation of the relative error score for the k^{th} method, $\mathrm{E}[r_{j,k}]$, and we estimated this bias using the sample mean $\bar{r}_k = \frac{1}{n}\sum_{j=1}^{n} r_{j,k}$. We also calculated the sample standard deviations of these error scores, $s_k = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{n}\left(r_{j,k} - \frac{1}{n}\sum_{j=1}^{n} r_{j,k}\right)^2}$, and the

square root of the mean squared relative errors (RMSRE), $\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(r_{j,k}\right)^2}$.

3.2.2 Simulation results

We found that as λ_2 increased across the simulation scenarios, Method CS1 produced increasingly negatively biased incidence estimates (Table 3.4). The value of λ_1 did not affect this method's results, because all the samples with $b_1 = 0$ can be classified as $W = -1$, regardless of whether B_2 is missing. Method CS2 estimated incidence with substantial positive bias in scenarios with $\lambda_2 < \lambda_1$. In scenarios where $\lambda_1 \approx \lambda_2$, this method showed less bias. Method CS3 had minimal biases across all the scenarios simulated.

Table 3.4: Performance of 3 methods for handling missing biomarker data for estimation of incidence from cross-sectional surveys

λ_1	λ_2	Method	Estimated bias (mean relative error, %)	SD(relative error) (%)	RMSRE (%)
0.25	0.25	CS1	-22	5	23
		CS2	0.4	6	6
		CS3	0.1	5	5
0.50	0.50	CS1	-47	6	47
		CS2	0.007	11	11
		CS3	-0.3	9	9
0.50	0.25	CS1	-22	5	23
		CS2	42	9	43
		CS3	0.1	5	5
0.50	0	CS1	0	0	0
		CS2	78	4	78
		CS3	0	0	0
0.20	0.10	CS1	-9	3	9
		CS2	11	4	12
		CS3	-0.001	3	3
0.10	0.05	CS1	-4	2	5
		CS2	5	3	6
		CS3	-0.05	2	2

In summary, Method CS1 was substantially biased for incidence estimation in the

scenarios in which a substantial fraction of the samples could not be classified, even under the optimistic assumption that B_2's missingness probability did not depend on B_1 ($\lambda_1 = \lambda_2$). Method CS2 was approximately unbiased only if $\lambda_1 \approx \lambda_2$. Method CS3 was nearly unbiased in all scenarios considered.

The standard deviations of the relative errors were comparable among the methods, and substantially smaller than the largest biases; hence RMSRE was dominated by the contributions from bias.

3.3 Discussion

The objective of this chapter was to consider some of the statistical challenges posed by incomplete biomarker data for cross-sectional incidence estimation. We examined this problem both in the context of estimating the mean window period of an MAA and in the context of estimating incidence from a cross-sectional survey using an MAA with a previously estimated mean window period. We evaluated three methods for handling missing data in each of these contexts, simulating the methods' performance across a range of six missingness mechanism scenarios.

Our main findings were that the "MAA determined" and "complete biomarker" methods (WP1/CS1 and WP2/CS2, respectively) produced substantially biased results in a range of plausible missingness conditions. The "MAA determined" methods had bias magnitudes associated with the fraction of samples that could not be classified by the MAA, whereas the "complete biomarker" methods had more bias when the probability of missing B_2 values depended strongly on the value of B_1. In contrast, the conditional methods which we proposed (WP3 and CS3) produced accurate results in all the scenarios that we considered.

However, we do note that while the conditional WP3 approach had small bias, it was consistently negative. Further study to identify any situations when the bias would be positive would be useful.

The results in this work highlight the perhaps surprising fact that even if the probability of missing biomarkers is a constant, simply ignoring unclassifiable samples can lead to bias. This result is due to the asymmetry between the positive (recent) and negative classifications; one requires the joint occurrence (intersection) of several events, while the other requires only that at least one event occur (union). Thus, the MAA classification can only be missing if all observed biomarkers indicate a positive result. Our proposed conditional likelihood analysis accounts for this asymmetry by modeling each biomarker separately. It thus avoids the biases that affect the other two methods.

There are several important extensions to this work worth considering in future research. First, we only assessed the performances of these methods for estimating the mean window period and incidence rate. We could also consider the shadow parameter discussed in Section 2.2.4 (E. H. Kaplan and Brookmeyer 1999; Brookmeyer 2010). Since the shadow is also a function of the probability curve $\phi(t)$, we expect that "MAA determined" and "complete biomarker" analyses would produce biased estimates of the shadow as well. We expect that our proposed conditional approach would produce approximately unbiased estimates of the shadow, since it estimates $\phi(t)$ without bias.

Another complication to consider is that the probability that B_2 is missing may depend on a different dichotomization of B_1 from the one used by the MAA. For example, in the HC-HIV data set, viral load was assayed with probability ≈ 1 if LAg-Avidity was below 3.0 OD-n,

and was assayed with a smaller probability otherwise. With such a dataset, when using Method WP3 and the LAg-Avidity < 1.5, viral load > 1000 MAA to estimate μ or incidence, we could condition on $B_1 < c$, where c is any value between the missingness cutoff (3.0) and the MAA cutoff (1.5), rather than conditioning on $b_{1i} = 1$. It seems that the optimal choice of c would be the largest value, in order to maximize the number of observations used to estimate the conditional model $P(b_2 = 1 \mid B_1 < c)$; smaller cutoffs might lead to instability in the estimation of the conditional model.

Finally, more complex scenarios, including MAAs using more than two biomarkers and MAAs with missing data in more than one biomarker, may require more sophisticated analyses. We only considered missingness in viral load, which is a relatively expensive biomarker. As new generations of potentially expensive assays for detecting incident infections are developed, it can be anticipated that situations will arise in which multiple biomarkers have missing values, due to scientific or resource constraints. It will be important to further develop and refine methods for addressing these challenges.

CHAPTER 4
Transporting MAA Calibration Results

In order to be useful for calibrating the cross-sectional incidence estimate, it is usually assumed that the calibration data set comes from a population in which the relationship between duration of infection and biomarker distributions matches the relationship in the target population. This assumption is most plausible when the calibration data set is collected from a source resembling the target population, such as the same demographic area at an earlier time point. Even in such cases, there may be differences between the calibration data set and the target population due to evolution of the epidemic. (Hallett et al. 2009) Over time, the pathogen might mutate, or the relative prevalence of different strains might change. Similarly, the distribution of innate biological responses to the pathogen in the population might change. Moreover, patterns of clinical treatment might change. Any such differences could alter the relationship between infection duration and biomarker values and would need to be accounted for in order to achieve accurate calibration for the target population.

In the HIV incidence setting, discrepancies between the target population and the calibration data set could occur for several reasons. First, there has been and continues to be increasing use of anti-retroviral therapy (ART) for HIV infected persons throughout most parts of the world. (Piot and Quinn 2013) Initiation of ART therapy is occurring earlier in the course of infection. ART induces viral suppression, and viral suppression may modify some biomarker levels because antibody titers would tend to decrease, thereby making longstanding infections resemble recent infections. (Laeyendecker et al. 2015) An initial data set collected years before widespread ART use may no longer be applicable to the current

target population. Second, the subtypes of HIV that are circulating in a population may evolve over time. (Hemelaar et al. 2011) HIV subtype may affect biomarker levels and their relationship to duration of infection. (Longosz et al. 2015; Kassanjee et al. 2014) The changing mix of subtypes can create discrepancies between the initial data set and the target populations.

If $\phi(t) \stackrel{\text{def}}{=} \Pr(Y = 1|T = t)$ differs between the target population and the source of the calibration data set, then estimates of μ and ψ based on straightforwardly estimating $\phi(t)$ from the calibration data set as described in Chapter 2 will not be consistent for the target population. One simple solution is to collect a new calibration data set that is representative of the current target population for the statistical analyses. However, that approach is expensive and could take considerable time, negating the advantages of the cross-sectional approach to incidence estimation. The objective of this chapter is to develop methods that address the discrepancy between the initial training data set and the target population for cross-sectional incidence estimation.

In this chapter, we show how an initial calibration data set which is not representative of the target population because of differences in the characteristics of the epidemic can be adjusted and still utilized for calibrating methods to perform cross-sectional incidence estimation in the target population. We consider a scenario in which there is one covariate whose distribution differs between the calibration data set and the target population. In Section 4.1, we define notation and assumptions for this scenario. We then propose several approaches for addressing this discrepancy: a curve averaging approach (Section 4.2), a sample weighting approach (Section 4.3), a resampling approach (Section 4.3.10), and a

multivariate biomarker modeling approach with curve averaging and potential outcomes modeling variations (Sections 4.4 and 4.5). We construct a simulation study to evaluate the methods in Section 4.6.

4.1 Scenario and notation

A calibration data set D consists of data on biological specimens from infected individuals. On the i^{th} biological sample, we have measurements on k biomarkers, which we denote by the k-dimensional vector $\boldsymbol{B}_i$. Without loss of generality, we consider the case of a single MAA for which μ and ψ will be estimated for the target population; if multiple MAAs need to be calibrated, the following procedures can be applied separately to each MAA under consideration in the calibration process. Let Y_i denote the MAA classification of the i^{th} biological sample; for this chapter, we assume no missing values, and let $Y_i = 1$ indicate MAA positive, and $Y_i = 0$ indicate MAA negative. We have a censoring interval for duration of infection associated with the i^{th} biological sample, which we denote as $[L_i, R_i]$. In addition, we have a binary variable X whose distribution we are concerned may have changed over time. For example, in HIV applications X can indicate if persons are virally suppressed. Alternatively, X could indicate if persons are infected with a particular HIV subtype. Changes in the distribution of X create a discrepancy between the initial data set and the target population. The probability that $X = 1$ may depend on other variables which are also recorded in D. We assume each of these variables takes discrete values and that a discrete-valued variable Z defines these strata. For example, one value of Z might correspond to females who have been infected for less than 2 years. Let $P_\tau(X|Z)$ and $P_\kappa(X|Z)$ represent the probabilities in stratum Z that $X = 1$ in the target population (subscript τ) and calibration

data set (subscript κ), respectively, and let $\phi_\tau(t) = P_\tau(Y = 1|T = t)$ and $\phi_\kappa(t) = P_\kappa(Y = 1|T = t)$ represent the corresponding probabilities of a positive ("recent") MAA classification. The issues we are addressing in this chapter arise when $P_\tau(X|Z)$ in the target population and $P_\kappa(X|Z)$ in the calibration data set differ for one or more values of Z.

4.1.1 Assumptions

We make several assumptions. First, we assume that $P_\tau(X|Z)$ is known. In practice, $P_\tau(X|Z)$ could sometimes be estimated, e.g., from simple surveys of antiretroviral use in the target population. Such surveys would not need the detailed information on biomarkers as required for the initial training data set. If surveys are not available, it is still useful to perform sensitivity analyses to examine how results would change under various assumed scenarios about $P_\tau(X|Z)$ in the target population. In this article, we will examine the consequences of error in the assumed values of $P_\tau(X|Z)$ for our methods.

Second, we assume that the initial calibration data set covers the entire domain of values in the target population. This assumption is analogous to the positivity assumption in causal inference settings. (Hernán 2012) For example, if the calibration data set included only samples collected within the first six months after infection, or did not include any virally suppressed samples, but the target population did, then it would not be possible to obtain reliable estimates from either of our adjustment procedures.

Third, we assume that the distribution of biomarker values, conditional on time since seroconversion and viral suppression status, does not change between the calibration data set and the target population; that is, we assume that $p_\tau(\boldsymbol{B}|X, Z, T) = p_\kappa(\boldsymbol{B}|X, Z, T)$ for all values of X, Z, T, where T is the infection duration. This assumption implies that all the

variables that describe the relevant differences between the initial and target population are identified. It is analogous to the conditional exchangeability and consistency assumptions in causal inference settings. (Greenland and Robins 1986; Cole and Frangakis 2009)

Fourth, we assume that $P_\tau(X|Z,T) = P_\tau(X|Z)$, i.e., that $X \perp\!\!\!\perp T|Z$. This assumption implies that the stratifying variable Z completely captures the dependence of X on T.

Fifth, we assume that the distribution of stratifying variable Z, conditional on infection duration t, does not change between the calibration data set and the target population; that is, we assume that $P_\tau(Z|T) = P_\kappa(Z|T)$; this is trivially true when Z is a function of time, such as an indicator variable for duration of infection less than two years.

A sixth assumption is made for the multivariate modeling approaches in Sections 4.4 and 4.5. For those approaches we assume that the form of $p_\tau(\boldsymbol{B}|X,Z,T)$ is known, including any necessary transformations of the biomarker scales and the functional form of the linear predictor model. The parametric assumptions are not required for the other approaches.

4.2 Curve averaging approach

We want to estimate μ_τ, the mean window period for the target population, whose definition we can decompose as follows:

$$\mu_\tau \stackrel{\text{def}}{=} \int_{t\geq 0} \phi_\tau(t)dt = \int_{t\geq 0} P_\tau(Y=1|T=t)dt = \int_{t\geq 0} \sum_{z\in\mathcal{R}(Z)} \sum_{x\in\{0,1\}} P_\tau(Y=1,X=x,Z=z|T=t)\,dt$$

$$= \int_{t\geq 0} \sum_{z\in\mathcal{R}(Z)} \sum_{x\in\{0,1\}} P_\tau(Y=1|X=x,Z=z,T=t)P_\tau(X=x|Z=z,T=t)P_\tau(Z=z|T=t)\,dt$$

$$= \int_{t\geq 0} \sum_{z\in\mathcal{R}(Z)} \sum_{x\in\{0,1\}} P_\tau(Y=1|X=x,Z=z,T=t)P_\tau(X=x|Z=z)P_\tau(Z=z|T=t)\,dt$$

The third equality follows from the law of total probability, the fourth from the definition of

conditional probability, and the fifth from the assumption that the stratum variable Z captures the dependence of X on t. If the calibration data set were collected from the population of interest, we could model the marginal probability $P_\tau(Y = 1|T = t)$ directly from the data and use that model to estimate μ. However, if we only have calibration data collected from a different population, with a different nuisance covariate distribution $P_\kappa(X = x|Z = z)$, the marginal model we estimate from the calibration data set may yield biased estimates of target population parameters.

To overcome this problem, we can instead estimate the conditional model $P_\kappa(Y|X, Z, T)$. Assumption 3 entails that $P_\kappa(Y|X, Z, T) = P_\tau(Y|X, Z, T)$ since Y (the binary MAA classification) is a deterministic function of $\boldsymbol{B}$ (the vector of biomarker assay values). We have directly assumed that $P_\tau(Z|T) = P_\kappa(Z|T)$. Thus, we can derive an estimate of $\phi_\tau(t)$ from $\hat{P}_\kappa(W|X, Z, T)$, using the distribution $P_\tau(X = x|Z = z)$ corresponding to the target population:

$$\hat{\phi}_\tau(t) = \hat{P}_\tau(Y = 1|T = t) = \sum_{z \in \mathcal{R}(Z)} \sum_{x \in \mathcal{R}(X)} \hat{P}_\kappa(Y = 1|x, z, t) P_\tau(x|z) P_\kappa(z|t) \qquad (4.1)$$

Then we can use this adjusted marginal model to estimate μ_τ as usual:

$$\hat{\mu}_\tau = \int_{t \geq 0} \hat{\phi}_\tau(t) dt \qquad (4.2)$$

This procedure is analogous to the causal inference technique known as g-computation, the g-formula, or standardization. (Robins 1986; Pearl 1995, 2010; Vansteelandt and Keiding 2011; Hernán and Robins 2019).

4.3 Sample weighting approach

Another approach to this problem is to treat the calibration data set as an imbalanced sample from the target population. We can use weighted maximum likelihood techniques to correct for this imbalance.

Specifically, let $n_1(z) \equiv \sum_{i=1}^{n} 1_{\{X_i=1, Z_i=z\}}$ be the number of observations with $X = 1$ in stratum z of the calibration data set, let $n_0(z) \equiv \sum_{i=1}^{n} 1_{\{X_i=0, Z_i=z\}}$ be the number of observations with $X = 0$, and let $n(z) = n_1(z) + n_0(z)$ be the total for that stratum. For each observation $(\boldsymbol{B}, X, Z, t)$, we can construct weights

$$w_X(Z) = \frac{P_\tau(X|Z)}{P_\kappa(X|Z)} \tag{4.3}$$

where $P_\kappa(X|Z) = n_X(Z)/n(Z)$. Then if $\mathcal{L}_i$ is the marginal likelihood $P_\tau(Y_i = y_i|T = t_i)$ of the i^{th} observation, the total weighted likelihood of the data set would be:

$$\mathcal{L}_W = \prod_{i \in 1:n} (\mathcal{L}_i)^{w_{X_i}(Z_i)}$$

Equivalently, if $\ell_i = \log \mathcal{L}_i$ is the log-likelihood of the i^{th} observation, then the weighted log-likelihood of the data set would be:

$$\ell_W = \sum_{i \in 1:n} w_{X_i}(Z_i)\, \ell_i$$

We would then find $\hat{\phi}$ by maximizing this weighted log-likelihood.

These weights effectively remove the calibration data set's association between Z and X (the denominator) and replace it with the target population's association (the numerator); thus the weighted count of observations with $X = 1$ in stratum z is:

$$\sum_{i=1}^{n} w_{X_i}(Z_i)\, 1_{\{X_i=1, Z_i=z\}} = n_1(z)\, w_1(z)$$

$$= n_1(z)\, \frac{P_\tau(X=1|Z=z)}{P_\kappa(X=1|Z=z)}$$

$$= n_1(z)\, \frac{P_\tau(X=x|Z=z)}{n_1(z)/n(z)}$$

$$= P_\tau(X=x|Z=z)\, n(z)$$

which is the expected count for a simple random sample from the target population, conditional on $n(z)$.

For another perspective, consider the calibration data set as if it were a selection-biased sample from the target population. Let $S=1$ denote the event that an observation is selected and let $S=0$ denote non-selection; then we can rewrite $P_\kappa(X=x|Z=z) = P_\tau(X=x|Z=z, S=1)$. Then:

$$w_x(z) = \frac{P_\tau(X=x|Z=z)}{P_\kappa(X=x|Z=z)}$$

$$= \frac{P_\tau(X=x|Z=z)}{P_\tau(X=x|Z=z, S=1)}$$

$$= P_\tau(X=x|Z=z)\, \frac{P_\tau(S=1|Z=z)}{P_\tau(X=x, S=1|Z=z)}$$

$$= P_\tau(S=1|Z=z)\, \frac{P_\tau(X=x|Z=z)}{P_\tau(X=x, S=1|Z=z)}$$

$$= P_\tau(S=1|Z=z)\, \frac{1}{P_\tau(S=1|X=x, Z=z)}$$

$$= \frac{P_\tau(S=1|Z=z)}{P_\tau(S=1|X=x, Z=z)}$$

Within each stratum $Z=z$, $P_\tau(S=1|Z=z)$ is a constant; thus $w_x(z)$ is approximately

proportional to the inverse probability of selection given $X = x$. These weights seem analogous to those used for inverse probability weighting (IPW) in causal inference and to those used for poststratification in finite-population survey sampling. (Hernán and Robins 2019; Lumley 2010; Westreich et al. 2017; Gelman 2007)

4.3.1 Resampling approach

As an alternative to using weights as factors in the log-likelihood, we can sample with replacement $P_\tau(X = x|Z = z)n(z)$ observations from the stratum $\{X_i = x, Z_i = z\}$; that is, we resample the number of observations we would expect from a sample of the target population, conditional on $n(z)$. After resampling observations for each combination of (x, z) values, we can compile the resampled observations into a new resampled data set, $\widetilde{D}$, with distribution $P(X, Z)$ matching the target population. We can treat this resampled data set as if it came from the target population and use it to estimate $\phi_\tau(t)$ and μ.

A connection between the weighting approach and the resampling approach can be seen in the weights. In the resampling approach, $1/n_x(z)$ is the probability that a given observation in category $\{X = x, Z = z\}$ will be selected, each time that we sample an observation from that category with replacement; thus, since there will be $n(z)P_\tau(x|z)$ resampled observations in that category in each resampling data set, $w_x(z) = n(z)P_\tau(x|z)/n_x(z)$ represents the expected number of times that each observation in $\{X = x, Z = z\}$ will be selected in a resampling data set.

The notable difference between the resampling approach and the weighting approach is the stochastic nature of the resampling approach. The resampling approach creates new data sets with predetermined proportions of suppressed specimens for each time stratum, so that

the resulting data sets will have a distribution similar to the target population. Thus the (unweighted) likelihood of each resampled data set should be close to the weighted likelihood of the original calibration data set. However, because the resampling process is stochastic, it introduces an extra source of variability into the resulting estimates of ϕ and μ. This variability can be reduced by generating multiple resampled data sets and merging the results, for example by using the median of the resulting $\hat{\mu}$ distribution as the final estimate. In contrast, the weighting approach is not stochastic; like the unadjusted approach, it produces a single weighted data set and corresponding $\hat{\phi}, \hat{\mu}$, for a given calibration data set and target population.

Simulations using this resampling approach are considered in our published work (D. Morrison et al. 2019); this method is not included in the simulation study below. It produces similar results to the sample weighting approach but requires more computation due to the need to repeatedly resample the data set.

4.4 Multivariate modeling and marginalization (MMM) approach

A third strategy for addressing discrepancies between the calibration data set and target populations is similar to the curve averaging approach, in that we will fit a model that conditions on viral suppression status X; however, instead of directly modeling the distribution of MAA classifications, $P(Y|X,Z,T)$, we will instead initially model the multivariate distribution of the individual biomarker values, $p(\boldsymbol{B}|X,Z,T)$. We can numerically integrate this model to derive $P(Y|X,Z,T)$:

$$P(Y = 1|X,Z,T) = \int Y_c(\boldsymbol{b})\, p(\boldsymbol{b}|X,Z,T) d\boldsymbol{b} \qquad (4.4)$$

We can then apply equations 4.1 and 4.2 to compute μ_τ, as we did in the curve averaging approach. Here, $Y_c(\boldsymbol{b})$ denotes the MAA classification variable, Y, expressed as a function of the vector of biomarker values, $\boldsymbol{b}$, and the vector of MAA cutoffs, $\boldsymbol{c}$; i.e., $Y_c(\boldsymbol{b}) = 1\{\forall j: b_j < c_j\}$. We will refer to this approach as multivariate modeling and marginalization (MMM).

More specifically, we can use a multivariate Gaussian model for the biomarker assay values, with mean function $\mathrm{E}(\boldsymbol{B}) = f_{\boldsymbol{\alpha}}(X, Z, T)$ and multivariate normal residual errors $\boldsymbol{\epsilon} = \boldsymbol{B} - \mathrm{E}(\boldsymbol{B}) \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$; here $\boldsymbol{\alpha}$ is the vector of mean function parameters (e.g., regression coefficients) and $\boldsymbol{\Sigma}$ is the covariance matrix of the residual errors. As an example, we applied this approach to the MAA "LAg < 2.8, BioRad < 40" using the calibration data set of biomarker data for subtype C infections combining the CAPRISA, FHI, and HPTN 039 cohorts described in Chapter 2. (Laeyendecker et al. 2018) In exploratory analysis, we determined that to achieve a good model fit with approximately Gaussian errors, it was best to transform the BioRad avidity biomarker from its original scale into a logit scale and to transform infection duration t onto a logarithmic scale. We thus fit a model the following mean function:

$$\mathrm{E}[(B_1, \operatorname{logit} B_2)'] = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 \log_{10} t + \boldsymbol{\alpha}_2 x \qquad (4.5)$$

where t is the duration of infection and x is viral suppression status (1 = suppressed, 0 = not). The maximum likelihood estimates were as follows:

$$\widehat{\boldsymbol{\alpha}} = \begin{pmatrix} 3.00 & 1.13 \\ 1.54 & 2.55 \\ -1.08 & -0.68 \end{pmatrix}$$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.563 & 0.413 \\ 0.413 & 1.040 \end{pmatrix}$$

Applying equations 4.4, 4.1, and 4.2 for a target population with the viral suppression distribution $P(X = x|Z = z) = \{30\%, t \leq 1 \text{ year}; 60\%, \mathrm{t} > 1 \text{ year}\}$, we would compute $\hat{\mu} =$

143.1 days. For a population with $P(X = x|Z = z) = \{15\%, t \leq 1 \text{ year}; 30\%, \text{t} > 1 \text{ year}\}$, we would instead compute $\hat{\mu} = 121.8$ days.

4.5 Potential outcomes weighting approaches

As an alternative to the marginalization step in the MMM approach (i.e., the application of equations 4.4, 4.1, and 4.2 to derive $\hat{\mu}_\tau$), we could instead adopt a potential outcomes perspective and consider, for each observation in the calibration data set, what biomarker values we would have observed, if person had the value of X that they did not experience in reality. (Neyman 1990; Rubin 1974) If we can estimate these counterfactual biomarker values, we can use them to create counterfactual data sets that might have been observed if the calibration data were drawn from the target population.

Let $\widehat{\boldsymbol{B}}_i(x)$ be the predicted value of $\boldsymbol{B}_i$ with $X_i = x$ holding t_i and Z_i fixed at their values as given in the observed calibration data set D. We can use these predictions to impute the potential biomarker values $(\boldsymbol{B}_i(1), \boldsymbol{B}_i(0))$ as follows. First, we assume that the observed value $\boldsymbol{B}_i$ equals the potential outcome with the observed X value; that is $\boldsymbol{B_i}(X_i) = \boldsymbol{B}_i$. Second, assuming multivariate normal residual errors independent of the covariates, we can estimate the counterfactual potential outcome:

$$\widetilde{\boldsymbol{B}}_i(x) = \boldsymbol{B}_i + (x - X_i)\left(\widehat{\boldsymbol{B}}_i(1) - \widehat{\boldsymbol{B}}_i(0)\right) \tag{4.6}$$

That is,

$$\widetilde{\boldsymbol{B}}_i(1) = \boldsymbol{B}_i + \left(\widehat{\boldsymbol{B}}_i(1) - \widehat{\boldsymbol{B}}_i(0)\right) \text{ if } X_i = 0$$

$$\widetilde{\boldsymbol{B}}_i(0) = \boldsymbol{B}_i - \left(\widehat{\boldsymbol{B}}_i(1) - \widehat{\boldsymbol{B}}_i(0)\right) \text{ if } X_i = 1$$

For example, consider a simple multivariate normal linear regression model of the form

$$\mathrm{E}[\boldsymbol{B}_i] = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 t_i + \boldsymbol{\alpha}_2 x_i$$

Then, the estimated counterfactual biomarker values $\widehat{\boldsymbol{B}}_i(1 - X_i)$ are:

$$\widetilde{\boldsymbol{B}}_i(1) = \boldsymbol{B}_i + \boldsymbol{\alpha}_2 \text{ if } X_i = 0$$

$$\widetilde{\boldsymbol{B}}_i(0) = \boldsymbol{B}_i - \boldsymbol{\alpha}_2 \text{ if } X_i = 1$$

The counterfactual imputation process produces two data sets, $D(1)$ and $D(0)$, each of which is a modified copy of the observed calibration data set D with the observed $\boldsymbol{B}_\mathrm{i}$ replaced with $\widetilde{\boldsymbol{B}}_i(1)$ or $\widetilde{\boldsymbol{B}}_i(0)$, respectively.

Given these potential outcomes, we want to construct data sets matching the target population's distribution $P_\tau(X|Z)$. We can accomplish this goal in several ways.

4.5.1 Complete potential outcomes weighting (CPOW) approach

First, we can assign weights $w_i(1) = P_\tau(X = 1|Z_i)$ to $D(1)$ and $w_i(0) = P_\tau(X = 0|Z_i)$ to $D(0)$, and concatenate these data sets. Given such an augmented data set, we have two options for analysis. First, we could perform weighted maximum likelihood estimation of $\phi_\tau(t) \stackrel{\text{def}}{=} P_\tau(Y = 1|T = t)$, as in Section 4.3. We call this procedure Complete Potential Outcomes Weighting (CPOW).

4.5.2 Complete potential outcomes sampling (CPOS) approach

Second, we can stochastically create a single data set $\widetilde{D}$ by selecting one of $\widetilde{\boldsymbol{B}}_i(1)$ or $\widetilde{\boldsymbol{B}}_i(0)$, with probabilities $P_\tau(X = 1|Z_i)$ and $P_\tau(X = 0|Z_i)$ respectively, identical to the weights used in CPOW. We call this procedure Complete Potential Outcomes Sampling (CPOS). As in the resampling method in Section 4.3.1, we can repeat this procedure multiple times, creating data sets $\widetilde{D}_1, \ldots, \widetilde{D}_K$, and merge the results (for example by taking the median of the $\hat{\mu}$s) to reduce the variability introduced by stochastic sampling. This approach is considered in (D.

Morrison et al. 2019); it is not included in the simulation study below. It produces similar results to CPOW but requires more computation due to the need to repeatedly sample from the potential outcomes.

4.5.3 Partial potential outcomes weighting (PPOW) approach

Both methods above consider both potential outcomes for every observation – hence the nomenclature "Complete". An alternative is to consider the counterfactual outcome only for a subset of the observations. Specifically, we propose the following weights/probabilities. For observations with $X_i = 1$ in the calibration data set D, assign to $\widetilde{\boldsymbol{B}}_i(0)$ the weight

$$w_{10}(Z_i) = \max\left\{0, \frac{P_\tau(X = 0|Z_i) - P_\kappa(X = 0|Z_i)}{P_\kappa(X = 1|Z_i)}\right\}$$

and assign to $\widetilde{\boldsymbol{B}}_i(1) = \boldsymbol{B}_i$ the weight $1 - w_{10}(Z_i)$. Analogously, for observations with $X_i = 0$ in the observed calibration data set D, assign to $\widetilde{\boldsymbol{B}}_i(1)$ the weight

$$w_{01}(Z_i) = \max\left\{0, \frac{P_\tau(X = 1|Z_i) - P_\kappa(X = 1|Z_i)}{P_\kappa(X = 0|Z_i)}\right\}$$

and assign to $\widetilde{\boldsymbol{B}}_i(0) = \boldsymbol{B}_i$ the weight $1 - w_{01}(Z_i)$. That is, if $X_i = 1$ and $P_\tau(X = 1|Z = Z_i) \geq P_\kappa(X = 1|Z = Z_i)$, only use the observed $\boldsymbol{B}_i$ (with weight 1). Likewise, if $X_i = 0$ and $P_\tau(X = 1|Z = Z_i) \leq P_\kappa(X = 1|Z = Z_i)$, only use the observed $\boldsymbol{B}_i$. For the remaining observations, we use both the observed value and the counterfactual value, with weights determined by the discrepancy between the calibration and target populations. In other words, assign weight $v_i(1) = X_i\big(1 - w_{10}(Z_i)\big) + (1 - X_i)w_{01}(Z_i)$ to potential outcome $\widetilde{\boldsymbol{B}}_i(1)$ and weight $v_i(0) = X_i\, w_{10}(Z_i) + (1 - X_i)\big(1 - w_{01}(Z_i)\big)$ to potential outcome $\widetilde{\boldsymbol{B}}_i(0)$. Then the expected weight assigned to $\widetilde{\boldsymbol{B}}_i(1)$, conditional on Z_i, is:

$$\mathrm{E}[v_i(1)|Z_i] = w_{01}(Z_i)\, P_{\kappa}(X = 0|Z_i) + \big(1 - w_{10}(Z_i)\big)\, P_{\kappa}(X = 1|Z_i)$$

and after substituting the expressions for $w_{01}(z)$ and $w_{10}(z)$ in the above equation, it can be seen that $\mathrm{E}[v_i(1)|Z_i] = P_{\tau}(X = 1|Z_i)$, the distribution in the target population. These weights can then be used either as likelihood weights or as sampling probabilities; we will call these two options Partial Potential Outcomes Weighting (PPOW) and Partial Potential Outcomes Sampling (PPOS), respectively. Here we will only consider PPOW, which is less computationally intensive.

Note that when the target population and calibration data set have the same distributions, i.e., when $P_{\tau}(X|Z) = P_{\kappa}(X|Z)$, the weights w_{10} and w_{01} both become 0, and only the observed outcomes are used. Hence in this case, PPOW and PPOS are equivalent to the unadjusted analysis. In contrast, CPOW and CPOS do not reduce to the unadjusted analysis in this case.

4.6 Simulation study

To compare the bias and precision of the proposed methods, we performed a simulation study. We used the multivariate biomarker model that we fit in Section 4.4 to define the following data-generating process:

$$t \sim \mathrm{Uniform}(0,12)$$

$$Z = 1_{\{t \leq 1\}}$$

$$p = p_1 Z + p_2(1 - Z)$$

$$X \sim \mathrm{Bernoulli}(p)$$

$$\boldsymbol{V} = (1, \log_{10}(t)\,, X)'$$

$$\boldsymbol{\alpha} = \begin{pmatrix} 3.00 & 1.13 \\ 1.54 & 2.55 \\ -1.08 & -0.68 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.563 & 0.413 \\ 0.413 & 1.040 \end{pmatrix}$$

$$(B_1, \text{logit}(B_2))' \sim N_2((\boldsymbol{V}'\boldsymbol{\alpha})', \boldsymbol{\Sigma})$$

$$Y = 1_{\{B_1 < c_1, B_2 < c_2\}}$$

The parameters p_1 and p_2 specify $P(X|Z)$, and thus these parameters define the differences between the target population and the population from which the calibration data set is sampled. We can also consider ranges of values of these parameters for each population. Note that in this specific scenario, $P(X|Z)$ reduces to a step function on infection duration, $P(X|t) = p_1 1_{\{t \leq 1\}} + p_2 1_{\{t > 1\}}$. This simplification occurs because we defined $Z|t = 1_{\{t \leq 1\}}$. In other scenarios, Z might have a more complex definition; for example, Z might represent combinations of gender and a discretization of infection duration. The parameters c_1, c_2 specify the MAA; for the following analyses, they are held fixed at $c_1 = 2.8, c_2 = 40$, which are the values determined as optimal for this pair of biomarkers in a previous study. (Brookmeyer, Konikoff, et al. 2013) Given this generating model, the true target estimand μ_τ can be calculated numerically using the marginalization method described in Section 4.4.

For each of several scenarios for $P_{\text{K}}(X|Z)$ – that is, for each of several pairs of (p_1, p_2) values – and for each of several sample sizes, we generated 2000 simulated calibration data sets. For each simulated data set, we applied each of the proposed analyses, as well as a naïve unadjusted analysis treating the calibration data as a representative sample of the target population, and we collected the resulting estimates of μ. We compared these estimates to the true value computed as described above, to estimate the bias, standard error, and mean

squared error of each approach.

In order to implement these analyses, functional forms needed to be chosen for the regression models. In our main simulations, we implemented the potential outcomes modeling approaches using the correct functional form for the multivariate biomarker model that matched the data-generating process – i.e., the infection duration t enters the model on the logarithmic scale and biomarker B_2 is Gaussian on the logit scale. The unadjusted analysis, curve averaging approach, sample weighting approach, and potential outcomes weighting approaches were all implemented with infection duration t entering the models $P(Y = 1|T = t)$ and $P(Y = 1|x, z, t)$ on a logarithmic scale. It is unclear what the "correct" functional form should be for these two models; given the data-generating model that we have assumed for this simulation, these functions are transformations of the underlying biomarker model which do not seem to have simple algebraic expressions:

$$P(Y = 1|x, z, t) = \int Y_c(\boldsymbol{b})\, p(\boldsymbol{b}|x, z, t)\, d\boldsymbol{b}$$

$$P(Y = 1|T = t) = \sum_{z \in \mathcal{R}(Z)} \sum_{x \in \mathcal{R}(X)} P(Y = 1|x, z, t)\, P(x|z)\, P(z|t)$$

As a sensitivity analysis, we also simulated all of the approaches with linear functional forms for T, and we simulated the multivariate modeling approaches (MMM, CPOW, and PPOW) with B_2 on its original, untransformed scale.

4.7 Results

Table 4.1 shows simulation results comparing the performances of the unadjusted analysis, curve averaging approach, sample weighting approach, and multivariate modeling and marginalization (MMM) approach for estimating the mean window period in the target

population, under several data-generating scenarios with varying assumptions about $P_\kappa(X|Z)$. In scenarios A and B, the calibration data set had lower levels of viral suppression than the target population, and the unadjusted analysis produced estimates with biases of approximately 22 days and 12 days, respectively. In contrast, all three adjustment approaches produced estimates with biases of less than 3 days in these scenarios. In scenario C, the calibration data set had the same levels of viral suppression as the target population, and all of the methods, including the unadjusted analysis, produced estimates with minimal bias.

The MMM approach produced estimates with standard errors that substantially smaller than those of the other analyses in all three scenarios. The curve averaging and sample weighting approaches resulted in substantially larger standard errors than the unadjusted analysis in scenarios A and B; in scenario C, these two approaches had standard errors on par with the unadjusted analysis.

Because it had minimal biases and the smallest standard errors, the MMM approach also produced the smallest RMSEs in all three scenarios, for every sample size considered. In contrast, the curve averaging and sample weighting approaches had larger RMSEs than the unadjusted approach in scenarios A and B for sample sizes of 250 and were on par with the unadjusted approach in RMSE for sample sizes of 500; the reductions in bias for these methods relative to the unadjusted approach only outweighed the increases in variance for the larger sample sizes. In scenario C, these approaches had RMSEs comparable to the unadjusted analysis for all sample sizes considered.

Table 4.2 compares the MMM approach with the CPOW and PPOW approaches. The MMM

approach had substantially better precision than CPOW and PPOW, but these variants still produced unbiased estimates, with SEs and RMSEs that were as small or smaller than those produced by the unadjusted analysis, the curve averaging approach, and the sample weighting approach.

Table 4.3 shows simulation results evaluating the robustness of MMM to modeling misspecifications. Bias increased substantially when infection duration was modeled on a linear scale instead of a logarithmic scale, and bias was even worse when B_2 was not correctly transformed to the logit scale. When both modeling errors were present, the average estimated mean window period shrunk to nearly zero, creating a massive bias; the standard errors shrunk in this case, but only because the estimated mean window periods were consistently close to 0.

Table 4.4 shows simulation results evaluating the robustness of CPOW to the same modeling misspecifications. Bias increased appreciably when B_2 was not correctly transformed to the logit scale, especially for scenario A, but not by nearly as much as for MMM. The bias also increased somewhat when infection duration was not correctly transformed to a logarithmic scale. When both modeling errors were present, the bias increased slightly further than when only B_2 was incorrectly transformed. Standard error was not substantially affected by these misspecifications. Even with both modeling errors, MMM had biases comparable to the unadjusted analysis in scenarios A and B, and RMSEs comparable to the curve averaging and sample reweighting approaches for sample sizes of 250 and 500 in all three scenarios. The bias introduced by misspecification only had substantial impact on RMSE for larger sample sizes.

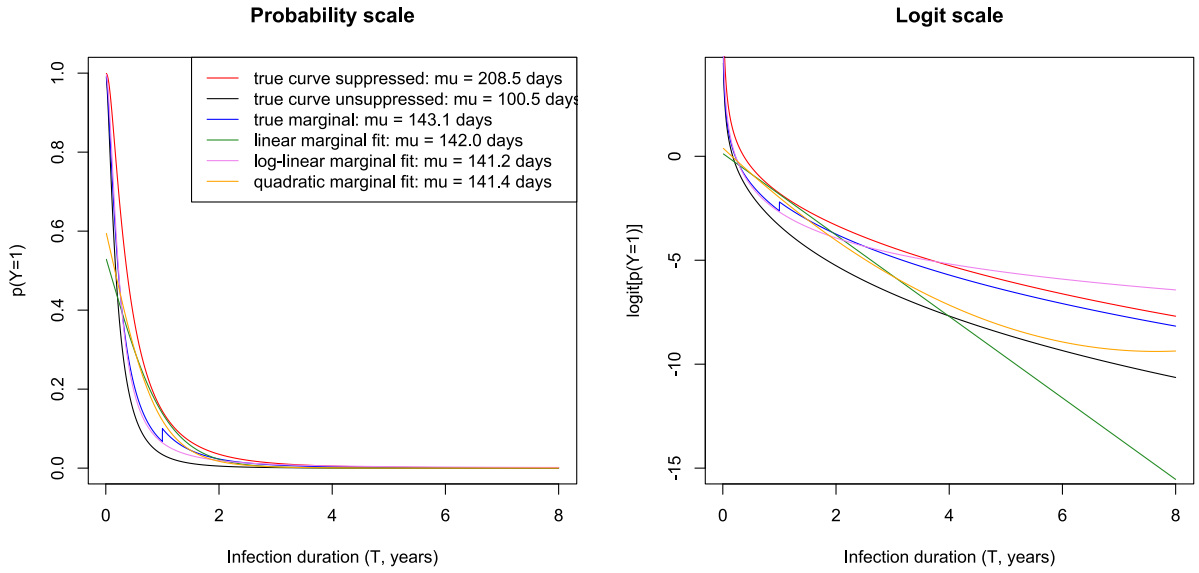
Table 4.5 shows simulation results evaluating the robustness of PPOW to the same misspecifications. The biases found for CPOW in Scenarios A and B are still present, but they are slightly smaller. However, the standard errors for PPOW are still larger than those for CPOW, so that the RMSE is larger for PPOW than CPOW for sample sizes of 250 or 500. Note that in Scenario C, PPOW is unbiased even when misspecified, because it reduces to the unadjusted analysis in this case, as discussed previously.

Table 4.6 shows simulation results evaluating the results of a linear functional form for infection duration, T, in the unadjusted analysis, curve averaging approach, and sample weighting approach. As discussed in Section 4.6, it is unclear what the ideal functional form should be for the model $P(Y = 1|x, z, t)$ used in the curve averaging approach or for the model $P(Y = 1|T = t)$ used in the unadjusted analysis, sample weighting approach, and potential outcomes weighting approaches. The results with a linear form are virtually identical to the corresponding ones in Table 4.1, indicating that these methods are relatively insensitive to functional form specification under the assumed data-generating model.

To further explore this issue, we graphed the three curves $P(Y = 1|X = 1, T = t)$, $P(Y = 1|X = 0, T = t)$, and $P(Y = 1|T = t)$, using the biomarker model parameter values listed above and the values $p_1 = 0.3, p_2 = 0.6$ for the target population's viral suppression distribution (Figure 4.1). We also included three additional curves, representing population-level marginal logistic regression models for $P(Y = 1|T)$. We estimated each of them using 10^5 simulated observations from the target distribution. The first model is linear in t, i.e. $\text{logit}(P(Y = 1|T = t)) = \alpha_0 + \alpha_1 t$. The second model is quadratic in t, i.e. $\text{logit}(P(Y = 1|T = t)) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$. The third model is log-linear in t, i.e.,

$\text{logit}(P(Y = 1|T = t)) = \alpha_0 + \alpha_1 \log_{10}(t)$. We graphed all curves on both the response (probability) and logit scales and provided the areas under the curve (integrated numerically from 0 to 10 years) in the legend. From the graphs and corresponding μ values, we can see that linear and quadratic models (green and orange) are both adequate to accurately estimate μ, even though they are not particularly good approximations to the true marginal curve (blue) on either scale.

Figure 4.1: MAA characteristics for transportability simulation model, on probability scale (left) and logit scale (right)



4.8 Discussion

Infectious disease prevention and evaluation research rely on accurate incidence estimates. Methodological challenges arise because of the difficulty in determining incidence from longitudinal cohorts of initially-uninfected persons and documenting infection acquisition. (Lagakos and Gable 2008) The cross-sectional approach addresses these challenges because incidence can be estimated without requiring longitudinal follow-up of persons. However, the cross-sectional approach does rely on an initial training data set to develop and calibrate the statistical methods to be used in cross-sectional surveys. The problem addressed in this

chapter is that the calibration data set may over time not reflect the current target population. We developed methods to adjust the analysis of the calibration data set in order to achieve unbiased estimation of the mean window period for the target population. These adjustment procedures could help avoid the time and expense of collecting a completely new training data set for each new target population. A critical assumption of the methods is that the variables that describe the relevant differences between the calibration and target population are identified. In our application the relevant variable was anti-retroviral treatment which results in viral suppression.

We proposed a variety of approaches for adjusting the calibration analysis: a curve averaging approach, a sample reweighting approach, and a multivariate modeling approach with several variations. We found that each of these methods produced estimates with negligible bias, as long as their underlying assumptions held true, whereas an unadjusted analysis produced estimates with substantial bias when the calibration data set's viral suppression distribution differed from the target population.

The adjustment methods' performances differed in precision: the multivariate modeling and marginalization (MMM) approach produced the smallest standard errors. This approach requires additional parametric assumptions not shared by the curve averaging and reweighting approaches; these assumptions resulted in lower standard errors, at the cost of vulnerability to bias when those assumptions are violated. Thus, this approach requires careful model fitting procedures to be reliable.

The complete and partial potential outcomes weighting (CPOW and PPOW) approaches offer a compromise between the strong modeling assumptions required by the MMM

approach and the weaker assumptions of the curve averaging and resampling approaches. They still make use of the assumptions of the MMM approach to gain precision but trade some of that precision for less vulnerability to bias from violations of those assumptions.

Overall, these findings suggest that the curve averaging or sample weighting approaches may be preferable in situations where the correct functional form for the multivariate biomarker model is unclear. However, for calibration data sets with small sample sizes or when the functional form for the multivariate model seems clear, the MMM approach with sensible modeling assumptions may be worthwhile. The use of flexible nonparametric methods such as splines or LOESS to fit the multivariate model may also improve the reliability of this method.

The curve averaging and sample weighting approaches had similar performance characteristics to each other; the choice between them may depend on whether the conditional model $P(Y|X,T)$ or the marginal model $P(Y|T)$ is easier to fit well to a given data set. The sample weighting approach also requires that there be at least one observation in the calibration data set for every combination of X and Z values; otherwise, the denominator of the weight $w_X(Z)$ [Eq. 4.1] for that combination is 0 and the weight itself is infinite. The curve averaging approach and MMM approach could still be used in such cases, although their reliability would be questionable since their predictions for those X and Z values would be extrapolations. Ideally, several analysis approaches should be employed in parallel, and the results should be compared to check for sensitivity to the specific assumptions of each method.

We assumed that only the levels of viral suppression differed between the training data

set and the target population. In reality, there may be several such relevant covariates, such as infection subtype, calendar time, and demographic factors that differ between the training and target populations. Further, the model for the biomarkers may involve interactions among these covariates. Future work could include extending these methods to such situations and would involve assuming or estimating a joint model for the covariates as a function of time, that is, replacing $P_T(X|Z)$ with $P_T(X_1, \ldots X_p|Z)$, and using this joint model to either resample the training data set or generate counterfactually adjusted data sets.

Here, we considered an MAA that did not include viral suppression status as a biomarker. However, it is possible to apply our methods to MAAs that do include viral suppression status, even when viral suppression is also the covariate X whose distribution needs to be adjusted.

A critical assumption of our methods for transporting results from the initial training dataset to the target population is the exchangeability assumption, that is, $P_\tau(Y|X, Z, T) = P_\kappa(Y|X, Z, T)$. An epidemic could evolve to violate this assumption. For example, if the virus mutates, new strains with different biological signatures could be introduced that might invalidate the assumption. Changes in clinical practice or in population characteristics could also invalidate this assumption; for example, if the mixture of causes of viral suppression (from anti-retroviral treatment versus innate resistance) evolves over time, then the statistical relationship between suppression and biomarker values may also shift. It is important to consider the validity of the exchangeability assumption using expert knowledge and any additional data that may be available from the target population.

While we have discussed model transportation approaches in the context of cross-

sectional incidence estimation, the proposed approaches are more general. They could be applicable in other situations where complex statistical analyses are conducted using an initial data set but those results may not be directly transportable to the new target population of interest.

Cross-sectional incidence methods have been successfully applied in many settings around the world. (Coates et al. 2014; Solomon et al. 2016) These methods require training data sets to develop and calibrate the methods. The approach we have proposed could offer a practical and cost-effective way to apply cross-sectional incidence methods to new target populations as the epidemic continues to evolve.

Table 4.1: Simulation results comparing the performance of adjustment procedures for estimating the mean window period in a target population. Bias, standard error (SE), and root mean squared error (RMSE) are given in days. Infection duration T is modeled on a logarithmic scale.

	Scenario					Unadjusted analysis			Curve averaging			Sample weighting			Multivariate modeling and marginalization ("MMM")		
	$P_K(X=1\|Z)$		$P_\tau(X=1\|Z)$														
	$Z=1$	$Z=2$	$Z=1$	$Z=2$	Sample size	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
A	15%	30%	30%	60%	250	-21.9	36.0	42.1	-3.7	49.3	49.5	-2.0	48.0	48.0	-1.5	20.7	20.7
					500	-21.2	24.0	32.0	-0.9	32.3	32.3	-0.8	32.3	32.3	-0.6	14.4	14.4
					2500	-22.0	10.4	24.4	-1.2	13.6	13.6	-1.0	13.7	13.7	-0.1	6.6	6.6
					10000	-21.7	5.3	22.3	-0.8	6.9	6.9	-0.5	6.9	6.9	0.1	3.3	3.3
B	15%	60%	30%	60%	250	-12.1	37.8	39.6	-0.5	42.1	42.1	-1.2	43.2	43.2	-1.7	19.6	19.6
					500	-11.6	25.4	27.9	0.1	27.9	27.9	-0.8	29.4	29.4	-0.6	14.0	14.0
					2500	-12.1	11.1	16.4	-0.4	11.9	12.0	-1.0	12.2	12.2	-0.1	6.4	6.4
					10000	-11.8	5.7	13.1	-0.2	6.1	6.1	-0.6	6.2	6.2	0.0	3.1	3.1
C	30%	60%	30%	60%	250	-0.8	38.2	38.2	-1.3	38.5	38.5	-1.2	38.2	38.2	-1.7	19.4	19.4
					500	-0.1	26.4	26.4	-0.2	26.2	26.2	-0.3	26.1	26.1	-0.6	13.9	13.9
					2500	-1.2	11.4	11.5	-0.9	11.2	11.3	-1.1	11.2	11.3	-0.1	6.3	6.3
					10000	-0.7	5.9	5.9	-0.5	5.8	5.8	-0.7	5.8	5.8	0.0	3.1	3.1

Table 4.2: Simulation results comparing the performance of three variations of the multivariate modeling adjustment approach for estimating the mean window period in a target population. Bias, standard error (SE), and root mean squared error (RMSE) are given in days.

	Scenario					**Multivariate modeling and marginalization (MMM)**			**Complete potential outcomes weighting (CPOW)**			**Partial potential outcomes weighting (PPOW)**		
	$P_\kappa(X=1\|Z)$		$P_\tau(X=1\|Z)$											
	$Z=1$	$Z=2$	$Z=1$	$Z=2$	**Sample size**	**Bias**	**SE**	**RMSE**	**Bias**	**SE**	**RMSE**	**Bias**	**SE**	**RMSE**
A	15%	30%	30%	60%	250	-1.5	20.7	20.7	-0.8	33.0	33.0	-0.9	35.4	35.4
					500	-0.6	14.4	14.4	-0.4	22.6	22.6	-0.2	24.2	24.2
					2500	-0.1	6.6	6.6	-1.0	9.8	9.8	-0.9	10.4	10.5
					10000	0.1	3.3	3.3	-0.5	5.0	5.0	-0.6	5.3	5.4
B	15%	60%	30%	60%	250	-1.7	19.6	19.6	-0.9	32.1	32.1	-0.9	36.3	36.3
					500	-0.6	14.0	14.0	-0.8	22.0	22.1	-0.4	24.7	24.7
					2500	-0.1	6.4	6.4	-0.9	9.6	9.7	-0.9	10.8	10.8
					10000	0.0	3.1	3.1	-0.6	5.0	5.0	-0.7	5.5	5.5
C	30%	60%	30%	60%	250	-1.7	19.4	19.4	-0.8	31.9	31.9	-0.8	38.2	38.2
					500	-0.6	13.9	13.9	-0.7	22.0	22.0	-0.1	26.4	26.4
					2500	-0.1	6.3	6.3	-1.0	9.6	9.6	-1.2	11.4	11.5
					10000	0.0	3.1	3.1	-0.6	4.9	5.0	-0.7	5.9	5.9

Table 4.3: Simulation results evaluating the robustness of the multivariate modeling and marginalization (MMM) approach for estimating mean window period under various model misspecifications. Bias, standard error (SE), and root mean squared error (RMSE) are given in days.

	Scenario					Infection duration (T) scale misspecified			Biomarker B_2 scale misspecified			Both misspecified		
	$P_\kappa(X=1\|Z)$		$P_\tau(X=1\|Z)$											
	$Z=1$	$Z=2$	$Z=1$	$Z=2$	Sample size	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
A	15%	30%	30%	60%	250	-56.2	30.8	64.1	-93.2	14.7	94.3	-134.4	8.0	134.7
					500	-56.4	21.4	60.3	-93.0	10.3	93.6	-135.5	5.3	135.6
					2500	-56.8	9.8	57.6	-93.0	4.6	93.1	-136.3	2.1	136.4
					10000	-56.6	4.8	56.8	-92.9	2.3	92.9	-136.4	1.0	136.5
B	15%	60%	30%	60%	250	-58.6	28.5	65.1	-86.5	15.3	87.9	-131.9	9.1	132.3
					500	-58.6	19.8	61.8	-86.2	10.8	86.9	-132.9	6.2	133
					2500	-58.8	9.2	59.5	-86.2	4.9	86.3	-133.7	2.6	133.8
					10000	-58.8	4.5	58.9	-86.1	2.4	86.2	-133.9	1.2	133.9
C	30%	60%	30%	60%	250	-55.3	29.5	62.7	-83.9	15.5	85.3	-130.2	10.1	130.6
					500	-55.3	20.5	59.0	-83.5	11.0	84.2	-131.2	6.9	131.4
					2500	-55.6	9.5	56.4	-83.5	4.9	83.7	-132.2	2.9	132.2
					10000	-55.5	4.7	55.7	-83.4	2.5	83.5	-132.3	1.4	132.3

Table 4.4: Simulation results evaluating the robustness of the complete potential outcomes weighting (CPOW) approach for estimating mean window period under various model misspecifications. Bias, standard error (SE), and root mean squared error (RMSE) are given in days.

	Scenario					Infection duration scale misspecified			Biomarker B_2 scale misspecified			Both misspecified		
	$P_\kappa(X=1\|Z)$		$P_\tau(X=1\|Z)$											
	$Z=1$	$Z=2$	$Z=1$	$Z=2$	Sample size	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
A	15%	30%	30%	60%	250	-4.6	34.7	35.0	-17.6	33.0	37.4	-18.9	34.5	39.3
					500	-4.2	24.1	24.4	-17.0	22.3	28.0	-18.2	23.9	30.0
					2500	-4.3	10.5	11.4	-17.5	9.7	20.0	-18.3	10.5	21.1
					10000	-3.6	5.4	6.5	-17.0	5.0	17.8	-17.7	5.4	18.5
B	15%	60%	30%	60%	250	-6.8	34.3	35.0	-12.2	34.2	36.3	-13.1	36.9	39.1
					500	-6.4	23.8	24.7	-11.7	23.1	25.9	-12.6	25.2	28.2
					2500	-6.1	10.6	12.2	-12.2	10.2	15.8	-12.7	11.2	16.9
					10000	-5.7	5.4	7.9	-11.7	5.2	12.8	-12.0	5.8	13.3
C	30%	60%	30%	60%	250	-3.9	34.2	34.4	-7.3	34.3	35.1	-7.2	36.8	37.5
					500	-3.2	23.9	24.1	-6.6	23.4	24.3	-6.2	25.4	26.2
					2500	-2.9	10.5	10.9	-7.3	10.2	12.6	-6.5	11.2	12.9
					10000	-2.4	5.5	6.0	-6.7	5.3	8.6	-5.7	5.8	8.2

Table 4.5: Simulation results evaluating the robustness of the partial potential outcomes weighting (PPOW) approach for estimating mean window period under various model misspecifications. Bias, standard error (SE), and root mean squared error (RMSE) are given in days.

	Scenario					Infection duration scale misspecified			Biomarker B_2 scale misspecified			Both misspecified		
	$P_\kappa(X=1\|Z)$		$P_\tau(X=1\|Z)$											
	$Z=1$	$Z=2$	$Z=1$	$Z=2$	Sample size	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
A	15%	30%	30%	60%	250	-3.8	36.9	37.1	-14.4	35.2	38.0	-15.6	36.4	39.6
					500	-3.2	25.8	26.0	-13.6	23.8	27.4	-14.9	25.2	29.3
					2500	-3.5	11.3	11.8	-14.4	10.3	17.7	-15.3	11.1	18.8
					10000	-2.9	5.8	6.5	-14	5.3	14.9	-14.7	5.7	15.8
B	15%	60%	30%	60%	250	-3.9	38.2	38.4	-7.4	37.0	37.7	-8.9	38.8	39.8
					500	-3.4	26.2	26.5	-6.7	25.1	26.0	-8.4	26.7	28.0
					2500	-3.4	11.7	12.1	-7.3	11.0	13.2	-8.4	11.9	14.5
					10000	-3.0	5.9	6.7	-7.0	5.6	9.0	-7.9	6.1	10.0
C	30%	60%	30%	60%	250	-0.9	39.7	39.7	-0.8	38.2	38.2	-0.9	39.7	39.7
					500	-0.3	27.9	27.9	-0.1	26.4	26.4	-0.3	27.9	27.9
					2500	-0.8	12.2	12.2	-1.2	11.4	11.5	-0.8	12.2	12.2
					10000	-0.1	6.3	6.3	-0.7	5.9	5.9	-0.1	6.3	6.3

Table 4.6: Simulation results evaluating the accuracy of the unadjusted, curve averaging, and sample weighting approaches for estimating mean window period, with infection duration T modeled on a linear scale. Bias, standard error (SE), and root mean squared error (RMSE) are given in days.

	Scenario					Unadjusted analysis (linear form)			Curve averaging (linear form)			Sample weighting (linear form)		
	$P_\kappa(X=1\|Z)$		$P_\tau(X=1\|Z)$											
	$Z=1$	$Z=2$	$Z=1$	$Z=2$	Sample size	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
A	15%	30%	30%	60%	250	-22.1	36.9	43.0	-1.3	50.5	50.5	-2.2	49.1	49.2
					500	-21.4	25.2	33.1	-0.4	33.5	33.5	-0.8	33.9	33.9
					2500	-21.9	11.2	24.6	-1.0	14.1	14.1	-0.6	14.5	14.5
					10000	-21.3	5.7	22.1	-0.4	7.1	7.1	0.1	7.3	7.3
B	15%	60%	30%	60%	250	-12.2	39.3	41.1	1.6	43.4	43.5	-1.3	44.8	44.8
					500	-11.7	26.9	29.4	1.4	29.0	29.0	-1.0	31.0	31.0
					2500	-11.7	12.0	16.7	0.8	12.7	12.7	-0.6	13.0	13.0
					10000	-11.2	6.1	12.8	1.1	6.4	6.5	0.0	6.6	6.6
C	30%	60%	30%	60%	250	-0.9	39.7	39.7	-1.0	39.5	39.5	-1.5	39.8	39.8
					500	-0.3	27.9	27.9	-0.5	27.6	27.6	-0.4	27.7	27.7
					2500	-0.8	12.2	12.2	-0.7	12.0	12.0	-0.7	12.0	12.0
					10000	-0.1	6.3	6.3	-0.1	6.1	6.1	-0.1	6.2	6.2

CHAPTER 5
Interval-censored seroconversion dates

Since seroconversion status is usually only tested periodically in longitudinal studies, seroconversion dates and durations of infection are typically interval-censored in calibration data sets, as discussed in Section 2.2. In the previous chapters and in the existing literature, this issue has been handled by uniform imputation over the censoring intervals. In this chapter, we present an alternative approach, using an incomplete-data perspective and the EM algorithm.

Several options exist for regression analysis with interval-censored covariates. One option is to treat the midpoints of the censoring intervals as if they were the observed values of the censored variable and to regress the outcome on these midpoints. We will refer to this approach as midpoint imputation. It has the appeal of simplicity, but we will demonstrate that it can lead to substantial bias when censoring intervals are wide.

A second option, discussed above, is to assume that the interval-censored variable has a uniform distribution over the censoring interval and to perform a multiple imputation analysis. In such a case a series of imputed data sets are constructed by selecting a random value from each censoring interval. Regression model coefficients are then estimated from each imputed data set, and these estimates are averaged to produce final estimates. (Konikoff et al. 2013) We will refer to this approach as uniform imputation. We will demonstrate that it can also lead to substantial bias, potentially more severe than midpoint imputation.

A third option is to simultaneously estimate the parameters of the regression model of

interest and the parameters of a model for the nuisance distribution of the interval-censored covariate (Hsiao 1983; Goggins et al. 1999; Gómez et al. 2003) We will refer to this approach as joint modeling.

In some cases, an interval-censored covariate is defined as a function of other variables, at least one of which is itself interval-censored. In the incidence estimation setting, the interval-censored covariate of interest is the duration of infection, defined as the time difference between the date of seroconversion, which is interval-censored, and the date of biomarker sample collection, which is recorded precisely. In this example, we could model the distribution of infection durations directly, ignoring the underlying seroconversion dates and biomarker measurement dates. To do so, the approach of Gómez et al (2003), referred to as the GEL approach, could be used. This approach aimed to model $p(Y = y|Z = z)$ when Z is censored in the interval $[Z_L, Z_R]$. The motivating example was a model of HIV viral load (Y) at start of secondary treatment, as a function of the time difference (Z) from primary treatment failure to start of secondary treatment. The following assumptions were made:

I. The data consist of n independent and identically distributed realizations of (Y, Z, Z_L, Z_R).

II. $p(Y = y|Z = z, Z_L = l, Z_R = r) = p_{\boldsymbol{\theta}}(Y = y|Z = z)$.

III. $p(Z = z|Z_L = l, Z_R = r) = 1\{z \in [l, r]\}\, p_{\boldsymbol{\omega}}(Z = z) \,/\, p_{\boldsymbol{\omega}}(Z \in [l, r])$.

IV. Z has a finite sample space $\boldsymbol{\mathcal{Z}} \subset \mathbb{R}$.

The distribution of Z was modeled non-parametrically; that is, as a multinomial distribution with no added assumptions. Let $\boldsymbol{\omega} = \{\omega(z) = p(Z = z)\}_{z \in \boldsymbol{\mathcal{Z}}}$ denote the parameters of this

distribution. Then the likelihood of Y, conditional on (Z_L, Z_R) and marginalizing over Z, is:

$$
\begin{aligned}
L(\boldsymbol{\omega}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} p_{\boldsymbol{\omega},\boldsymbol{\theta}}(y_i | l_i, r_i) \\
&= \prod_{i=1}^{n} \sum_{z \in \mathcal{Z}} p\big(y_i \big| Z_i = z, Z_{L_i} = l_i, Z_{R_i} = r_i\big)\, p(Z_i = z | l_i, r_i) \\
&= \prod_{i=1}^{n} \sum_{z \in \mathcal{Z}} p_{\boldsymbol{\theta}}(y_i | Z_i = z)\, 1\{z \in [l_i, r_i]\}\, p_{\boldsymbol{\omega}}(Z_i = z) / p_{\boldsymbol{\omega}}(Z_i \in [l_i, r_i]) \\
&= \left[\prod_{i=1}^{n} \{p_{\boldsymbol{\omega}}(Z_i \in [l_i, r_i])\}^{-1}\right] \prod_{i=1}^{n} \sum_{z \in \mathcal{Z} \cap [l_i, r_i]} p_{\boldsymbol{\theta}}(y_i | Z_i = z)\, \omega(z)
\end{aligned}
$$

This analysis focused on the partial likelihood

$$
L^*(\boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{z \in \mathcal{Z} \cap [l_i, r_i]} p_{\boldsymbol{\theta}}(y_i | Z_i = z)\, \omega(z)
$$

omitting the term $\prod_{i=1}^{n} \{p_{\boldsymbol{\omega}}(Z_i \in [l_i, r_i])\}^{-1} = \prod_{i=1}^{n} \{\sum_{z \in \mathcal{Z}} 1\{z \in [l_i, r_i]\}\, \omega(z)\}^{-1}$. The analysis sought to maximize $L^*\big(\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}}\big)$ by iterating between updates to $\widehat{\boldsymbol{\omega}}$ ("Step A") and updates to $\widehat{\boldsymbol{\theta}}$ ("Step B").

Step A in turn consisted of a further iteration between the following steps:

A[i]: For each $z \in \mathcal{Z}$, compute $p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(Z_i = z | l_i, r_i, y_i)$ using Bayes' Theorem and the assumptions:

$$
\begin{aligned}
p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(Z_i = z | l_i, r_i, y_i) &= \frac{p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}\big(y_i \big| Z_i = z, Z_{L_i} = l_i, Z_{R_i} = r_i\big) p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}\big(Z_i = z \big| Z_{L_i} = l_i, Z_{R_i} = r_i\big)}{\sum_{z \in \mathcal{Z}} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}\big(y_i \big| Z_i = z, Z_{L_i} = l_i, Z_{R_i} = r_i\big)\, p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}\big(Z_i = z \big| Z_{L_i} = l_i, Z_{R_i} = r_i\big)} \\
&= \frac{p_{\widehat{\boldsymbol{\theta}}}(y_i | z)\, 1\{z \in [l_i, r_i]\}\, p_{\widehat{\boldsymbol{\omega}}}(Z_i = z) / p_{\widehat{\boldsymbol{\omega}}}(Z \in [l, r])}{\sum_{z \in \mathcal{Z}} p_{\widehat{\boldsymbol{\theta}}}(y_i | z)\, 1\{z \in [l_i, r_i]\}\, p_{\widehat{\boldsymbol{\omega}}}(Z_i = z) / p_{\widehat{\boldsymbol{\omega}}}(Z \in [l, r])}
\end{aligned}
$$

$$= \frac{1\{z \in [l_i, r_i]\}\, p_{\widehat{\boldsymbol{\theta}}}(y_i|z)\, \widehat{\omega}(z)}{\sum_{z \in \mathcal{Z} \cap [l_i, r_i]}\, p_{\widehat{\boldsymbol{\theta}}}(y_i|z)\, \widehat{\omega}(z)}$$

A[ii]: For each $z \in \mathcal{Z}$, update:

$$\widehat{\omega}(z) \leftarrow \frac{1}{n} \sum_{i=1}^{n} p_{\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}}}(Z_i = z | l_i, r_i, y_i)$$

Steps A[i] and A[ii] were iterated until the relative norm difference $\left\|\boldsymbol{\omega}^{(new)} - \boldsymbol{\omega}^{(old)}\right\| / \left\|\boldsymbol{\omega}^{(old)}\right\|$ was less than a tolerance value.

Step B consisted of the update:

$$\widehat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}} \log L^*(\widehat{\boldsymbol{\omega}}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \left\{ \sum_{z \in \mathcal{Z} \cap [l_i, r_i]} p_{\boldsymbol{\theta}}(y_i|Z_i = z)\, \widehat{\omega}(z) \right\}$$

The maximizing value has a closed-form solution if $p_{\boldsymbol{\theta}}(y_i|z)$ is a linear model; otherwise, it can be found by numerical methods, such as the Broyden–Fletcher–Goldfarb–Shannon (BFGS) quasi-Newton method. (Langohr and Gómez Melis 2014; Bolker and R Core Team 2020)

Steps A and B were iterated until the sum of relative norm differences

$$\frac{\left\|\boldsymbol{\omega}^{(new)} - \boldsymbol{\omega}^{(old)}\right\|}{\left\|\boldsymbol{\omega}^{(old)}\right\|} + \frac{\left\|\boldsymbol{\theta}^{(new)} - \boldsymbol{\theta}^{(old)}\right\|}{\left\|\boldsymbol{\theta}^{(old)}\right\|}$$

was less than a tolerance value.

A limitation of the GEL approach for our application is that it does not account for calendar time, which is an important factor to consider in incidence estimation. The probability that a given individual becomes infected at a particular point in time depends on the contemporary population prevalence of infectious individuals with whom they might

interact.

In this chapter, we propose an alternative to the GEL approach: we can model the distribution of seroconversion dates directly and then derive the distribution of infection durations from this model. Using this approach, we derive a simpler estimation procedure than the GEL approach's procedure: we remove a second loop nested inside the main iteration loop, and we replace a Quasi-Newton maximization step with a faster step using Fisher scoring. (Lange 2010)

5.1 Notation

In this chapter, we use the following notation. A calibration data set D consists of the following random variables observed for each of N participants: the date when participant i enrolled in the study, E_i; the date of participant i's last seronegative test, L_i; the date of participant i's first seropositive test, R_i; the vector of participant i's sample collection dates after seroconversion, $\boldsymbol{O}_i = (O_{i1}, \dots, O_{in_i})$; and the vector of MAA classification outcomes for participant i's blood samples collected after seroconversion, $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})$. Each Y_{ij} is binary; $Y_{ij} = 1$ indicates a positive classification, and $Y_{ij} = 0$ indicates a negative classification. In addition, we define an unobserved variable: the date when participant i first seroconverted, S_i. We also define the following variable transformations to represent the time differences between a participant's actual seroconversion date and their biomarker sample collection dates: $T_{ij} = O_{ij} - S_i$ and $\boldsymbol{T}_i = (T_{i1}, \dots, T_{in_i})$. The corresponding observed values are e_i, l_i, r_i, $\boldsymbol{o}_i = (o_{i1}, \dots, o_{in_i})$, $\boldsymbol{y}_i = (y_{i1}, \dots, y_{in_i})$, s_i, $t_{ij} = o_{ij} - s_i$, and $\boldsymbol{t}_i = (t_{i1}, \dots, t_{in_i})$, respectively.

5.2 Assumptions

To enable our joint modeling analysis, we make nine modeling assumptions about the relationships among these variables. These assumptions can be grouped in several ways. Assumptions 1-3 are also used by the midpoint and uniform imputation approaches, while Assumptions 4-9 are used only for joint modeling. Assumptions 1, 3, 4, 5, and 6 specify independence relationships among the variables in our data set; they enable us to decompose the joint likelihood into a hierarchical model structure amenable to estimation. In this hierarchical model, the sub-model for Y_{ij} (the binary MAA classification of the sample collected on date O_{ij}) depends only on T_{ij} (the elapsed time since seroconversion), and the sub-model for S_i (the seroconversion date) depends only on E_i (the enrollment date). Assumptions 2-4 characterize the distribution of the outcome, $\boldsymbol{Y}_i$, assumptions 5 and 6 characterize the distributions of the enrollment and follow-up observation dates $(E_i, L_i, R_i, \boldsymbol{O}_i)$, and assumptions 8 and 9 characterize the distribution of the interval-censored covariate, S_i. Assumption 7 distinguishes the parameter sets for the various sub-models.

First, we assume that the participants' data are independently and identically distributed; that is, $(E_i, L_i, R_i, \boldsymbol{O}_i, \boldsymbol{Y}_i, S_i) \underset{iid}{\sim} p(e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i, s_i)$.

Second, we assume that $p(\boldsymbol{y}_i|\boldsymbol{t}_i)$ has a functional form which would be estimable if $\boldsymbol{t}_i$ were observed precisely; for example, a generalized additive model (Hastie and Tibshirani 1990)

Third, we assume that longitudinally repeated MAA classifications of the same individual are mutually independent, conditional on the duration of infection at the time of sample collection; i.e., $p(\boldsymbol{y}_i|\boldsymbol{t}_i) = \prod_{j\in 1:n_i} p(y_{ij}|\, t_{ij})$. This assumption is unnecessary when the

regression parameters are the only estimands of interest; in such cases, we can use any regression model $p(\boldsymbol{y}_i|\boldsymbol{t}_i)$ which we could fit with an uncensored covariate, including models with random effects, autoregressive parameters, or other forms of autocorrelation. This assumption is necessary for our particular motivating application because we want to transform the estimated regression model $\hat{p}(y_{ij}|t_{ij})$ into an estimated mean window period $\hat{\mu}$; if random effects or other forms of correlation are added to this model, there is no longer a straightforward way to use the model to estimate μ. This issue is further discussed in. Note that this assumption and the preceding two are not specific to the joint modeling approach; they are also relied upon for midpoint imputation and uniform imputation when we use these approaches for our application.

If individuals never re-entered the MAA-positive state after exiting, then this assumption would be clearly false, and a survival model for time-to-MAA-negativity should be used instead of the Bernoulli model that we are proposing here. The survival model approach to mean window period estimation has been considered elsewhere (Hanson et al. 2016) However, depending on the MAA being used, it is possible for individuals to return to the MAA-positive state long after exiting, especially as infections progress toward AIDS and the immune response weakens (Brookmeyer 2010) In order to allow for this possibility, we chose to consider the Bernoulli model.

Fourth, we assume that conditional on the vector of infection durations corresponding to the dates of biomarker sample collection, the corresponding MAA classification is independent of the participant's enrollment date, seroconversion date, seroconversion censoring interval endpoints, and biomarker sample collection dates; that is,

$p(\boldsymbol{y}_i|e_i, l_i, r_i, \boldsymbol{o}_i, s_i) = p(\boldsymbol{y}_i|\boldsymbol{t}_i)$.

Fifth, we assume that the follow-up dates through the first seropositive test are independent of the actual seroconversion date, conditional on enrollment date; that is, $p(l_i, r_i|e_i, s_i) = c(l_i, r_i; e_i)\ 1\{s_i \in [l_i, r_i]\}$, where $c(l_i, r_i; e_i)$ is the probability, conditional on $E_i = e_i$, that participant i's pre-seroconversion follow-up schedule includes tests at l_i followed by r_i. Given such a schedule, $L_i = l_i$ and $R_i = r_i$ if and only if $s_i \in [l_i, r_i]$. This assumption entails that $p(s_i|e_i, l_i, r_i) = p(s_i|S_i \in [l_i, r_i], e_i)$:

$$
\begin{aligned}
p(s_i|e_i, l_i, r_i) &= \frac{p(l_i, r_i|e_i, s_i)\ p(s_i|e_i)}{\sum_{s_i \in \mathcal{R}(S_i)} p(l_i, r_i|e_i, s_i)\ p(s_i|e_i)} \\
&= \frac{c(l_i, r_i; e_i)\ 1\{s_i \in [l_i, r_i]\}\ p(s_i|e_i)}{c(l_i, r_i; e_i) \sum_{s_i \in \mathcal{R}(S_i)} 1\{s_i \in [l_i, r_i]\}\ p(s_i|e_i)} \\
&= \frac{1\{s_i \in [l_i, r_i]\}\ p(s_i|e_i)}{\sum_{s_i \in \mathcal{R}(S_i)} 1\{s_i \in [l_i, r_i]\}\ p(s_i|e_i)} \\
&= \frac{1\{s_i \in [l_i, r_i]\}\ p(s_i|e_i)}{p(S_i \in [l_i, r_i]|e_i)} \\
&= \frac{p(S_i = s_i, S_i \in [l_i, r_i]|e_i)}{p(S_i \in [l_i, r_i]|e_i)} \\
&= p(S_i = s_i|S_i \in [l_i, r_i], e_i)
\end{aligned}
\tag{5.1}
$$

Note that the right-hand side of this equality is determined by $p(s_i|e_i)$, so both sides only depend on the parameters of $p(s_i|e_i)$. This relationship is referred to as non-informative censoring, and it is frequently assumed in analyses of interval-censored data (Gómez et al. 2003; Sun 2006) It is plausible for study designs with prespecified follow-up testing schedules, but it might be violated, for example, if study participants can request an earlier test date when they feel sick or believe they may have been exposed; e.g., after high-risk

behaviors.

Sixth, we assume that conditional on enrollment date and the censoring interval endpoints, the post-seroconversion observation dates are independent from the actual seroconversion date; that is, $p(\boldsymbol{o}_i|e_i, l_i, r_i, s_i) = p(\boldsymbol{o}_i|e_i, l_i, r_i)$. An equivalent but perhaps less intuitive formulation of this assumption is $p(s_i|e_i, l_i, r_i, \boldsymbol{o}_i) = p(s_i|e_i, l_i, r_i)$; we will make use of both formulations.

Seventh, we assume that the conditional distributions $p(e_i)$, $p(s_i|e_i)$, $p(l_i, r_i|e_i, s_i)$, $p(\boldsymbol{o}_i|e_i, l_i, r_i)$, and $p(y_{ij}|t_{ij})$ are characterized by disjoint parameter sets; we denote the parameters of $p(s_i|e_i)$ and $p(y_{ij}|t_{ij})$ by $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$, respectively.

Eighth, we assume that the seroconversion date, S_i, has a countable sample space, $\mathcal{S} = \{s_1 < s_2 < \cdots\} \subset \mathbb{R}^+$, consisting of an evenly-spaced grid of dates starting with $s_1 = \min_{i \in 1:N} l_i$ and including at least one date from every censoring interval in the data set; the spacing of this grid, $\gamma = s_{k+1} - s_k$, can be as small as computationally feasible, thus approximating a continuous distribution. This assumption is a simplification of the true data-generating process, but it greatly simplifies the subsequent analysis; it enables us to compute expectations over the possible seroconversion dates as sums, rather than as integrals which may not be analytically solvable. It should also be noted that in most cases, the exact clock time of blood sample collection is not recorded; thus, the observed data are already effectively discretized at the day level.

Ninth, we assume that $p_{\boldsymbol{\omega}}(s_i|e_i) = 1\{s_i \geq e_i\}\, \omega(s_i) \prod_{u \in \mathcal{S} \cap [e_i, s_i)} \{1 - \omega(u)\}$; that is, conditional on enrollment date E_i, the distribution of S_i is analogous to a non-homogeneous shifted geometric distribution, with parameter set $\boldsymbol{\omega} = \{\omega(s) =$

$p(S_i = s|S_i \geq s, E_i = e_i)\,; s \in \mathcal{S}\}$, indexed by calendar time s. Calendar time, rather than time since enrollment or time until the first seropositive test date, is chosen as the basis for the parametrization because the risk of infection is viewed as a function of the contemporary population disease prevalence.

Using these assumptions, we will now decompose the joint likelihood of the observed data into a hierarchical model. We will then maximize the decomposed joint likelihood using the well-known EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 2007)

5.3 Approach

If S_i were observed, then the likelihood contribution from each participant's MAA classification data would be $\mathcal{L}_i^* = p(e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i, s_i)$. This expression is the "complete-data likelihood" for individual i. Since S_i is not observed, we apply the law of total probability to marginalize $\mathcal{L}_i^*$ over s_i and obtain the "observed-data likelihood":

$$\mathcal{L}_i = p(e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i) = \sum_{s_i \in \mathcal{S}} p(e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i, s_i) = \sum_{s_i \in \mathcal{S}} \mathcal{L}_i^*$$

The statistical objective is to model $\phi(t) = p(Y_{ij} = 1|T_{ij} = t)$; in order to express $\mathcal{L}_i$ in a form involving $p(y_{ij}|t_{ij})$, we decompose $\mathcal{L}_i^*$ into a hierarchical model reflecting the study design, and we simplify this decomposition using our assumptions 2, 4, 5, and 6:

$$\begin{aligned}\mathcal{L}_i^* &= p(e_i)\; p(s_i|e_i)\; p(l_i, r_i|e_i, s_i)\; p(\boldsymbol{o}_i|e_i, l_i, r_i, s_i)\; p(\boldsymbol{y}_i|e_i, l_i, r_i, \boldsymbol{o}_i, s_i) \\ &= p(e_i)\; p_{\boldsymbol{\omega}}(s_i|e_i)\; p(l_i, r_i|e_i, s_i)\; p(\boldsymbol{o}_i|e_i, l_i, r_i) \prod_{j \in 1:n_i} p_{\boldsymbol{\theta}}(y_{ij}|t_{ij})\end{aligned}$$

Correspondingly, the observed-data likelihood contribution for participant i is:

$$\mathcal{L}_i = p(e_i)\, p(\boldsymbol{o}_i|e_i, l_i, r_i) \sum_{s_i \in \mathcal{S}} p_{\boldsymbol{\omega}}(s_i|e_i)\, p(l_i, r_i|e_i, s_i) \prod_{j \in 1:n_i} p_{\boldsymbol{\theta}}(y_{ij}|t_{ij})$$

The observed-data likelihood for the data set is then $\mathcal{L} = \prod_{i \in 1:N} \mathcal{L}_i$, and the observed-data log-likelihood is:

$$\ell = \sum_{i \in 1:N} \log p(e_i) + \log p(\boldsymbol{o}_i|e_i, l_i, r_i) + \log \left\{ \sum_{s_i \in \mathcal{S}} p_{\boldsymbol{\omega}}(s_i|e_i)\, p(l_i, r_i|e_i, s_i) \prod_{j \in 1:n_i} p_{\boldsymbol{\theta}}(y_{ij}|t_{ij}) \right\}$$

5.3.1 Estimation procedure

We would like to estimate $p_{\boldsymbol{\theta}}(y_{ij}|t_{ij})$ by maximizing the observed-data log-likelihood, ℓ, but the third term of ℓ is the logarithm of a sum and thus is challenging to maximize directly. Fortunately, the theory of the EM algorithm proves that for a given parametrized complete-data model $p_{\boldsymbol{\Psi}}(e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i, s_i)$ and initial parameter estimate $\widehat{\boldsymbol{\Psi}}$, iterating the following two steps will monotonically increase the observed-data likelihood $\mathcal{L}$ toward a local maximum or saddlepoint; the latter can be escaped by randomly perturbing the converged solution (McLachlan and Krishnan 2007) In the E step, we calculate the expectation of the complete-data log-likelihood, conditional on the observed data and assuming that the parameters of the distribution of the unobserved variables, given the observed variables, are equal to the current parameter estimates $\widehat{\boldsymbol{\Psi}}$:

$$Q\big(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\big) = \sum_{i \in 1:N} \mathrm{E}_{\widehat{\boldsymbol{\Psi}}}[\log \mathcal{L}_i^*(\boldsymbol{\Psi})\,|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i] = \sum_{i \in 1:N} \sum_{s_i \in \mathcal{S}} \log\{\mathcal{L}_i^*(\boldsymbol{\Psi})\}\, p_{\widehat{\boldsymbol{\Psi}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

In the M step, we maximize $Q(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}})$ over the possible values of $\boldsymbol{\Psi}$ and update $\widehat{\boldsymbol{\Psi}}$ to this new value; the function $Q(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}})$ is often easier to maximize than the observed-data log-likelihood. We will now specify this algorithm for our data analysis problem.

E step:

To complete the E step, we need to solve for $p_{\hat{\boldsymbol{\Psi}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$. Applying Bayes' Theorem and our assumptions, we find:

$$p_{\hat{\boldsymbol{\Psi}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i) = \frac{p_{\hat{\boldsymbol{\Psi}}}(\boldsymbol{y}_i|e_i, l_i, r_i, \boldsymbol{o}_i, s_i)\, p_{\hat{\boldsymbol{\Psi}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i)}{\sum_{s_i \in \mathcal{S}} p_{\hat{\boldsymbol{\Psi}}}(\boldsymbol{y}_i|e_i, l_i, r_i, \boldsymbol{o}_i, s_i)\, p_{\hat{\boldsymbol{\Psi}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i)}$$

$$= \frac{p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}_i|\boldsymbol{t}_i)\, p_{\hat{\boldsymbol{\omega}}}(s_i|S_i \in [l_i, r_i], e_i)}{\sum_{s_i \in \mathcal{S}} p_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}_i|\boldsymbol{t}_i)\, p_{\hat{\boldsymbol{\omega}}}(s_i|S_i \in [l_i, r_i], e_i)}$$

We can calculate $p_{\hat{\boldsymbol{\omega}}}(s_i|S_i \in [l_i, r_i], e_i)$ in that expression as follows:

$$p_{\boldsymbol{\omega}}(S_i = s_i|S_i \in [l_i, r_i], E_i = e_i) = \frac{p_{\boldsymbol{\omega}}(S_i = s_i, S_i \in [l_i, r_i]|E_i = e_i)}{p_{\boldsymbol{\omega}}(S_i \in [l_i, r_i]|E_i = e_i)}$$

$$= \frac{p_{\omega}(S_i \in [l_i, r_i]|S_i = s_i, E_i = e_i)\, p_{\omega}(S_i = s_i|E_i = e_i)}{p_{\omega}(S_i \in [l_i, r_i]|E_i = e_i)}$$

$$= \frac{1\{s_i \in [l_i, r_i]\}\, p_{\omega}(S_i = s_i|E_i = e_i)}{p_{\omega}(S_i \in [l_i, r_i]|E = e_i)}$$

$$= 1\{s_i \in [l_i, r_i]\} \frac{p_{\omega}(S_i = s_i, S_i \geq l_i|E_i = e_i)}{p_{\omega}(S_i \leq r_i, S_i \geq l_i|E_i = e_i)}$$

$$= 1\{s_i \in [l_i, r_i]\} \frac{p_{\omega}(S_i = s_i|S_i \geq l_i, E_i = e_i)}{p_{\omega}(S_i \leq r_i|S_i \geq l_i, E_i = e_i)} \frac{p_{\omega}(S_i \geq l_i, E_i = e_i)}{p_{\omega}(S_i \geq l_i, E_i = e_i)}$$

$$= 1\{s_i \in [l_i, r_i]\} \frac{p_{\omega}(S_i = s_i|S_i \geq l_i, E_i = e_i)}{p_{\omega}(S_i \leq r_i|S_i \geq l_i, E_i = e_i)}$$

$$= 1\{s_i \in [l_i, r_i]\} \frac{p_{\omega}(S_i = s_i|S_i \geq l_i, E_i = e_i)}{1 - p_{\omega}(S_i > r_i|S_i \geq l_i, E_i = e_i)}$$

$$= 1\{s_i \in [l_i, r_i]\} \frac{1\{s_i \geq e_i\}\, \omega(s_i) \prod_{u \in \boldsymbol{s} \cap [\max(l_i, e_i), s_i)} (1 - \omega(u))}{1 - \prod_{u \in \boldsymbol{s} \cap [\max(l_i, e_i), r_i]} (1 - \omega(u))}$$

$$= \ 1\{s_i \in [l_i, r_i], s_i \geq e_i\} \frac{\omega(s_i) \prod_{u \in \boldsymbol{s} \cap [\max(l_i, e_i), s_i)} (1 - \omega(u))}{1 - \prod_{u \in \boldsymbol{s} \cap [\max(l_i, e_i), r_i]} (1 - \omega(u))}$$

Thus, to perform the E step we only need estimates for $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$.

M step:

Expanding the term $\log \mathcal{L}_i^*(\boldsymbol{\Psi})$ in $Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$, we have:

$$\log \mathcal{L}_i^*(\boldsymbol{\Psi}) = \log p(e_i) + \log p_{\boldsymbol{\omega}}(s_i|e_i) + \log p(l_i, r_i|e_i, s_i) + \log p(\boldsymbol{o}_i|e_i, l_i, r_i) + \sum_{j \in 1:n_i} \log p_{\boldsymbol{\theta}}(y_{ij}|t_{ij})$$

By assumption 5, each of these terms involves a disjoint set of parameters; thus to maximize $Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$, we can maximize each term's expectation separately. Further, the terms $\log p(e_i)$ and $\log p(\boldsymbol{o}_i|e_i, l_i, r_i)$ do not involve s_i, so their expectations are merely the original terms. They can be maximized immediately in the first M step and do not need to be revisited in subsequent iterations. Since we are not interested in these distributions and the E step does not require the parameters of these distributions, we can ignore the details of specifying and maximizing them. For the same reason, we can also ignore the term $\log p(l_i, r_i|e_i, s_i)$. Thus, maximizing $Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$ over $\boldsymbol{\Psi}$ reduces to maximizing $Q^*\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right) = Q_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right) + Q_{\boldsymbol{\omega}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$ over $\boldsymbol{\omega}$ and $\boldsymbol{\theta}$, respectively, where:

$$Q_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right) = \sum_{i \in 1:N} \sum_{s_i \in \mathcal{S}} \sum_{j \in 1:n_i} \log\left\{p_{\boldsymbol{\theta}}\left(y_{ij}\middle|t_{ij}\right)\right\} p_{\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

$$Q_{\boldsymbol{\omega}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right) = \sum_{i \in 1:N} \sum_{s_i \in \mathcal{S}} \log\{p_{\boldsymbol{\omega}}(s_i|e_i)\} p_{\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

That is, the M step can be subdivided into two parallel sub-steps, $\widehat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}} Q_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$ and $\widehat{\boldsymbol{\omega}} \leftarrow \arg\max_{\boldsymbol{\omega}} Q_{\boldsymbol{\omega}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$. Assuming a discrete distribution for the interval-censored covariate ensures that these expressions are tractable to maximize.

We can maximize $Q_{\boldsymbol{\omega}}\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\right)$ analytically. For $s, u \in \boldsymbol{\mathcal{S}} \cap [e_i, \infty)$, we have:

$$\log p_{\boldsymbol{\omega}}(s_i|e_i) = \log \omega(s_i) + \sum_{v \in \boldsymbol{\mathcal{S}} \cap [e_i, s_i)} \log\big(1 - \omega(v)\big)$$

$$\frac{d}{d\omega(u)} \log p_{\boldsymbol{\omega}}(s_i|e_i) = 1\{u = s_i\} \frac{1}{\omega(u)} - 1\{u \in [e_i, s_i)\} \frac{1}{1 - \omega(u)}$$

$$\frac{d}{d\omega(u)} Q_{\boldsymbol{\omega}}\big(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\big) = \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} \left(\frac{d}{d\omega(u)} \log p_{\boldsymbol{\omega}}(s_i|e_i) \right) p_{\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

As a temporary shorthand, let $\mathscr{p} = p_{\widehat{\boldsymbol{\omega}}, \widehat{\boldsymbol{\theta}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$; then, setting $\frac{d}{d\omega(u)} \mathrm{Q}_{\omega}\big(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}\big) = 0$ and solving for $\omega(u)$, we have:

$$\sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} \left(\frac{d}{d\omega(u)} \log p_{\boldsymbol{\omega}}(s_i|e_i) \right) \mathscr{p} = 0$$

$$\Leftrightarrow \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} \left(1\{u = s_i\} \frac{1}{\omega(u)} - 1\{u \in [e_i, s_i)\} \frac{1}{1 - \omega(u)} \right) \mathscr{p} = 0$$

$$\Leftrightarrow \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} 1\{u = s_i\} \frac{1}{\omega(u)} \mathscr{p} - 1\{u \in [e_i, s_i)\} \frac{1}{1 - \omega(u)} \mathscr{p} = 0$$

$$\Leftrightarrow \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} 1\{u = s_i\} \frac{1}{\omega(u)} \mathscr{p} - \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} 1\{u \in [e_i, s_i)\} \frac{1}{1 - \omega(u)} \mathscr{p} = 0$$

$$\Leftrightarrow \frac{1}{\omega(u)} \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} 1\{u = s_i\} \mathscr{p} = \frac{1}{1 - \omega(u)} \sum_{i \in 1:N} \sum_{s_i \in \boldsymbol{\mathcal{S}}} 1\{u \in [e_i, s_i)\} \mathscr{p}$$

$$\Leftrightarrow \big(1 - \omega(u)\big) \sum_{i} \sum_{s_i} 1\{u = s_i\} \mathscr{p} = \omega(u) \sum_{i} \sum_{s_i} 1\{u \in [e_i, s_i)\} \mathscr{p}$$

$$\Leftrightarrow \left(\sum_{i} \sum_{s_i} 1\{u = s_i\} \mathscr{p} - \omega(u) \sum_{i} \sum_{s_i} 1\{u = s_i\} \mathscr{p} \right) = \omega(u) \sum_{i} \sum_{s_i} 1\{u \in [e_i, s)\} \mathscr{p}$$

$$\Leftrightarrow \sum_{i} \sum_{s_i} 1\{u = s_i\} \mathscr{p} = \omega(u) \sum_{i} \sum_{s_i} \big(1\{u \in [e_i, s_i)\} + 1\{u = s_i\}\big) \mathscr{p}$$

$$\Leftrightarrow \sum_i \sum_{s_i} 1\{u = s_i\} \mathcal{p} = \omega(u) \sum_i \sum_{s_i} 1\{u \in [e_i, s_i]\}\, \mathcal{p}$$

$$\Leftrightarrow \sum_i \sum_{s_i} 1\{u = s_i\} \mathcal{p} = \omega(u) \sum_i \sum_{s_i} 1\{e_i \le u\} 1\{s_i \ge u\}\, \mathcal{p}$$

$$\Leftrightarrow \sum_i \sum_{s_i} 1\{u = s_i\} \mathcal{p} = \omega(u) \sum_i 1\{e_i \le u\} \sum_{s_i} 1\{s_i \ge u\}\, \mathcal{p}$$

$$\Leftrightarrow \sum_i \sum_{s_i} 1\{u = s_i\}\, p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(s_i | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

$$= \omega(u) \sum_i 1\{e_i \le u\} \sum_{s_i} 1\{s_i \ge u\}\, p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(s_i | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

$$\Leftrightarrow \sum_i p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(u | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i) = \omega(u) \sum_i 1\{e_i \le u\}\, p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(S_i \ge u | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$$

$$\Leftrightarrow \omega(u) = \frac{\sum_{i \in 1:N} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(u | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)}{\sum_{i \in 1:N} 1\{e_i \le u\}\, p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(S_i \ge u | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)}$$

Thus, we have the following closed-form update formula:

$$\widehat{\omega}(u) \leftarrow \frac{\sum_{i \in 1:N} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(S_i = u | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)}{\sum_{i \in 1:N} 1\{e_i \le u\} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(S_i \ge u | e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)}$$

In this update, we estimate the probability of seroconverting on day s, conditional on not seroconverting prior to s, as the sum of the probabilities that each participant seroconverted at day s, divided by the sum of the probabilities that each participant was at risk of seroconverting on day s. These estimates resemble the factors of the Kaplan-Meier product-limit estimate of a survival function (E. L. Kaplan and Meier 1958) To ensure computational stability for the time points in the latest censoring interval of a data set, we can add a small offset to the denominator of this update formula, such as 0.1. This offset results in a small amount of regularization.

The objective function for updating $\widehat{\boldsymbol{\theta}}$ is equivalent to the log-likelihood of a weighted regression model, where the data points are the possible completions of our observed data and the weights are the probabilities of those completions, given the observed data and the current parameter estimates. Thus, if we have assumed $p_{\boldsymbol{\theta}}(y_{ij}|t_{ij})$ is a generalized additive model, we can find the maximizing value using optimization algorithms such as Fisher scoring, implemented in standard software such as the "bigglm()" function in R (Lumley 2013; R Core Team 2019)

In summary, the M step reduces to two parallel sub-steps:

$$\widehat{\omega}(s) \leftarrow \frac{\sum_{i\in 1:N} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(S_i = s|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)}{\sum_{i\in 1:N} 1\{e_i \leq s\} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(S_i \geq s|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)} \text{ for each } s \in \mathcal{S} \tag{5.2}$$

$$\widehat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{i\in 1:N} \sum_{s_i\in\mathcal{S}} \sum_{j\in 1:n_i} \log\{p_{\boldsymbol{\theta}}(y_{ij}|t_{ij})\} p_{\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\theta}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i) \tag{5.3}$$

These sub-steps are computed separately from each other; the updated values of $\widehat{\boldsymbol{\omega}}$ are not used in the update to $\widehat{\boldsymbol{\theta}}$ in the same EM iteration, and vice versa. Each parameter is updated only once per EM iteration.

5.3.2 Convergence criteria

We have now completely specified the E and M steps of an EM algorithm. To assess convergence of the algorithm, we can monitor the relative change in the observed-data likelihood between iterations k and $k+1$:

$$\Delta_{\mathcal{L}}^{(k)} = \frac{\mathcal{L}^{(k+1)} - \mathcal{L}^{(k)}}{\mathcal{L}^{(k)}} = \frac{\mathcal{L}^{(k+1)}}{\mathcal{L}^{(k)}} - 1$$

Alternatively, we can monitor the absolute change in the observed data log-likelihood:

$$\Delta_{\ell}^{(k)} = \log \mathcal{L}^{(k+1)} - \log \mathcal{L}^{(k)} = \log\left(\frac{\mathcal{L}^{(k+1)}}{\mathcal{L}^{(k)}}\right) = \log\left(\Delta_{\mathcal{L}}^{(k)} + 1\right)$$

Since $\lim_{x \to 0}(\log\{x+1\} - x) = 0$, these metrics are asymptotically equivalent. We can calculate $\Delta_{\mathcal{L}}^{(k)}$ as follows:

$$\Delta_{\mathcal{L}}^{(k)} = \frac{\mathcal{L}^{(k+1)}}{\mathcal{L}^{(k)}} - 1 = \frac{\prod_{i \in 1:N} \mathcal{L}_i^{(k+1)}}{\prod_{i \in 1:N} \mathcal{L}_i^{(k)}} - 1 = \prod_{i \in 1:N} \frac{\mathcal{L}_i^{(k+1)}}{\mathcal{L}_i^{(k)}} - 1$$

$$= \prod_{i \in 1:N} \frac{p(e_i)p(\boldsymbol{o}_i|e_i, l_i, r_i) \sum_{s_i \in \mathcal{S}} \left[p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i)\, p(l_i, r_i|e_i, s_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})\right]}{p(e_i)p(\boldsymbol{o}_i|e_i, l_i, r_i) \sum_{s_i \in \mathcal{S}} \left[p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i)\, p(l_i, r_i|e_i, s_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})\right]} - 1$$

$$= \prod_{i \in 1:N} \frac{\sum_{s_i \in \mathcal{S}} p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i)\, p(l_i, r_i|e_i, s_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})}{\sum_{s_i \in \mathcal{S}} p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i)\, p(l_i, r_i|e_i, s_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})} - 1$$

$$= \prod_{i \in 1:N} \frac{\sum_{s_i \in \mathcal{S}} p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i)\, c(l_i, r_i; e_i)\, 1\{s_i \in [l_i, r_i]\} \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})}{\sum_{s_i \in \mathcal{S}} p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i)\, c(l_i, r_i; e_i)\, 1\{s_i \in [l_i, r_i]\} \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})} - 1$$

$$= \prod_{i \in 1:N} \frac{c(l_i, r_i; e_i) \sum_{s_i \in \mathcal{S}} p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i)\, 1\{s_i \in [l_i, r_i]\} \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})}{c(l_i, r_i; e_i) \sum_{s_i \in \mathcal{S}} p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i)\, 1\{s_i \in [l_i, r_i]\} \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})} - 1$$

$$= \prod_{i \in 1:N} \frac{\sum_{s_i \in \mathcal{S}} 1\{s_i \in [l_i, r_i]\}\, p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})}{\prod_{i \in 1:N}\left(\sum_{s_i \in \mathcal{S}} 1\{s_i \in [l_i, r_i]\}\, p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})\right)} - 1$$

$$= \prod_{i \in 1:N} \frac{\sum_{s_i \in \mathcal{S} \cap [l_i, r_i]} p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})}{\sum_{s_i \in \mathcal{S} \cap [l_i, r_i]} p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})} - 1$$

Correspondingly, the change in log-likelihood reduces to:

$$\Delta_{\ell}^{(k)} = \ell^{(k+1)} - \ell^{(k)} = \sum_{i=1}^{n} \log \frac{\sum_{s_i \in \mathcal{S} \cap [l_i, r_i]} p_{\widehat{\boldsymbol{\omega}}^{(k+1)}}(s_i|e_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k+1)}}(y_{ij}|t_{ij})}{\sum_{s_i \in \mathcal{S} \cap [l_i, r_i]} p_{\widehat{\boldsymbol{\omega}}^{(k)}}(s_i|e_i) \prod_{j \in 1:n_i} p_{\widehat{\boldsymbol{\theta}}^{(k)}}(y_{ij}|t_{ij})}$$

Note that we have canceled $p(e_i)$, $p(\boldsymbol{o}_i|e_i, l_i, r_i)$, and $c(l_i, r_i; e_i)$ out of this expression; hence it is computable, even though we never actually parametrized or estimated those functions.

In practice, we found it difficult to choose a tolerance level for Δ_ℓ that was sufficient for small sample sizes and necessary for large sample sizes; for large sample sizes, the increase in log-likelihood can still seem substantial for many iterations after $\widehat{\boldsymbol{\theta}}$ is no longer changing appreciably. As a computational shortcut for the simulation study below, we used a relatively lax convergence cutoff for the change in log-likelihood, $\Delta_\ell^{(k)} < 0.1$, and we added another metric based on the relative change in $\widehat{\boldsymbol{\theta}}$: $\Delta_{\boldsymbol{\theta}}^{(k)} = \max_{j \in 1:p} |\hat{\theta}_j^{(k+1)} - \hat{\theta}_j^{(k)}| / |\hat{\theta}_j^{(k)}| < 0.0001$.

Note that this metric does not include $\widehat{\boldsymbol{\omega}}$. We had two reasons for only considering the change in $\widehat{\boldsymbol{\theta}}$ and not in $\widehat{\boldsymbol{\omega}}$. First, the purpose of our analysis is to estimate μ, which is a function of $\boldsymbol{\theta}$ and not $\boldsymbol{\omega}$; we consider $\boldsymbol{\omega}$ to be a nuisance parameter which we only need in order to account for the uncertainty about the precise value of our covariate due to interval-censoring. For our particular motivating application, we might have even ignored $\widehat{\boldsymbol{\theta}}$ and judged convergence based on the change in $\hat{\mu}$, which is the estimate we are ultimately most interested in; our implementation of the algorithm includes this metric as an alternative option. However, for the purposes of this chapter, we chose to present the criterion based on $\widehat{\boldsymbol{\theta}}$**,** since in other applications of this approach, the regression parameters may be the estimands of primary interest. Second, the final few $\hat{\omega}(s)$ estimates (ordered by calendar time) take many iterations to converge, because they are estimated using only the data from the last participants to seroconvert. However, this instability also does not substantially affect the overall likelihood, since these parameters only affect the likelihood contributions from those last few participants.

Even if we considered the full set of parameters $(\boldsymbol{\omega}, \boldsymbol{\theta})$, the theory of the EM algorithm does not guarantee that the relative or absolute change in the parameters decreases

monotonically. Thus, even if $\widehat{\boldsymbol{\theta}}$ has not changed much for several iterations, these estimates may start changing substantially again in a later iteration. Therefore, we used both the likelihood convergence criterion $\Delta_{\ell}^{(k)}$ and the parameter convergence criterion $\Delta_{\boldsymbol{\theta}}^{(k)}$, and we stopped our algorithm only when both criteria fall below pre-specified cutoffs.

5.3.3 Uncertainty quantification

As discussed in Chapter 2, the bootstrap approach has been used in combination with uniform imputation to estimate standard errors and construct confidence intervals for $\hat{\mu}$. (Efron 1979; Konikoff et al. 2013) The bootstrap can also be used with joint modeling. To preserve the longitudinal structure of the data set, bootstrap resampling is performed at the participant level, with all observations from a resampled participant included in the bootstrapped data set as many times as that individual is sampled, with a new synthetic ID generated each time an individual is resampled. Once a bootstrap data set has been generated, the analysis can be run on this data set, producing an estimate of μ. This process is repeated for e.g. 1000 bootstrap data sets, generating a corresponding number of bootstrapped $\hat{\mu}$ estimates. The standard error of $\hat{\mu}$ can then be estimated using the standard deviation of the bootstrapped $\hat{\mu}$ estimates, and a 95% confidence interval can be constructed using the 2.5% and 97.5% quantiles of the distribution of bootstrapped $\hat{\mu}$ estimates. We demonstrate this approach in the simulation analyses below. The same process can also be used to generate standard errors and confidence intervals for $\widehat{\boldsymbol{\theta}}$ when the regression parameters are the estimands of interest. The bootstrap is computationally expensive, especially when combined with an EM algorithm for each bootstrapped data set, but it is feasible, especially since the bootstrapped data sets can be analyzed simultaneously in

parallel, if sufficient computing resources are available.

5.4 Comparison of GEL approach and current approach

Our data structure $(E, L, R, \boldsymbol{O}, \boldsymbol{Y}, S)$ is an extension of the GEL approach's data structure (Y, Z, Z_L, Z_R); note that Z is analogous to $\boldsymbol{T} = \boldsymbol{O} - S$, Z_L is analogous to $\boldsymbol{O} - R$, and Z_R is analogous to $\boldsymbol{O} - L$. Accordingly, with minor modifications to account for repeated measurements on each participant, the GEL approach could be performed on our data set. Our approach allows a similar but simpler estimation procedure.

5.4.1 Similarities

Our approach borrows several key ideas from the GEL approach, including iteratively maximizing the likelihood, the discrete approximation for the sample space of the interval-censored covariate, and an adaptation of the assumptions to our setting; our assumptions 1, 2, 3, and 6 are analogous to assumptions I-IV of the GEL approach, respectively. Our E step is analogous to step A[i] in the GEL approach; both consist of calculating the distribution of the censored covariate, conditional on the observed variables and the current parameter estimates.

5.4.2 Differences in model specification

As discussed above, the GEL approach directly models the distribution of Z, whereas our approach indirectly models the distribution of T as a function of the distributions of $\boldsymbol{O}$ and S. We model S using a non-homogenous geometric distribution, conditional on study enrollment date and with parameters indexed by calendar date; in contrast, the GEL approach modeled Z using a multinomial distribution with parameters indexed by the time difference between the date of the censored event and the outcome measurement date. Our

modeling choice was motivated by the characteristics of our intended application: we view the risk of infection at a given time point as a function of enrollment date and the time-varying population prevalence starting from that date, interacting with individual risk behaviors.

5.4.3 Differences in estimation procedure

The updates for $\widehat{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\theta}}$ in the M step are analogous to step 0 and B of the GEL approach, respectively, but there are noteworthy differences. We perform the update for $\widehat{\boldsymbol{\omega}}$ once per update to $\widehat{\boldsymbol{\theta}}$, avoiding the inner loop of Steps A[i] and A[ii]. Furthermore, our update for $\widehat{\boldsymbol{\theta}}$ maximizes the expectation of the logarithm of the outcome distribution, whereas the GEL approach maximizes the logarithm of the expectation. Consequently, the GEL approach for generalized linear models cannot use Fisher scoring to find the MLEs; instead, a version of the Broyden–Fletcher–Goldfarb–Shannon (BFGS) Quasi-Newton method is used. Using Fisher scoring instead of BFGS resulted in a substantial speed increase for our analysis. Also, in the GEL approach, the expectation in step A[ii] does not condition on y_i.

5.5 Simulation study

To evaluate the performance characteristics of our joint modeling approach, as well as midpoint imputation and uniform imputation, we developed a data-generating model, which we used to produce simulated data sets. We designed the data-generating model based on the FHI 360 HC-HIV and GS studies, described in Section 2.4.2.

5.5.1 Data-generating model

We began by specifying a cohort size, N_0. We considered two sizes: $N_0 = 4500$ participants, to resemble the HC-HIV study, and $N_0 = 100{,}000$ participants, to examine the methods'

large-sample properties.

Next, for each participant, we simulated the enrollment date E_i from a discrete uniform distribution on the first 366 dates starting from the study start date. We assigned each participant a corresponding study exit date, $F_i = E_i + 3650$ days, representing the end of follow-up ten years after enrollment.

We assumed that only a small fraction of the population is at risk of infection; specifically, we assumed that each cohort participant has a $\pi = 0.05$ probability of being at risk. We assigned each simulated participant an at-risk status A_i as a Bernoulli random variable with $p(A_i = 1) = \pi$. Simulated individuals with $A_i = 0$ have no chance of contributing observations to the calibration data set, since they will never seroconvert. For the participants with $A_i = 1$, we assumed a time-to-event model for the distribution of seroconversion dates with a linearly changing instantaneous hazard rate $\lambda(t) = \alpha + \beta t$, where t is time since the study start date (in years), α is the hazard of seroconversion at study start (events per person-year), and β is the change in hazard per calendar year (events per person-year^2). We considered the following seven pairs of α and β values: (0, 0.5), (0, 1), (0, 2), (1, 0), (1, 0.5), (10, 0), and (10, 0.5).

From this hazard model, we calculated the inverse survival function $\mathrm{G}_i^{-1}(u)$ as follows. Letting t_0 denote the study start date and $t(s) = (s - t_0)/365$ denote the elapsed time (in years) from t_0 to s, this hazard model leads to the following participant-specific cumulative hazard function:

$$\Lambda_i\big(t(s)\big) = \int_{u=t(e_i)}^{t(s)} (\alpha + \beta u)du = \left[\alpha t(s) \; + \frac{\beta}{2}\{t(s)\}^2\right] - \left[\alpha t(e_i) + \frac{\beta}{2}\{t(e_i)\}^2\right]$$

The corresponding survival function is $\mathrm{G}_i(s) = p(S_i \geq s | E_i = e_i) = \exp\{-\Lambda_i(t(s))\}$. If $\beta \neq 0$, then by the quadratic formula:

$$\mathrm{G}_i^{-1}(u) = t_0 + \frac{-\alpha \pm \sqrt{\alpha^2 - 2\beta\{\log(u) - [\alpha t(e_i) + \beta\{t(e_i)\}^2/2]\}}}{\beta} \times 365 \text{ days/year}$$

For $\beta = 0.5$, the positive square-root is the one of interest, since otherwise $\mathrm{G}_i^{-1}(u) < t_0$. For $\beta = 0$, the distribution reduces to a shifted exponential distribution with rate parameter α; then the survival function is $\mathrm{G}_i(s) = \exp[-\alpha\{t(s) - t(e_i)\}]$, and the inverse survival function is $\mathrm{G}_i^{-1}(u) = t_0 + [t(e_i) - \{log(u)/\alpha\}] \times 365$ days/year. We then simulated the seroconversion date $S_i = \mathrm{G}_i^{-1}(U_i)$, where U_i has a standard continuous uniform distribution.

We considered two protocols for pre-seroconversion follow up: testing every 12 weeks (84 days), as in the HC-HIV study, or testing every year (365 days). Let $\delta \in \{84, 365\}$ denote this parameter. For each design, we assumed that there would be a small amount of random deviation from the protocol in scheduling each test; specifically, each test is scheduled $\delta + D_{ij}$ days after the last test, where D_{ij} is simulated from a discrete uniform distribution on the integers $\{-7, \ldots, 7\}$. These offsets, combined with the variation in study enrollment dates, entail that the resulting censoring intervals are not limited to a mutually exclusive set of calendar-time intervals; instead, the participants' censoring intervals can partially overlap with each other. This modeling choice is realistic and helps avoid edge cases in which the EM algorithm struggled to converge. The seroconversion interval is defined as $[L_i, R_i]$, where L_i is the date of the last test before S_i, and R_i is the date of the first test after S_i. Simulated participants for whom seroconversion is not detected before the end of their follow-up duration (i.e., $R_i > F_i$) do not contribute any observations to the calibration data set and

were removed from the subsequent data-generation steps.

The first post-seroconversion blood sample collection date is the date when seroconversion is detected; that is, $O_{i1} = R_i$. The subsequent collection dates $\{O_{ij};\ j \in 2{:}\, n_i\}$ follow the GS Study protocol of visits at 4, 8, and 12 weeks after R_i and then at 12-week intervals, continuing until 10 years after enrollment in the pre-seroconversion phase of the study (F_i). For simplicity, scheduling offsets were not implemented for these dates; we did not see any reason why such offsets would meaningfully alter the results.

For each observation date O_{ij}, we calculated the corresponding time (in years) since seroconversion, $T_{ij} = \left(O_{ij} - S_i\right)/365$. We then simulated an MAA classification, Y_{ij}, from the Bernoulli distribution with success probability $\phi(t) = p(Y_{ij} = 1|T_{ij} = t) = (1 + \exp\{-(\theta_0 + \theta_1 t)\})^{-1}$, where $\theta_0 = 0.986$ is the log-odds of MAA-positive biomarker assay measurements on the date of seroconversion, and $\theta_1 = -3.88$ is the change per year since seroconversion in the log-odds of MAA-positive biomarkers; these parameters were estimated using midpoint imputation (for convenience) from the Clade C data set described in Section 2.4, consisting of 2,442 samples from the CAPRISA 002, FHI-360 GS, and HPTN 039-01 cohort studies (Laeyendecker et al. 2018) The corresponding mean window period μ is approximately 122.6 days:

$$\begin{aligned}
\mu &= \int_{t=0}^{\infty} \phi(t)dt \\
&= \int_{t=0}^{\infty} (1 + \exp\{-(\theta_0 + \theta_1 t)\})^{-1} dt \\
&= \left[\frac{\log\{\exp(\theta_0 + \theta_1 t) + 1\}}{\theta_1}\right]_{t=0}^{\infty}
\end{aligned}$$

$$= -\frac{\log\{\exp(\theta_0) + 1\}}{\theta_1}$$

$$\approx 122.6 \text{ days}$$

We used this target parameter value to assess our data analysis methods' accuracy.

For this simulation, the functional form of $\phi(t)$ was specified as a generalized linear model with a Bernoulli outcome distribution, a logistic link function, and with the log-odds linear in t; this form was chosen for simplicity and because it permitted a closed-form expression for μ. In practice, we would typically use a logistic model with a higher-order polynomial or spline function for the log odds, and we would integrate $\hat{\phi}(t)$ numerically to calculate $\hat{\mu}$. Using polynomials or splines for the linear component of the model allows the fitted curve $\hat{\phi}(t)$ to be flexible and data-adaptive. The use of a logistic link function is not crucial; probit or identity link functions could also be used. With a sufficiently flexible form for the linear component, any shape for $\hat{\phi}(t)$ is possible for any of these link functions.

5.5.2 Simulation analysis

We implemented the data-generating model, as well as the midpoint imputation, uniform imputation, and joint modeling analyses, in the R statistical computing environment, version 3.6.1, starting from code implementing the GEL approach (R Core Team 2019; Langohr and Gómez Melis 2014) We generated 1000 simulated data sets for each combination of cohort size $N_0 \in \{4500, 10^5\}$ participants, mean follow-up interval $\delta \in \{84, 365\}$ days, and hazard function $\lambda(t) \in \{0 + 0.5t, 0 + t, 0 + 2t, 1 + 0t, 1 + 0.5t, 10 + 0t, 10 + 0.5t\}$ events per person-year. With each data set, we performed midpoint imputation, uniform imputation with 100 imputed data sets, and our joint modeling analysis; all three analyses used the correct logistic functional form to model $\phi(t) = p(Y_{ij} = 1|T_{ij} = t)$. We stopped the EM

algorithm when $\Delta_{\ell}^{(k)} < 0.1$ and $\Delta_{\boldsymbol{\theta}}^{(k)} < 0.0001$. Each analysis produced an estimate $\hat{\mu}_i$ for each data set $i \in \{1{:}\,1000\}$. Accordingly, for each analysis we estimated bias and standard error by $\widehat{\mathrm{E}}[\hat{\mu} - \mu] = \bar{\hat{\mu}} - \mu$ and $\sqrt{\widehat{\mathrm{E}}[(\hat{\mu} - \mathrm{E}[\hat{\mu}])^2]} = \sqrt{(n-1)^{-1} \sum_{i=1}^{n} (\hat{\mu}_i - \bar{\hat{\mu}})^2}$, respectively, where $\bar{\hat{\mu}} = n^{-1} \sum_{i=1}^{n} \hat{\mu}_i$. Results are listed in Tables 5.1 and 5.2.

To apply our joint modeling approach, we needed to choose a spacing width γ for the grid of possible seroconversion dates, $\mathcal{S}$, in the seroconversion date model. The choice of γ determines the number of dates in $\mathcal{S}$, and equivalently the number of parameters in $\boldsymbol{\omega}$ which must be estimated; smaller values of γ require more parameters. For the scenarios with mean pre-seroconversion follow-up interval $\delta = 84$ days, we chose $\gamma = 1$ day (Table 5.1). For the scenarios with $\delta = 365$ days, the simulations took too long to run with $\gamma = 1$ day, given our computational resources and current software implementation, so we instead performed the analysis with $\gamma = 7$ days (Table 5.2). To determine whether different choices of γ affected bias and variance, we also performed the joint analysis with γ values of 7, 28, and 42 days for the scenarios with $\delta = 84$ days, and with γ values of 28 and 42 days for the scenarios with $\delta = 365$ days (Table 5.3).

Enrollment dates are routinely recorded in cohort studies. Unfortunately, in our particular data sets, we did not have enrollment dates available. As a possible solution, we considered assuming that all participants had enrolled prior to the start of the earliest censoring interval. To test this strategy using our simulation framework, we generated a modified version of each simulated data set, with the enrollment dates, e_i, overwritten to equal $\min_i l_i$. We then applied our joint modeling analysis to these modified data sets (Table

5.4); note that midpoint and uniform imputation are unaffected by these data modifications.

To demonstrate our proposed bootstrap approach to uncertainty quantification, we applied it to an example simulated data set from the scenario with $N_0 = 4500$ cohort participants, $\delta = 365$ days between pre-seroconversion follow-up visits, and hazard rate $\lambda(t) = 1 + 0.5t$ events per person-year.

5.6 Simulation results

Table 5.1 shows simulation results for scenarios with cohorts of $N_0 = 4500$ participants. Our joint modeling analysis consistently produced estimates with biases of less than 9 days off from the target value, $\mu = 122.6$ days, across all combinations of follow-up interval widths (δ) and hazard functions ($\alpha + \beta t$) that we tested. In contrast, midpoint imputation produced estimates with biases of up to 76 days off from the target value, and uniform imputation produced estimates with biases of up to 99 days off from the target value. Standard errors were comparable for all three methods in most scenarios; they were substantially larger for joint modeling in the scenarios with wide censoring intervals and fast hazard rates, but even in these scenarios, the increase in standard error was less than the reduction in bias, relative to the other methods.

The biases for midpoint imputation ranged from moderately positive (+12.0 days) to very negative (-75.7 days), whereas the biases for uniform imputation ranged from negligibly positive (+0.7 days) to very negative (-98.7 days). The cause of these biases is explored in detail in Section 5.7. In short: the distribution of the seroconversion date conditional on the enrollment date and censoring interval, $p(s_i|e_i, l_i, r_i)$, is right-skewed in the scenarios with $\alpha \in \{1,10\}$; these methods incorrectly assume that this distribution is

uniform or at least symmetric, and the violation of this assumption produces bias: these methods tend to overestimate the seroconversion date and hence underestimate the elapsed time from seroconversion until biomarker collection, resulting in an underestimate of μ. In contrast, for the scenarios with $\alpha = 0$, $p(s_i|e_i, l_i, r_i)$ can be left-skewed, right-skewed, or symmetric, depending on where the censoring interval, (l_i, r_i), is located relative to the peak of $p(s_i|e_i)$. In these scenarios, the bias introduced by uniform or midpoint imputation can be either positive or negative, depending on the hazard function's slope, β, and the mean follow-up interval width, δ.

Table 5.1: Simulation results: bias and standard error of estimates for μ, by method, in scenarios with cohort size N_0 = 4500.

Simulation parameters		*Bias of $\hat{\mu}$ (days)*			*Standard error of $\hat{\mu}$ (days)*		
δ: mean pre-seroconversion follow-up interval (days)	$\alpha + \beta t$: hazard rate (events / person-year)	Midpoint imputation	Uniform imputation	Joint modeling	Midpoint imputation	Uniform imputation	Joint modeling
84	$0 + 0.5t$	0.2	0.7	-0.3	4.4	4.4	4.5
	$0 + t$	-0.1	0.5	-0.1	4.3	4.3	4.3
	$0 + 2t$	-0.4	0.1	0.3	4.3	4.3	4.4
	$1 + 0t$	-0.5	0.1	0.2	4.3	4.3	4.4
	$1 + 0.5t$	-0.9	-0.4	0.1	4.4	4.4	4.4
	$10 + 0t$	-10.0	-9.5	-0.7	4.3	4.3	4.6
	$10 + 0.5t$	-10.2	-9.7	-0.6	4.3	4.3	4.6
365	$0 + 0.5t$	12.0	-19.3	-1.9	7.8	7.2	8.3
	$0 + t$	6.1	-25.4	-0.9	8.0	7.0	8.9
	$0 + 2t$	-5.7	-37.5	0.1	8.2	6.7	10.5
	$1 + 0t$	-2.4	-33.9	-0.7	8.1	6.9	9.7
	$1 + 0.5t$	-8.6	-40.0	-0.7	8.3	6.7	10.3
	$10 + 0t$	-75.1	-98.3	8.7	10.0	3.2	35.8
	$10 + 0.5t$	-75.7	-98.7	8.0	10.1	3.2	36.1

Table 5.2 shows simulation results for scenarios with cohorts of $N_0 = 100{,}000$ participants. The estimated biases in these scenarios are nearly identical to those in table 1, except that joint modeling no longer results in any substantial bias, even for $\delta = 365$ and $\alpha = 10$; it appears that the bias we observed for joint modeling when $N_0 = 4500$ was only a finite-sample phenomenon. The standard errors are mostly negligible at this cohort size for

all three methods, except again for joint modeling in the scenarios with $\delta = 365$ and $\alpha = 10$.

Table 5.2: Simulation results: bias and standard error of estimates for μ, by method, in scenarios with cohort size N_0 = 100,000.

Simulation parameters		*Bias of $\hat{\mu}$ (days)*			*Standard error of $\hat{\mu}$ (days)*		
δ: mean pre-seroconversion follow-up interval (days)	$\alpha + \beta t$: hazard rate (events / person-year)	Midpoint imputation	Uniform imputation	Joint modeling	Midpoint imputation	Uniform imputation	Joint modeling
84	$0 + 0.5t$	0.3	0.9	0.0	0.9	0.9	1.0
	$0 + t$	0.0	0.6	0.0	0.9	0.9	0.9
	$0 + 2t$	-0.6	0.0	0.0	0.9	0.9	0.9
	$1 + 0t$	-0.6	0.0	0.0	0.9	0.9	0.9
	$1 + 0.5t$	-0.9	-0.3	0.0	0.9	0.9	0.9
	$10 + 0t$	-10.1	-9.6	0.0	0.8	0.8	0.9
	$10 + 0.5t$	-10.4	-9.9	0.0	0.8	0.8	0.9
365	$0 + 0.5t$	12.0	-19.3	-0.2	1.7	1.5	1.7
	$0 + t$	6.0	-25.3	-0.2	1.6	1.5	1.8
	$0 + 2t$	-6.1	-37.5	-0.1	1.7	1.4	2.1
	$1 + 0t$	-3.2	-34.3	-0.4	1.7	1.4	2.0
	$1 + 0.5t$	-8.9	-40.0	-0.3	1.7	1.4	2.0
	$10 + 0t$	-77.8	-98.5	0.0	2.0	0.7	7.9
	$10 + 0.5t$	-78.3	-98.9	-0.2	2.0	0.7	8.0

Table 5.3 examines the effects of changing the joint modeling approach's seroconversion model grid width tuning parameter, γ, in the scenarios with cohort size $N_0 = 4500$ and hazard rate slope $\beta = 0.5$. Scenarios with $\beta = 0$ produced nearly identical results (not shown). We only observed substantial effects of γ on bias or standard error for the scenario with $\alpha = 10$ events per person year. In the scenario with $\delta = 365$ and $\alpha = 10$, there was a noticeable bias-variance trade-off: larger values of γ produced larger biases but smaller standard errors.

Table 5.4 examines the consequences of incorrectly assuming that all participants enrolled prior to the start of the earliest censoring interval. This assumption led to biases and standard errors approximately equal to those produced by uniform imputation in Table 1; clearly, it is an unsafe assumption.

Table 5.3: Simulation results: bias and standard error of joint modeling approach estimate for μ, by seroconversion model grid width, in scenarios with cohort size $N_0 = 4500$ and hazard rate slope $\beta = 0.5$.

δ: mean pre-seroconversion follow-up interval (days)	$\alpha + \beta t$: hazard rate (events / person-year)	γ: Seroconversion model grid width	Bias of $\hat{\mu}$ (days)	Standard error of $\hat{\mu}$ (days)
84	$0 + 0.5t$	1	-0.3	4.5
		7	-0.2	4.5
		28	-0.1	4.5
		42	0.0	4.5
	$1 + 0.5t$	1	0.1	4.4
		7	0.1	4.4
		28	-0.1	4.4
		42	-0.3	4.5
	$10 + 0.5t$	1	-0.6	4.6
		7	-0.9	4.6
		28	-2.0	4.6
		42	-3.3	4.6
365	$0 + 0.5t$	1	-1.9	8.3
		7	-1.6	8.3
		28	-1.0	8.3
		42	-0.3	8.3
	$1 + 0.5t$	1	-0.7	10.3
		7	-0.8	10.3
		28	-0.9	10.2
		42	-0.9	10.2
	$10 + 0.5t$	1	8.0	36.1
		7	2.3	34.2
		28	-10.2	29.9
		42	-16.7	27.7

Table 5.4: Simulation results: bias and standard error of joint modeling estimate for μ with incorrect enrollment dates, in scenarios with cohort size $N_0 = 4500$

δ: mean pre-seroconversion follow-up interval (days)	$\alpha + \beta t$: hazard rate (events / person-year)	Bias of $\hat{\mu}$ (days)	Standard error of $\hat{\mu}$ (days)
84	$0 + 0.5t$	-0.8	4.5
	$0 + t$	-1.0	4.3
	$0 + 2t$	-1.2	4.3
	$1 + 0t$	-1.5	4.4
	$1 + 0.5t$	-1.8	4.4
	$10 + 0t$	-10.7	4.3
	$10 + 0.5t$	-10.8	4.4
365	$0 + 0.5t$	-10.2	8.0
	$0 + t$	-18.0	8.2
	$0 + 2t$	-34.6	7.7
	$1 + 0t$	-26.6	9.4
	$1 + 0.5t$	-36.6	9.2
	$10 + 0t$	-98.2	3.3
	$10 + 0.5t$	-98.6	3.3

To illustrate the joint modeling approach, Figure 5.1 shows an example of an estimated cumulative distribution function for seroconversion date, given enrollment on the first day of the study (dotted line). We produced this estimate by applying our joint modeling analysis to a data set which we simulated using the data-generating model described in Section 5.5.1, for the following scenario: initial cohort study size $N_0 = 100{,}000$ participants, mean pre-seroconversion follow-up interval $\delta = 365$ days, true hazard rate at study start $\alpha = 1$ event per person-year, true hazard rate changing by $\beta = 0.5$ events per person-year^2, and joint model seroconversion grid spacing width $\gamma = 7$ days. The random number generator was initialized with seed = 1. We also show the true data-generating cumulative distribution function (solid line). In this figure, the estimated survival curve is very close to the true survival curve; due to the large sample size, there is very little variance left in the estimated curve.

Figure 5.1: Estimated cumulative distribution function and data-generating cumulative distribution function for seroconversion date, given enrollment on the first day of the study, for a simulated data set.

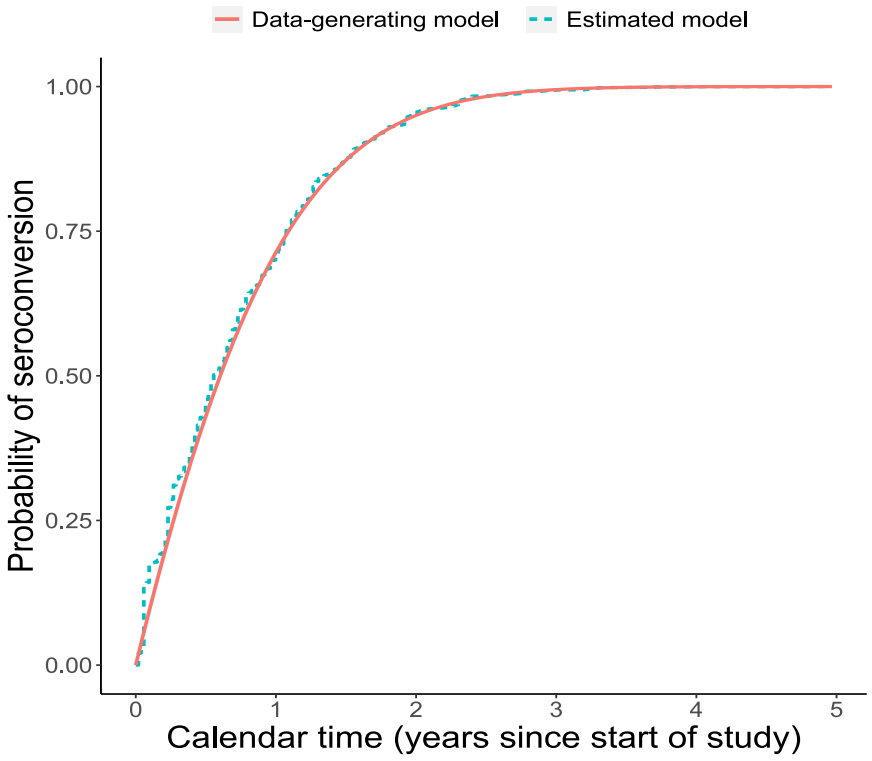
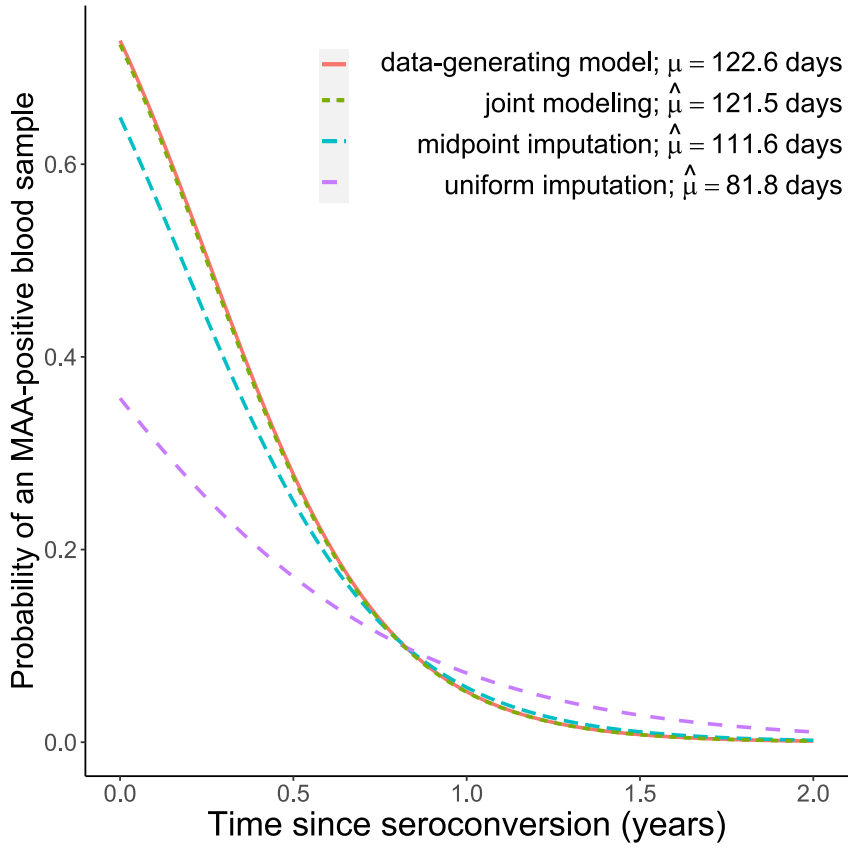


Figure 5.2 shows an example of $\hat{\phi}$ and $\hat{\mu}$ estimates produced by applying the midpoint imputation, uniform imputation, and joint modeling approaches to the same simulated data set used in Figure 5.1. There is very little sampling variance left at this sample size, so discrepancies between the data-generating model and the three estimated models can be attributed predominantly to bias. For this simulated data set, our joint modeling approach produced a $\hat{\phi}$ estimate nearly identical to the data-generating model's; the corresponding estimate $\hat{\mu} = 121.5$ days is only 1.1 days below the target value derived from the data-generating model, $\mu = 122.6$ days. In contrast, midpoint imputation moderately underestimated $\phi(t)$ for the first 9 months after seroconversion and underestimated μ by 11 days. Uniform imputation substantially underestimated $\phi(t)$ for the first 9 months, and moderately overestimated $\phi(t)$ for the next two years; the resulting $\hat{\mu} = 81.8$ days

underestimated μ by more than 40 days.

Figure 5.2: Estimated probability of MAA-positive biomarkers as a function of time since seroconversion, by method, for a simulated data set.



When we performed the bootstrap procedure from Section 5.3.3 on an example data set from the scenario with $N_0 = 4500$ cohort participants, $\delta = 365$ days between pre-seroconversion follow-up visits, and hazard rate $\lambda(t) = 1 + 0.5t$ events per person-year, the resulting bootstrap confidence interval was (117.3, 164.0), and the corresponding estimated standard error was 11.6, which is comparable to the estimate generated from the full set of simulations for this scenario, 10.3 (Table 1).

5.7 Distribution of seroconversion date, conditional on seroconversion window

We can understand why uniform imputation and midpoint imputation suffer from biases in some of our simulation scenarios by examining the distribution of the seroconversion date, conditional on the seroconversion censoring interval and enrollment date, $p(s_i|e_i, l_i, r_i)$.

In our simulation's data-generating model, the follow-up dates through the first seropositive test are independent of the actual seroconversion date, conditional on enrollment date; that is, Assumption 5 holds, and thus Eq. 5.1, derived in Section 5.2, also holds:

$$p(s_i|e_i, l_i, r_i) = \frac{1\{s_i \in [l_i, r_i]\}\, p(s_i|e_i)}{p(S_i \in [l_i, r_i]|e_i)} \propto 1\{s_i \in [l_i, r_i]\}\, p(s_i|e_i)$$

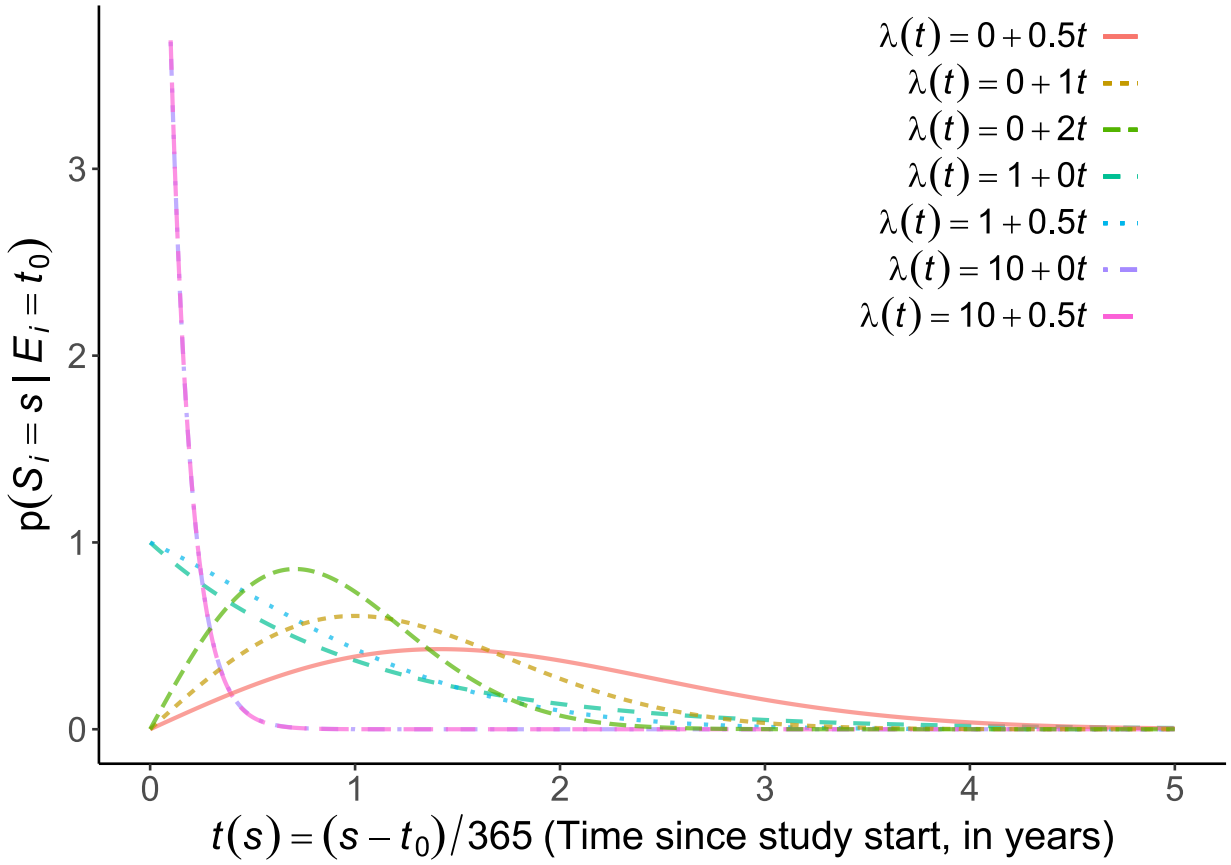
That is, the density of the seroconversion date, conditional on the seroconversion censoring interval, is proportional to the original seroconversion date density, truncated to the seroconversion censoring interval. This proportionality allows us to understand the conditional density by examining the unconditional density.

Consider the case of a participant who enrolls at the start of the study, i.e., $E_i = t_0$. We can derive an analytic expression for the unconditional density of the seroconversion date, using the relationships among the hazard, density, and survival functions:

$$\begin{aligned} p(S_i = s|E_i = t_0) &= \lambda\big(t(s)\big)P(S_i \geq s|E_i = t_0) \\ &= \lambda\big(t(s)\big)\exp\left\{-\int_{u=0}^{t(s)} \lambda(u)du\right\} \\ &= \big(\alpha + \beta t(s)\big)\exp\left(-\left[\alpha t(s) + \frac{\beta}{2}\{t(s)\}^2\right]\right) \end{aligned}$$

This density is shown in Figure 5.3 for each seroconversion hazard function that we considered. We can see that the four scenarios with $\alpha > 0$ all have monotonically decreasing densities for seroconversion date, whereas the three scenarios with $\alpha = 0$ have densities which increase to a maximum and then decrease.

Figure 5.3: Simulation data-generating probability density functions for seroconversion date, by hazard function, given enrollment at study start.



We can determine whether and when such a maximum will occur, as a function of α and β, by taking the derivative of the density and setting that derivative equal to 0:

$$\frac{d}{dt(s)}p(S_i = s|E_i = t_0) = \frac{d}{dt(s)}\left[\lambda(t(s))\exp\left\{-\int_{u=0}^{t(s)}\lambda(u)du\right\}\right]$$

$$= \exp\left\{-\int_{u=0}^{t(s)}\lambda(u)\,du\right\}\frac{d}{dt(s)}\lambda(t(s)) \;+\; \lambda(s)\frac{d}{dt(s)}\exp\left\{-\int_{u=0}^{t(s)}\lambda(u)\,du\right\}$$

$$= \exp\left\{-\int_{u=0}^{t(s)} \lambda(u)\, du\right\} \frac{d}{dt(s)} \lambda\big(t(s)\big)$$

$$+ \lambda\big(t(s)\big) \exp\left\{-\int_{u=0}^{t(s)} \lambda(u) du\right\} \frac{d}{dt(s)} \left(-\int_{u=0}^{t(s)} \lambda(u) du\right)$$

$$= \exp\left\{-\int_{u=0}^{t(s)} \lambda(u)\, du\right\} \frac{d}{dt(s)} \lambda\big(t(s)\big) + \lambda\big(t(s)\big) \exp\left\{-\int_{u=0}^{t(s)} \lambda(u) du\right\} \left(-\lambda\big(t(s)\big)\right)$$

$$= \exp\left\{-\int_{u=0}^{t(s)} \lambda(u)\, du\right\} \left[\left\{\frac{d}{dt(s)} \lambda\big(t(s)\big)\right\} - \left\{\lambda\big(t(s)\big)\right\}^2\right]$$

$$= p(S_i \geq s | E_i = t_0) \left[\beta - \big(\alpha + \beta t(s)\big)^2\right]$$

$$= p(S_i \geq s | E_i = t_0)(\beta - [\alpha^2 + 2\alpha\beta t(s) + \beta^2 \{t(s)\}^2])$$

$$= p(S_i \geq s | E_i = t_0)[-\beta^2 \{t(s)\}^2 - 2\alpha\beta t(s) + (\beta - \alpha^2)]$$

The first factor is strictly positive and can be ignored. The second term is quadratic in s, so by the quadratic formula, its roots are:

$$t(s) = \frac{2\alpha\beta \pm \sqrt{4\alpha^2\beta^2 + 4\beta^2(\beta - \alpha^2)}}{-2\beta^2} = \frac{2\alpha\beta \pm 2\beta\sqrt{\alpha^2 + (\beta - \alpha^2)}}{-2\beta^2} = \frac{-\alpha \pm \sqrt{\beta}}{\beta}$$

We know that α must be nonnegative (since it is the hazard rate at the start of the study). Since we are considering a hazard that begins at $t(s) = 0$, we are only interested in positive roots. So, the density will have a maximum at $t(s) = \big(\sqrt{\beta} - a\big)/\beta$ if $\alpha < \sqrt{\beta}$; i.e., if $\beta > \alpha^2$.

From Figure 5.3, we can see that with a linear hazard function, the conditional density of S within a censoring interval can be left-skewed, right-skewed, or symmetric, depending on α, β, and the position of the censoring interval. However, even for scenarios in which the density is clearly right-skewed and monotonically decreasing, the conditional density may be approximately uniform, if the width of the censoring interval is sufficiently narrow in comparison with the slope of the density; for example, Figure 5.4 shows the same seven

densities, zoomed in on the first 84 days after study start; as shown above, these curves are proportional to the seroconversion densities conditional on a censoring interval [0, 84]. Given an interval of this width, the scenarios with $\alpha = 1$ have essentially uniform densities for the seroconversion date, conditional on the seroconversion censoring interval, whereas the scenarios with $\alpha = 0$ or $\alpha = 10$ still have substantially non-symmetric densities (left-skewed and right-skewed, respectively) for censoring intervals consisting of the first 84 days.

Figure 5.4: Simulation data-generating probability density functions for seroconversion date, by hazard function, given enrollment at study start, for the first 84 days after study start.

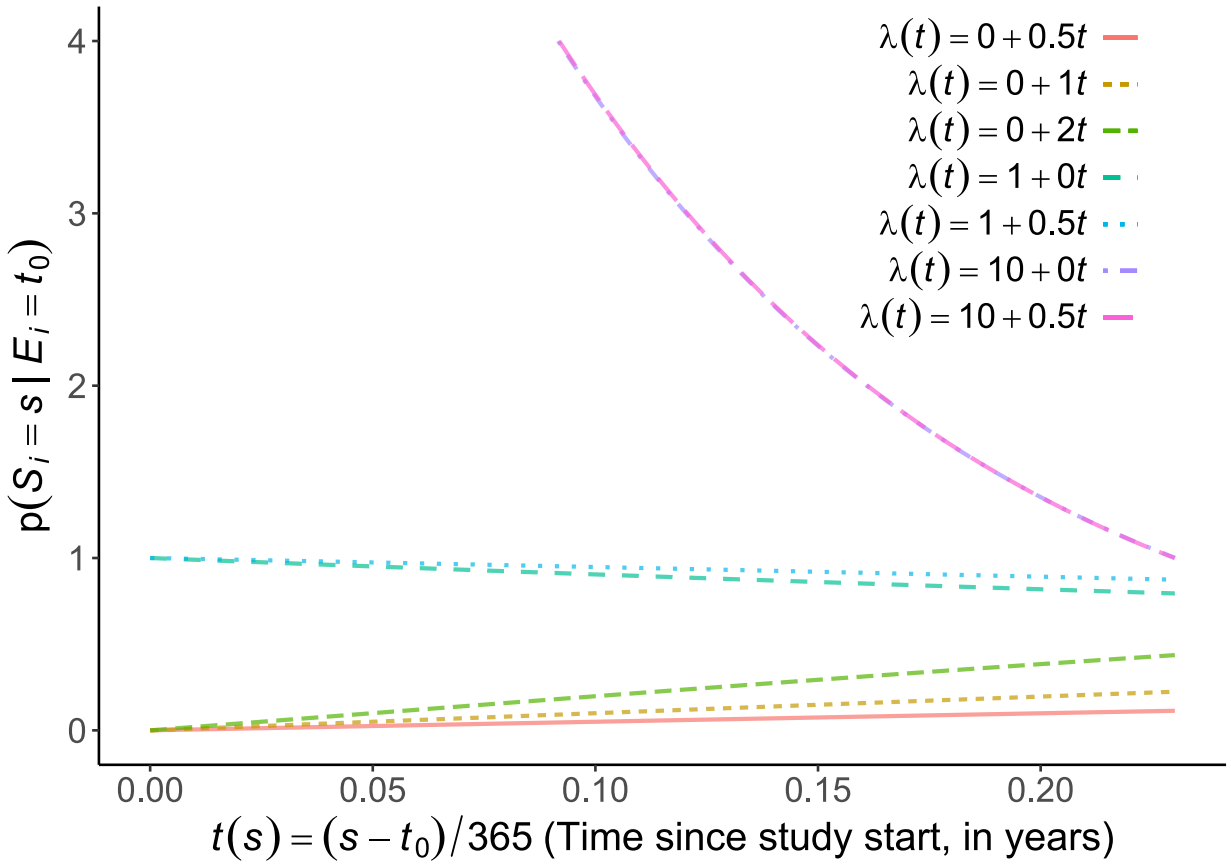
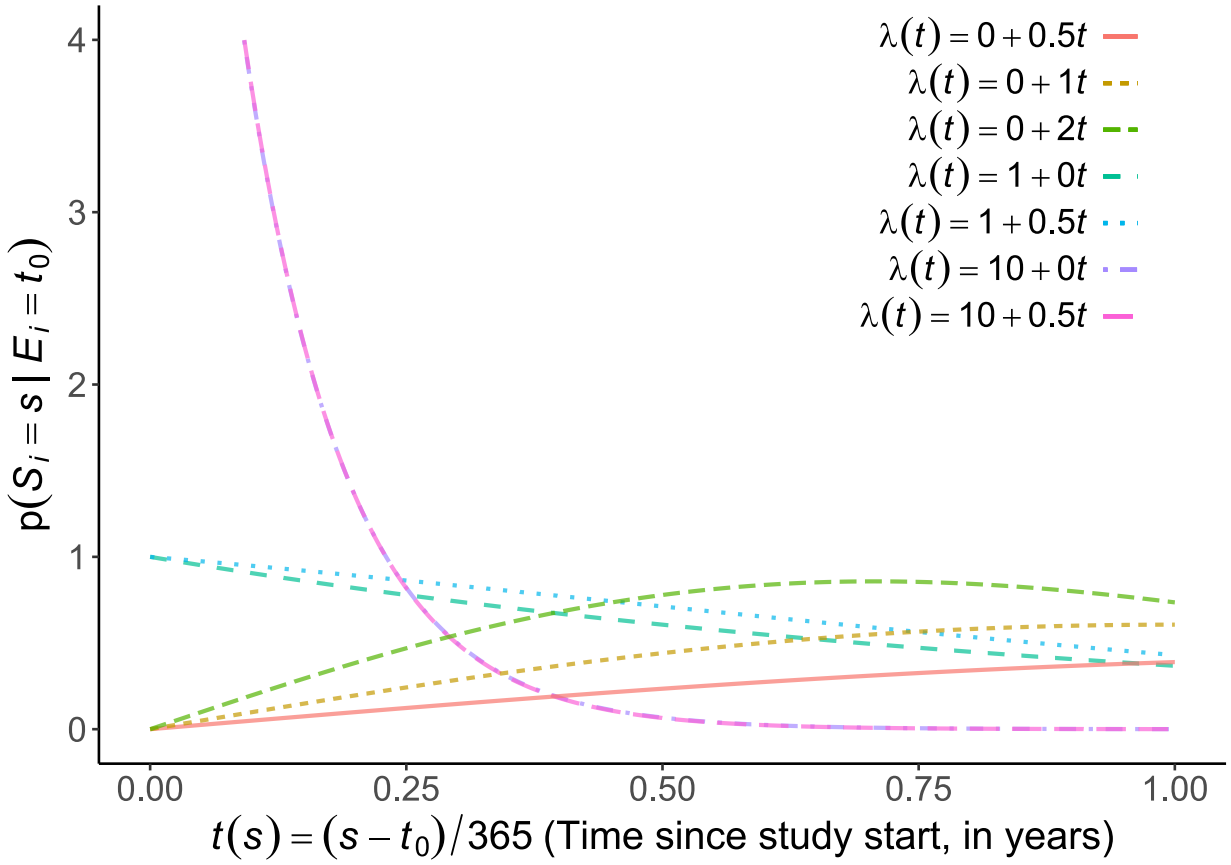


Figure 5.5 shows the same seven densities, zoomed in on the first 365 days after study start; we can see that given a censoring interval of this size and position, the scenarios with

$\alpha = 1$ now have slightly right-skewed densities for the seroconversion date, and the density corresponding to $\lambda(t) = 0 + 2t$ is now nonmonotone but still left-skewed.

Figure 5.5: Simulation data-generating probability density functions for seroconversion date, by hazard function, given enrollment at study start, for the first 365 days after study start.



These results indicate why the midpoint and uniform imputation approaches had more bias in the scenarios with $\alpha = 10$ or $\delta = 365$ (δ is the mean pre-seroconversion follow-up interval length); these approaches assume a uniform or at least symmetric density for the seroconversion date within the seroconversion censoring interval, which is approximately correct if that interval is sufficiently narrow relative to the hazard rate but can be substantially incorrect otherwise. When the distribution is heavily right-skewed, these methods will tend to overestimate the seroconversion date and thus underestimate the

duration of infection at the date of sample collection, resulting in a negative bias in the resulting estimates of μ. When the slope of the density is nonmonotone, the bias from these methods could go in either direction, depending on the exact hazard function and the censoring interval widths, as seen in our simulation results (Tables 5.1-5.2).

5.8 Difficulties in calculating μ for outcome models with autocorrelation

The third assumption in our analysis is that longitudinally repeated MAA classifications of the same individual are mutually independent, conditional on the duration of infection at the time of sample collection; i.e. $p(\boldsymbol{y}_i|\boldsymbol{t}_i) = \prod_{j\in 1:n_i} p(y_{ij}|\, t_{ij})$. In practice, longitudinal biomarker observations may exhibit substantial within-individual correlation; however, it is hoped that by using an appropriate functional form for the relationship between time and MAA classification, any such autocorrelation can be removed. This assumption is not necessary for the joint modeling approach in general; it is necessary for our motivating application, regardless of whether joint modeling, midpoint imputation, or uniform imputation is used, because it enables us to identify the marginal distribution $\phi(t) = p(Y = 1|\, T = t)$ with $p(y_{ij}|\, t_{ij})$, which is then used to compute μ as described above.

To see how models with autocorrelation pose difficulties for our motivating application, consider the example of adding a random effect on the regression intercept. Our joint modeling approach can be straightforwardly extended to such a scenario, by replacing the generalized additive model $p(y_{ij}|\, t_{ij})$ with $p(y_{ij}|\, t_{ij}, u_i)\, p(u_i)$, where u_i is the individual-specific random effect. We could fit this model by maximum likelihood if seroconversion date S_i were directly observable; hence we can fit it using joint modeling and the EM algorithm: in the M step, the estimates of the fixed effects and variance components are updated via

maximum likelihood, and in the E step, the marginal likelihood based on these estimates, $\hat{p}(y_{ij}|\, t_{ij}) = \mathrm{E}[\hat{p}(y_{ij}|\, t_{ij}, U_i)|t_{ij}]$, is used to update $p_{\hat{\boldsymbol{\Psi}}}(s_i|e_i, l_i, r_i, \boldsymbol{o}_i, \boldsymbol{y}_i)$. We could also fit this model using midpoint imputation or uniform imputation. Regardless of which estimation approach we use, if $p(y_{ij}|t_{ij}, u_{\boldsymbol{i}})$ has a nonlinear link function, then the marginal distribution $\phi(t) = p(Y_{ij} = 1|T_{ij} = t) = E[p(Y_{ij} = 1|T_{ij} = t, U_{\boldsymbol{i}})|T_{ij} = t]$ is no longer equivalent with the value predicted by the fixed effects. For example, given the logistic mixed-effects model $p\big(Y_{ij} = 1\big|T_{ij}, U_i\big) = \mathrm{expit}\big\{\theta_0 + \theta_1 T_{ij} + U_i\big\}$, $U_i \underset{iid}{\sim} N(0, \sigma_u^2)$, $U_i \perp\!\!\!\perp \boldsymbol{T}_i$, we find:

$$E\big[p\big(Y_{ij} = 1\big|T_{ij} = t, U_i\big)|T_{ij} = t\big] = \int_{u_i \in \mathbb{R}} \mathrm{expit}\{\theta_0 + \theta_1 t + u_i\}\, p(U_i = u_i)\, du_i$$

$$\neq \mathrm{expit}\{\theta_0 + \theta_1 t\}$$

Future work could involve solving this problem; generalized estimating equations may be helpful.

5.9 Discussion

The simulation results for midpoint imputation and uniform imputation showed the potential for substantially underestimating or overestimating μ; the biases from these methods were nearly identical between the smaller and larger sample sizes, indicating that in some scenarios, these methods are asymptotically inconsistent. In contrast, joint modeling produced asymptotically consistent estimates in all scenarios considered, as long as accurate study enrollment dates were available for analysis. When participants were incorrectly assumed to have all enrolled prior to the first censoring interval, our method no longer produced accurate estimates and instead performed similarly to uniform imputation.

In the scenarios with mean pre-seroconversion follow-up interval $\delta = 84$ days, midpoint

imputation and uniform imputation resulted in similar amounts of bias. In the scenarios with $\delta = 365$ days, uniform imputation resulted in substantially more severe biases than midpoint imputation. It may seem surprising that uniform imputation led to more bias than midpoint imputation, since the expected value of a uniform variable is the interval midpoint. However, it should be noted that both the resulting model estimate $\hat{\phi}(t)$ and mean window period estimate $\hat{\mu} = \int_{t=0}^{\infty} \hat{\phi}(t)dt$ are nonlinear functions of the imputed seroconversion dates; furthermore, uniform imputation involves averaging the estimated parameters of $\hat{\phi}(t)$ across the multiply-imputed data sets, on the log-odds scale. Future work could include further study of these biases. Additionally, midpoint imputation resulted in larger standard error than uniform imputation in many scenarios; these differences could also be further investigated.

In this analysis, we assumed that the follow-up dates through the first seropositive test are independent of the actual seroconversion date, conditional on enrollment date. In practice, depending on the study protocols, the follow-up dates might deviate from the planned schedules if study participants can request an earlier test date when they feel sick or believe they may have been recently exposed – for example, after high-risk behaviors. In such cases, our assumption of independence between the follow-up dates and the actual seroconversion date would be invalid, and the joint modeling analysis presented in this chapter might produce biased estimates. To handle such a scenario, the model used in the analysis would need to be changed accordingly; such an extension would be worthwhile for future work.

In our simulation study, we assumed that the MAA classification model $\hat{\phi}(t)$ used in the

analysis was correctly specified to match the functional form of the data-generating model. In practice, the correct functional form would be unknown and model fitting would be required. The model fitting process would be complicated by the fact that the covariate t_{ij} is not known precisely, making graphical approaches to regression modeling diagnostics more challenging to apply. Exploration of best practices for model fitting in this setting and evaluation of the consequences of mis-specification would also be worthwhile for future work.

Future work could also include combining the joint modeling approach for interval-censored seroconversion dates with a survival analysis, mixed-effects modeling, or functional data analysis approach for modeling MAA classifications, and the effects of model mis-specification relative to the data-generating process could be quantified. For example, if the onset date of the MAA-negative state were precisely observable, then the joint model and EM estimation procedure proposed here could be straightforwardly combined with a time-to-event model for onset of MAA-negativity by redefining Y_i as the MAA-negative onset date; then a time-to-event model for $p(y_i|s_i)$ would be substituted in place of $p(\boldsymbol{y}_i|\boldsymbol{t}_i)$ in the likelihood decomposition, and the EM algorithm would otherwise remain the same. However, in practice the MAA-negative onset dates would also be interval-censored, which could further complicate the analysis. A Markov model approach allowing repeated transitions between the MAA-positive and MAA-negative states might also be of interest.

In both the analysis model and the simulation data-generating model described in this analysis, the seroconversion date hazard function was assumed not to vary within the subset of individuals at any risk of infection. In our implementation, we have further extended the

seroconversion date model to accommodate stratification by baseline characteristics. For example, if participants were recruited from several clinics serving different populations, we might estimate clinic-specific hazard functions. Furthermore, if data has been combined from multiple cohort studies, we might stratify by cohort. Future work could include modeling unobserved or time-varying risk factors. We also assumed that there was no possibility of drop-out prior to the protocol-defined study exit date; unmodeled associations between infection risk and drop-out risk could lead to bias. Future work could explore such effects.

We may be able to improve precision for small samples if we can assume a simple parametric model for the hazard function, such as a low-order polynomial with random effects by participant. Such an assumption could substantially reduce the number of parameters that need to be estimated for the seroconversion date model, relative to our current non-parametric approach.

While our analysis was motivated by and tailored to a specific example, the joint modeling approach is more general. The interval-censored covariate need not be a function of a time-to-event variable; other forms of inexact measurement also result in interval-censoring. Joint modeling could be useful in those contexts as well.

CHAPTER 6
Conclusion

6.1 Challenges Addressed

In this dissertation, we addressed three statistical challenges which are frequently encountered when using cross-sectional biomarker surveys to estimate infectious disease incidence: analyzing data sets with incomplete biomarker data, transporting MAA calibration estimates to new populations and epidemiological conditions, and accounting for interval-censored seroconversion dates in calibration data sets. These challenges all required inferences about indirectly-observed probability distributions.

In Chapter 3, we considered a data set in which one of two biomarkers used in an MAA was incompletely assayed, resulting in missing MAA classifications. We needed to extrapolate from the observed incomplete data distribution to the unobserved distribution which would have been observed if every sample had all been assayed for both biomarkers. We assumed that the incompletely-assayed biomarker's missingness status was independent of its underlying distribution, conditional on the other, completely assayed biomarker indicating a recent infection. This assumption motivated a hierarchical model which produced accurate estimates of the mean window period and incidence rate in simulation scenarios. In contrast, a single model fit using all the observations that could be classified produced biased estimates, whenever there was substantial missingness in biomarker. Single models fit using the subset of samples for which all biomarkers were assayed also produced biased results when the probability of assaying the second biomarker depended on the value of the first biomarker.

In Chapter 4, we needed to extrapolate from an MAA calibration data set to a target

population with different patterns of viral suppression conditional on duration of infection, resulting in a different mean window period. Here, we presented several estimation approaches borrowed from the causal inference literature. The "curve averaging" approach modeled MAA classifications conditional on viral suppression and duration of infection using the calibration data set, and then marginalized this model using the target population's distribution of viral suppression. The "sample weighting" approach constructed weights to account for the differences in viral suppression between the calibration data set and target population and then used these weights to analyze the calibration data, either by weighted maximum likelihood analysis or by weighted resampling. The "multivariate modeling and marginalization" and "potential outcomes modeling" approaches modeled the multivariate distribution of the biomarker assay values conditional on viral suppression and duration of infection, using the calibration data set; the MMM approach then marginalized this model, whereas the potential outcomes approaches used this model to estimate the counterfactual biomarker values that would have been observed for each observation under opposite viral suppression conditions. The potential outcomes approaches then modeled the corresponding MAA classifications conditional on duration of infection using either weighted likelihood or weighted sampling.

The proposed approaches in Chapter 4 all relied on a number of assumptions – most notably, that the distribution of viral suppression in the target population is known or estimable, and that the distribution of biomarker values, conditional on viral suppression and duration of infection, is equivalent between the calibration data set and the target population. The MMM and potential outcomes modeling approaches additionally required assuming a functional form for the conditional distribution of biomarker values, whereas the

other two approaches only assumed a functional form for the conditional distribution of MAA classifications. All of these approaches produced mean window period estimates with minimal bias when their assumptions were valid, but the first two approaches resulted in substantially larger standard errors than the multivariate modeling approaches. On the other hand, when the additional assumptions of the multivariate modeling approaches were violated, its performance suffered accordingly. Hence, the choice of analysis should depend on whether the added assumptions of the multivariate modeling approaches are defensible for a particular analysis.

In Chapter 5, we considered the question of how best to handle interval-censored seroconversion dates in calibration data sets. Here, we needed to extrapolate from the observed distribution of seroconversion censoring intervals to the unobserved distribution of seroconversion dates. To do so, we made a number of assumptions, most notably that the dates of follow-up visits prior to diagnosis are independent of the actual seroconversion date. These assumptions led us to a joint modeling approach based on the EM algorithm which produced accurate estimates of the mean window period in simulation scenarios. In contrast, an analysis approach using the seroconversion censoring interval midpoint as an estimate of the seroconversion date produced substantially biased estimates in scenarios with wide censoring intervals. An approach using uniform imputation over the censoring interval also performed poorly in these scenarios. However, in the scenarios with relatively narrow censoring intervals and lower hazard rates, all three approaches performed similarly; in such cases, the added computational requirements of the joint modeling approach would be unnecessary.

6.2 Future Work

These methods all have avenues for further development. As discussed in Chapter 4, the missing data analysis could be extended to consider more complex missingness mechanisms, possibly involving incomplete measurements in more than one biomarker. The transportability analyses in Chapter 5 could be extended to accommodate more complex differences between the calibration data set and the target population, possibly involving a vector of mediating covariates whose relationships with infection duration have changed. The joint modeling analysis of interval-censored seroconversion dates in Chapter 6 should be extended to include additional covariates in the outcome sub-model, so that it can be combined with the methods in Chapter 5. Furthermore, the EM algorithm for the joint modeling approach is computationally intensive, especially when combined with bootstrapping to produce uncertainty estimates. A more efficient implementation of the algorithm, or an alternative approach to quantifying uncertainty, would be valuable. Similarly, a non-saturated model for the distribution of seroconversion dates might reduce computation time and improve precision.

6.3 Closing Thoughts

The methods presented in this dissertation were all motivated by the application of calibrating and performing HIV incidence estimation using cross-sectional surveys of biomarker prevalence; however, these methods are more generally applicable. Incomplete data, extrapolation to new populations, and interval-censored covariates are frequently-encountered challenges in statistical analysis.

Moreover, the cross-sectional survey-based approach to incidence estimation has

applicability for other diseases than HIV. The key aspect of HIV infection that makes incidence estimation particularly challenging is its frequently lengthy pre-symptomatic period, which means that cases may be several years old by the time they are diagnosed; hence the rate of new diagnoses in a population constitutes a lagged and temporally blurred indicator of the incidence rate. Other diseases with long latent periods may also benefit from this approach to incidence estimation. Even for diseases with shorter latent periods, this approach may be useful when it is crucial to rapidly detect changes in incidence. For example, during the coronavirus-19 (COVID-19) pandemic, this approach could be deployed by identifying individuals who are COVID-positive according to PCR or antigen testing but not yet symptomatic and using the prevalence of these individuals in weekly or daily cross-sectional snapshots to estimate incidence. We hope that the methods presented in this dissertation will encourage the use of the cross-sectional approach to incidence estimation in a variety of contexts and will help address the inevitable real-world complications in the data collection process.

Appendix: Consolidated Notation List

κ	Calibration population
λ	Biomarker missingness probability
μ	Mean duration of MAA-positive infection ("mean window period")
τ	Target population
ψ	Shadow of MAA
$\phi(t)$	Probability of MAA-positive infection, conditional on time: $P(Y = 1 \mid T = t)$
B, B_1, B_2	Biomarker variables
b	Indicator variable for a biomarker being inside its cutoff for recent classification (1 = recent, 0 = not recent).
c, c_1, c_2	Biomarker cutoff values for "recent" status
E	Study enrollment date
$h(s)$	Incidence rate at time s, $p(S = s \mid S \geq s)$
L	Left endpoint of censoring interval for seroconversion date S
m	Indicator of biomarker missingness (1: missing, 0: observed)
N_u	Number of uninfected individuals in a cross-sectional biomarker survey
N_x	Number of infected individuals in a cross-sectional biomarker survey
N	Number of individuals in a data set
n_i	Number of biomarker samples collected for individual i
$\boldsymbol{O}$	Vector of post-seroconversion observation dates when biomarker samples were collected

$P(A)$	The probability mass of event A
$p(A)$	The probability density of event A
R	Right endpoint of censoring interval for seroconversion date S
S	Seroconversion date (the date when an individual would first be diagnosed with the condition of interest if tested).
T	Elapsed time since seroconversion ("duration of infection")
t_0	The calendar date at which a cross-sectional survey is performed.
$t_{\max}$	Time point after seroconversion beyond which we assume that $\phi(t) \approx 0$
V	Number of MAA-positive individuals in a cross-sectional biomarker survey
$W(s)$	The population in which an individual is living at calendar time s.
X	In Chapter 2: Seroconversion status (1 = seropositive, 0 = seronegative) In Chapter 4: Covariate(s) mediating the relationship between infection duration and biomarker distribution, e.g., anti-retroviral usage
Y	Multi-assay algorithm (MAA) recency classification: 1 = "recent", 0 = "non-recent"
Z	Variables mediating the relationship between infection duration (T) and X

References

Abdool Karim, Q., S.S. Abdool Karim, J.A. Frohlich, A.C. Grobler, C. Baxter, L.E. Mansoor, A.B.M. Kharsany, S. Sibeko, K.P. Mlisana, Z. Omar, T.N. Gengiah, S. Maarschalk, N. Arulappan, M. Mlotshwa, L. Morris, and D. Taylor. 2010. "Effectiveness and Safety of Tenofivir Gel, an Antiretroviral Microbicide, for the Prevention of HIV in Women." *Science* 329(5996): 1168–74.

Birnbaum, A. 1962. "On the Foundations of Statistical Inference." *Journal of the American Statistical Association* 57(298): 269–306.

Bolker, B., and R Core Team. 2020. "Bbmle: Tools for General Maximum Likelihood Estimation."

Brookmeyer, R. 1997. "Accounting for Follow-up Bias in Estimation of Human Immunodeficiency Virus Incidence Rates." *Journal of the Royal Statistical Society: Series A* 160(1). Wiley Online Library: 127–40.

———. 2010. "On the Statistical Accuracy of Biomarker Assays for HIV Incidence." *Journal of Acquired Immune Deficiency Syndrome* 54(4): 406–14.

Brookmeyer, R., J. Konikoff, O. Laeyendecker, and S.H. Eshleman. 2013. "Estimation of HIV Incidence Using Multiple Biomarkers." *American Journal of Epidemiology* 177(3): 264–72.

Brookmeyer, R., O. Laeyendecker, D. Donnell, and S.H. Eshleman. 2013. "Cross-Sectional HIV Incidence Estimation in HIV Prevention Research." *Journal of Acquired Immune Deficiency Syndrome* 63(0 2): 1–13.

Brookmeyer, R., and T. Quinn. 1995. “Estimation of Current Human Immunodeficiency Virus Incidence Rates from a Cross-Sectional Survey Using Early Diagnostic Tests.” *American Journal of Epidemiology* 141(2): 166–72.

Busch, M.P., C.D. Pilcher, T.D. Mastro, J. Kaldor, G. Vercauteren, W. Rodriguez, C. Rousseau, T.M. Rehle, A. Welte, M.D. Averill, and J.M. Garcia Calleja. 2010. “Beyond Detuning: 10 Years of Progress and New Challenges in the Development and Application of Assays for HIV Incidence Estimation.” *AIDS* 24(18): 2763–71.

Celum, C., and A. Wald. 2004. “HPTN 039-01-Ancillary Prospective Cohort Study of HPTN 039 Seroconverters: The Effect of HSV-2 Suppression on HIV-1 Viral Set Point.”

Coates, T.J., M. Kulich, D.D. Celentano, C.E. Zelaya, S. Chariyalertsak, A. Chingono, G. Gray, J.K.K. Mbwambo, S.F. Morin, L. Richter, M. Sweat, H. van Rooyen, N. McGrath, A. Fiamma, O. Laeyendecker, E. Piwowar-Manning, G. Szekeres, D. Donnell, and S.H. Eshleman. 2014. “Effect of Community-Based Voluntary Counselling and Testing on HIV Incidence and Social and Behavioural Outcomes (NIMH Project Accept; HPTN 043): A Cluster-Randomised Trial.” *The Lancet Global Health* 2(5): 267–77.

Cole, S.R., H. Chu, and R. Brookmeyer. 2007. “Confidence Intervals for Biomarker-Based Human Immunodeficiency Virus Incidence Estimates and Differences Using Prevalent Data.” *American Journal of Epidemiology* 165(1): 94–100.

Cole, S.R., and C.E. Frangakis. 2009. “The Consistency Statement in Causal Inference: A Definition or an Assumption?” *Epidemiology* 20(1): 3–5.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. “Maximum Likelihood from Incomplete

Data Via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1): 1–38.

Dobson, A.J., and A.G. Barnett. 2008. *An Introduction to Generalized Linear Models*. 3rd Editio. Boca Raton, FL: Chapman and Hall/CRC.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7(1): 1–26.

Freeman, J., and G.B. Hutchison. 1980. "Prevalence, Incidence and Duration." *American Journal of Epidemiology* 112(5): 707–23.

Garrett, N.J., L. Werner, N. Naicker, V. Naranbhai, S. Sibeko, N. Samsunder, C. Gray, C. Williamson, L. Morris, Q. Abdool Karim, and S.S. Abdool Karim. 2015. "HIV Disease Progression in Seroconvertors from the CAPRISA 004 Tenofovir Gel Pre-Exposure Prophylaxis Trial." *Journal of Acquired Immune Deficiency Syndrome* 68(1): 55–61.

Gelman, A. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22(2): 153–64.

Goggins, W.B., D.M. Finkelstein, and A.M. Zaslavsky. 1999. "Applying the Cox Proportional Hazards Model When the Change Time of a Binary Time-Varying Covariate Is Interval Censored." *Biometrics* 55(2): 445–51.

Gómez, G., A. Espinal, and S.W. Lagakos. 2003. "Inference for a Linear Regression Model with an Interval-Censored Covariate." *Statistics in Medicine* 22(3): 409–25.

Gordis, L. 2014. *Epidemiology*. 5th ed. Philadelphia, PA: Elsevier.

Greenland, S., and J.M. Robins. 1986. "Identifiability, Exchangeability, and Epidemiological

Confounding." *International Journal of Epidemiology* 15(3): 413–19.

Hallett, T.B., P. Ghys, T. Bärnighausen, P. Yan, and G.P. Garnett. 2009. "Errors in 'BED'-Derived Estimates of HIV Incidence Will Vary by Place, Time and Age." *PLoS ONE* 4: 5720.

Hanson, D.L., R. Song, S. Masciotra, A. Hernandez, T.L. Dobbs, B.S. Parekh, S.M. Owen, and T.A. Green. 2016. "Mean Recency Period for Estimation of HIV-1 Incidence with the BED-Capture EIA and Bio-Rad Avidity in Persons Diagnosed in the United States with Subtype B." *PLoS ONE* 11(4): 1–9.

Hastie, T.J., and R.J. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall.

Hemelaar, J., E. Gouws, P.D. Ghys, and S. Osmanov. 2011. "Global Trends in Molecular Epidemiology of HIV-1 during 2000-2007." *AIDS* 25: 679–89.

Hernán, M.A. 2012. "Beyond Exchangeability: The Other Conditions for Causal Inference in Medical Research." *Statistical Methods in Medical Research* 21(1): 3–5.

Hernán, M.A., and J.M. Robins. 2019. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC, forthcoming.

Hsiao, C. 1983. "Regression Analysis with a Categorized Explanatory Variable." In *Studies in Econometrics, Time Series, and Multivariate Statistics*, edited by Samuel Karlin, Takeshi Amemiya, and Leo A. Goodman, 93–130. San Diego, CA: Academic Press.

Kaplan, E.H., and R. Brookmeyer. 1999. "Snapshot Estimators of Recent HIV Incidence Rates." *Operations Research* 47(1): 29–37.

Kaplan, E.L., and P. Meier. 1958. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53(282): 457–81.

Kassanjee, R., C.D. Pilcher, S.M. Keating, S.N. Facente, E. McKinney, M.A. Price, J.N. Martin, S. Little, F.M. Hecht, E.G. Kallas, A. Welte, M.P. Busch, and G. Murphy. 2014. "Independent Assessment of Candidate HIV Incidence Assays on Specimens in the CEPHIA Repository." *AIDS* 28: 2439–49.

Konikoff, J. 2015. Cross-Sectional HIV Incidence Estimation: Techniques and Challenges. Ph.D. Dissertation. University of California at Los Angeles, Los Angeles, CA.

Konikoff, J., R. Brookmeyer, A.F. Longosz, M.M. Cousins, C. Celum, S.P. Buchbinder, G.R. Seage, G.D. Kirk, R.D. Moore, S.H. Mehta, J.B. Margolick, J. Brown, K.H. Mayer, B.A. Koblin, J.E. Justman, S.L. Hodder, T.C. Quinn, S.H. Eshleman, and O. Laeyendecker. 2013. "Performance of a Limiting-Antigen Avidity Enzyme Immunoassay for Cross-Sectional Estimation of HIV Incidence in the United States." *PLoS ONE* 8(12): 1–9.

Laeyendecker, O., R. Brookmeyer, M.M. Cousins, C.E. Mullis, J. Konikoff, D. Donnell, C. Celum, S.P. Buchbinder, G.R. Seage, G.D. Kirk, S.H. Mehta, J. Astemborski, L.P. Jacobson, J.B. Margolick, J. Brown, T.C. Quinn, and S.H. Eshleman. 2012. "HIV Incidence Determination in the United States: A Multiassay Approach." *Journal of Infectious Diseases* 207(2): 232–39.

Laeyendecker, O., J. Konikoff, D.E. Morrison, R. Brookmeyer, J. Wang, C. Celum, C.S. Morrison, Q. Abdool Karim, A.E. Pettifor, and S.H. Eshleman. 2018. "Identification and Validation of a Multi-Assay Algorithm for Cross-Sectional HIV Incidence Estimation in Populations with Subtype C Infection." *Journal of the International AIDS Society* 21(2): 1–7.

Laeyendecker, O., A.D. Redd, M. Nason, A.F. Longosz, Q.A. Karim, V. Naranbhai, N. Garrett, S.H. Eshleman, S.S. Abdool Karim, and T.C. Quinn. 2015. "Antibody Maturation in Women

Who Acquire HIV Infection While Using Antiretroviral Preexposure Prophylaxis." *Journal of Infectious Diseases* 212(5): 754–59.

Lagakos, S.W., and A.R. Gable. 2008. "Challenges to HIV Prevention — Seeking Effective Measures in the Absence of a Vaccine." *New England Journal of Medicine* 358(15): 1543–45.

Lange, K. 2010. *Numeric Analysis for Statisticians*. New York: Springer.

Langohr, K., and G. Gómez Melis. 2014. "Estimation and Residual Analysis with R for a Linear Regression Model with an Interval-Censored Covariate." *Biometrical Journal* 56(5): 867–85.

Longosz, A.F., C.S. Morrison, P.-L. Chen, H.H. Brand, E. Arts, I. Nankya, R.A. Salata, T.C. Quinn, S.H. Eshleman, and O. Laeyendecker. 2015. "Comparison of Antibody Responses to HIV Infection in Ugandan Women Infected with HIV Subtypes A and D." *AIDS Research and Human Retroviruses* 31: 421–27.

Lumley, T. 2010. *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ: John Wiley & Sons, Inc.

———. 2013. "Biglm: Bounded Memory Linear and Generalized Linear Models."

Mastro, T.D. 2013. "Determining HIV Incidence in Populations: Moving in the Right Direction." *Journal of Infectious Diseases* 207(2): 204–6.

McLachlan, G.J., and T. Krishnan. 2007. *The EM Algorithm and Extensions: Second Edition*. Hoboken, NJ: John Wiley & Sons.

Morrison, C.S., P.L. Chen, I. Nankya, A. Rinaldi, B. Van Der Pol, Y.R. Ma, T. Chipato, R. Mugerwa,

M. Dunbar, E. Arts, and R.A. Salata. 2011. “Hormonal Contraceptive Use and HIV Disease Progression among Women in Uganda and Zimbabwe.” *Journal of Acquired Immune Deficiency Syndromes* 57(2): 157–64.

Morrison, C.S., B.A. Richardson, F. Mmiro, T. Chipato, D.D. Celentano, J. Luoto, R. Mugerwa, N. Padian, S. Rugpao, J.M. Brown, P. Cornelisse, and R.A. Salata. 2007. “Hormonal Contraception and the Risk of HIV Acquisition.” *Aids* 21(1): 85–95.

Morrison, D., O. Laeyendecker, and R. Brookmeyer. 2019. “Cross-Sectional HIV Incidence Estimation in an Evolving Epidemic.” *Statistics in Medicine* 38(19): 3614–27.

———. 2021. “Regression with Interval-Censored Covariates: Application to Cross-Sectional Incidence Estimation.” *Biometrics* online(April 17): biom.13472.

Morrison, D., O. Laeyendecker, J. Konikoff, and R. Brookmeyer. 2018. “Cross-Sectional HIV Incidence Estimation with Missing Biomarkers.” *Statistical Communications in Infectious Diseases* 10(1). De Gruyter: 1–10.

Neyman, J. 1990. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. [Master’s Thesis, 1923; Excerpts Reprinted in English, DM Dabrowska and TP Speed, Translators].” *Statistical Science* 5(4): 465–80.

Pearl, J. 1995. “Casual Diagrams for Empirical Research.” *Biometrika* 82(4): 669–710.

———. 2010. “An Introduction to Causal Inference.” *The International Journal of Biostatistics* 6(2).

Piot, P., and T.C. Quinn. 2013. “Response to the AIDS Pandemic — A Global Health Model.” *New England Journal of Medicine* 369: 1180.

R Core Team. 2019. “R: A Language and Environment for Statistical Computing.” *Vienna, Austria*.

Rehle, T., L. Johnson, T. Hallett, M. Mahy, A. Kim, H. Odido, D. Onoya, S. Jooste, O. Shisana, A. Puren, B. Parekh, and J. Stover. 2015. “A Comparison of South African National HIV Incidence Estimates: A Critical Appraisal of Different Methods.” *PLoS ONE* 10(7).

Reid, S.E., J.Y. Dai, J. Wang, B.N. Sichalwe, G. Akpomiemie, F.M. Cowan, S. Delany-MOretlwe, J.M. Baeten, J.P. Hughes, A. Wald, and C. Celum. 2010. “Pregnancy, Contraceptive Use, and HIV Acquisition in HPTN 039: Relevance for HIV Prevention Trials Among African Women.” *Journal of Acquired Immune Deficiency Syndromes* 53(5): 606–13.

Robins, J. 1986. “A New Approach To Causal Inference in Period-Application To Control of The.” *Mathematical Modelling* 7: 1393–1512.

Royall, R.M. 1986. “The Effect of Sample Size on the Meaning of Significance Tests.” *American Statistician* 40(4): 313–15.

Rubin, D.B. 1974. “Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66(5): 688–701.

Solomon, S.S., S.H. Mehta, A.M. McFall, A.K. Srikrishnan, S. Saravanan, O. Laeyendecker, P. Balakrishnan, D.D. Celentano, S. Solomon, and G.M. Lucas. 2016. “Community Viral Load, Antiretroviral Therapy Coverage, and HIV Incidence in India: A Cross-Sectional, Comparative Study.” *The Lancet HIV* 3(4): 183–90.

Sun, J. 2006. *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.

Vansteelandt, S., and N. Keiding. 2011. “Invited Commentary: G-Computation-Lost in Translation?” *American Journal of Epidemiology* 173(7): 739–42.

Westreich, D., J.K. Edwards, C.R. Lesko, E. Stuart, and S.R. Cole. 2017. “Transportability of Trial Results Using Inverse Odds of Sampling Weights.” *American Journal of Epidemiology* 186(8): 1010–14.