Title
Cross-Sectional HIV Incidence Estimation: Techniques and Challenges

Author
Konikoff, Jacob

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Cross-Sectional HIV Incidence Estimation: Techniques and Challenges

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Jacob Moss Konikoff

2015

Abstract of the Dissertation

Cross-Sectional HIV Incidence Estimation: Techniques and Challenges

by

Jacob Moss Konikoff

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2015

Professor Ronald S. Brookmeyer, Chair

Tracking and surveillance of the HIV epidemic depend on accurate estimation of the number of new infections in the population. The rate at which these infections occur, known as the incidence, is also critical for effectively designing, targeting, and evaluating prevention efforts. Incidence can be estimated through cross-sectional surveys by using biomarkers, such as HIV viral load, CD4 cell count, and recently developed serologic assays, which define and mark people in an early disease stage. The total number of individuals found in this stage, that is possessing markers of recent infection, gives a snapshot of how the epidemic is progressing.

We explore how these biomarkers should best be combined to define this early disease stage by examining how the definition influences the bias and variability of the cross-sectional incidence estimator. These calculations depend on estimating the probability that persons will remain in the early disease stage t years after seroconversion. We present two different approaches for estimating this probability curve. Once we have defined viable methods for combining these biomarkers we derive the sample sizes needed to conduct one or more cross-sectional surveys and explore how missing biomarker data should be handled in the context of implementing these surveys.

The dissertation of Jacob Moss Konikoff is approved.

Marjan Javanbakht

Catherine Ann Sugar

Robert Erin Weiss

Ronald S. Brookmeyer, Committee Chair

University of California, Los Angeles

2015

To my mother and father who have
always encouraged and challenged me;

and

to my wife
whose steadfast dedication,
unparalleled kindness,
and unshakable moral compass
continue to inspire me.

Table of Contents

Acknowledgments

This work was supported by the National Institutes of Health Grant R01-AI095068 as well as the National Institutes of Health Training Grant T32-AI-007370. A version of some of the material found in Chapters 1, 2, and 3 has been published in *The Journal of Infectious Diseases* (Laeyendecker et al., 2012), *The American Journal of Epidemiology* (Brookmeyer et al., 2013), and *Plos One* (Konikoff et al., 2013). A version of Chapter 4 has been accepted for publication by *Biometrics* (Konikoff and Brookmeyer). Some of the concepts in Chapter 5 are new approaches meant to handle statistical challenges first presented in *The Journal of Clinical Microbiology* (Cousins, Konikoff, Laeyendecker, et al., 2014) and *Plos One* (Cousins, Konikoff, Sabin, et al., 2014). While many authors contributed to these manuscripts, the statistical work, which is the focus of this dissertation, was jointly conducted with Dr. Ron Brookmeyer. Dr. Brookmeyer serves as a principal investigator for the National Institutes of Health Grant R01-AI095068.

I would like to express my sincere gratitude to my committee for their feedback and support throughout the preparation of this dissertation. In particular I am deeply indebted to Professor Ron Brookmeyer whose guidance and insight have been invaluable.

I would also like to thank my loving family, especially my wife Adina, my parents Sidney and Anita, my sisters Naomi and Riki, and my in-laws Mickey and Rena. Finally, I want to recognize Eran Barnoy and Jeremy Cytryn for profoundly enriching my graduate studies.

Vita

2010	B.S. in Mathematics and B.A. in Economics received from the University of Maryland, College Park.
2011-2015	Graduate Student Researcher at the Department of Biostatistics, University of California, Los Angeles.
2014	Advancement to Candidacy, Department of Biostatistics, University of California, Los Angeles.

Publications and Presentations

Konikoff, J., and Toplosky, S. (2010). Analysis of simplified DES algorithms. *Cryptologia*, 34(3), 211-224.

Brookmeyer, R., and **Konikoff, J.** (2011). Statistical considerations in determining HIV incidence from changes in HIV prevalence. *Statistical Communications in Infectious Diseases*, 3(1), article 9.

Laeyendecker, O., Brookmeyer, R., Cousins, M. M., Mullis, C. E., **Konikoff, J.,** Donnell, D., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Mehta S. H., Astemborksi, J., Jacobson, L. P., Margolick, J. B., Brown J., Quinn, T. C., and Eshleman S. H. (2013). HIV incidence determination in the United States: A multi-assay approach. *Journal of Infectious Diseases*, 207(2), 232-239. doi: 10.1093/infdis/jis659

Brookmeyer, R., **Konikoff J.,** Laeyendecker, O., and Eshleman, S. H. (2013). Estimation of HIV incidence using multiple biomarkers. *American Journal of Epidemiology*, 49(10), 264-272.

Konikoff, J., Brookmeyer, R., Longosz, A. F., Cousins, M. M., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Moore, R. D., Mehta S. H., Margolick, J. B., Brown J., Mayer, K. H., Koblin, B. A., Justman, J. E., Hodder, S. L., Quinn, T. C., Eshleman S. H., and

Laeyendecker, O. (2013). Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of HIV incidence in the United States. *PloS One*, 8(12), e82772. doi: 10.1371/journal.pone.0082772

Cousins, M. M., **Konikoff, J.,** Laeyendecker, O., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Moore, R. D., Mehta S. H., Margolick, J. B., Brown J., Mayer, K. H., Koblin, B. A., Wheeler, D., Justman, J. E., Hodder, S. L., Quinn, T. C., Brookmeyer, R., and Eshleman S. H. (2014). HIV diversity as a biomarker for HIV incidence estimation: including a high-resolution melting diversity assay in a multiassay algorithm. *Journal of Clinical Microbiology*, 52(1), 115-121. doi: 10.1128/JCM.02040-13

Cousins, M. M., **Konikoff, J.,** Sabin, D., Khaki, L., Longosz, A. F., Laeyendecker, O., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Moore, R. D., Mehta S. H., Margolick, J. B., Brown J., Mayer, K. H., Koblin, B. A., Wheeler, D., Justman, J. E., Hodder, S. L., Quinn, T. C., Brookmeyer, R., and Eshleman S. H. (2014). A comparison of two measures of HIV diversity in multi-assay algorithms for HIV incidence estimation. *PloS One*, 9(6), e101043. doi: 10.1371/journal.pone.0101043

Konikoff, J. and Brookmeyer R. (Accepted for publication April 2015). Sample size methods for estimating HIV incidence from cross-sectional surveys. *Biometrics.*

Konikoff, J. and Brookmeyer R. (July 2012). "Estimating HIV incidence from changes in HIV prevalence." Presented at the Joint Statistical Meetings, San Diego, CA.

Konikoff, J. and Brookmeyer R. (June 2014). "Sample size methods for cross-sectional HIV incidence and trend estimation." Presented at the Annual Meeting of the Western North American Region of the International Biometric Society at the University of Hawaii, Manoa.

Konikoff, J. and Brookmeyer R. (March 2015). "Sample size methods for estimating HIV incidence from cross-sectional surveys." Presented at the Annual Meeting of the Eastern North American Region of the International Biometric Society, Miami, Florida.

CHAPTER 1

Introduction

1.1 Background

Monitoring of the HIV epidemic depends critically on the ability to measure new infections in the population. Reliable estimates of incidence, the rate at which new infections emerge from the susceptible part of the population, additionally empower researchers seeking to target and evaluate prevention efforts. Moreover, since HIV infection is followed by an asymptomatic phase lasting for years, major changes in the epidemic will manifest themselves in accurate estimates of incidence long before they are felt by the healthcare system. Incidence estimation can thus serve as an early warning system, alerting public health officials about upcoming changes in the disease burden and allowing for the allocation of resources to properly engage the leading edge of the epidemic.

Biomarker-based assays have been used to measure HIV incidence from cross-sectional studies for more than two decades (Brookmeyer and Quinn, 1995). They offer a promising alternative to longitudinal studies, which can be difficult to implement, particularly in highly stigmatized and marginalized populations. In these settings selection bias and loss to follow up are profound concerns that can theoretically be circumvented by taking a large random sample of the population at a single point in time (Brookmeyer, 2010a).

The original idea for the cross-sectional study was to use a biomarker, the presence of HIV p24 antigen in the blood, to determine which samples in the survey came from infected individuals. Samples which tested positive for p24 antigen were then tested further with an antibody assay to determine if they came from individuals infected in the recent past. Since individuals, on average, show detectable levels of the virus for 22.5 days before they

test positive for HIV antibodies (Brookmeyer and Quinn, 1995) we say that individuals who are positive for p24 antigen but negative for HIV antibodies are in an early disease stage associated with recent infection. While this methodology can be used to ascertain the rate at which new infections emerge from the susceptible part of the population, the cost is prohibitive. The short duration individuals spend in the early disease stage makes the number of individuals found in this state highly variable. Large samples are needed to control this variability and accurately measure incidence. Further, testing for p24 antigen is expensive. While the much cheaper antibody assay can be run first, and the test for p24 antigen second, this will not substantially reduce costs as most populations are predominately uninfected and are thus antibody negative.

These limitations can theoretically be solved by changing the definition of the early disease stage used for estimating incidence to a stage of infection individuals enter *after* seroconverting. In this framework, infected individuals who have not yet seroconverted are grouped with the uninfected individuals and jointly considered the at risk group for entering this early disease stage. Individuals who test positive for HIV antibodies, a much smaller group than those who test negative for HIV antibodies, are further tested and classified as either in or out of the early disease stage based on one or more biomarkers. For example, the Centers for Disease Control (CDC) developed and promoted the BED IgG capture enzyme immunoassay (BED-CEIA) specifically for this purpose. In this case the early disease stage is defined as an assay result below the cutoff of 0.8 optical density units (ODn) (Dobbs et al., 2011). The assay works by assessing the proportion of IgG antibodies directed against certain HIV antigens. Low results are thus indicative of a less mature immune response. Recent work has estimated that, on average, people spend close to 200 days in this early disease stage with an assay result below 0.8 ODn (Hargrove et al., 2012).. This assay has been used all over the world to estimate HIV incidence (Dobbs et al., 2011).

Despite the theoretical promise, the Joint United Nations Programme on HIV/AIDS (UNAIDS) found that the BED-CEIA overestimates HIV incidence and cautioned against its use (UNAIDS Epidemiology Reference Group Secretariat, 2005). It was discovered that some individuals remained in the early disease state (below 0.8 ODn) for several years after

seroconversion. While most individuals will develop a strong enough immune response to score above 0.8 ODn within one year of seroconverting, a percentage of the population will only begin to test above 0.8 ODn after many years. Moreover, the immune system may deteriorate with the onset of advanced disease allowing individuals to reenter the early disease stage. When this occurs the disease stage is not indicative of recent infection and the incidence estimate will not convey information about current incidence trends.

The issues with the BED-CEIA do not invalidate the general method but rather highlight the importance of properly defining an early disease stage. In the next section we will look more closely at the cross-sectional incidence estimator's derivation. By examining the estimator we will develop criteria for specifying an early disease stage to estimate incidence accurately.

1.2 Framework and Notation

At calendar time t we conduct a cross-sectional survey of n persons to produce an estimate of the incidence I. Blood or other biological specimens are collected from each person. Those specimens are analyzed for various biomarkers to classify each person into one of three groups. A standard HIV antibody test divides the n collected samples into N_u samples from individuals considered uninfected and N_i samples from infected individuals. The infected samples are then further divided into X samples from people in the early disease stage and $N_i - X$ samples from individuals who have progressed out of this early disease stage. While n is fixed before the survey, N_u, N_i, and X are all random variables whose realizations we denote by n_u, n_i, and x.

We formally define the incidence, I, as the expected number of new infections per uninfected persons per year in a specific population. Then $I = g(0)(1-p)^{-1}$ where p denotes the HIV prevalence at the time of the survey and $g(t)$ is the expected number of new infections per year, divided by the size of the population, t years *before* the cross-sectional survey occurs. The denominator $(1-p)^{-1}$ scales $g(0)$ so the rate is per the uninfected proportion of the population. This makes incidence a hazard rate insofar as it is the instantaneous event

rate conditional upon survival (remaining uninfected) up to the time of the survey.

The cross-sectional incidence estimator used in the literature assumes that $g(0) = \pi\mu^{-1}$ (Kaplan and Brookmeyer, 1999), where π is the probability of being in the early disease stage at the time of the cross-sectional survey and μ is the average amount of time individuals spend in this early disease stage. The cross-sectional incidence estimator is therefore not estimating I but $I_c = \pi\{(1-p)\mu\}^{-1}$. If individuals spend at most M years in the early disease stage, $I = I_c$ when $g(t)$ has remained constant over the time period stretching back M years (Kaplan and Brookmeyer, 1999). This is similar to a constant incidence assumption over those M years, the difference being that the new infections are compared to the entire population instead of the at risk group. When the original methodology was proposed this assumption was reasonable because M was small and incidence is always approximately constant over a short enough time period.

When the survey is conducted we estimate I_c with the following rule or estimator $\widehat{I_c} = X(N_u\mu)^{-1}$. The estimator follows from the fact that

$$\frac{1}{1-\hat{p}} = \frac{\text{\# of individuals sampled}}{\text{\# of individuals who test HIV negative}}$$

and

$$\hat{\pi} = \frac{\text{\# of individuals in the early disease stage}}{\text{\# of individuals sampled}}.$$

When the BED-CEIA is used by itself some individuals remain below 0.8 ODn for several years. It is not reasonable to assume that $g(t)$ is constant over this time period. Therefore, $I \neq I_c$ and the cross-sectional incidence estimator is not measuring the current incidence.

In this case I_c can still be used to approximate incidence, albeit no longer the instantaneous hazard rate at the time of the survey. It can be shown that $\widehat{I_c} = X(N_u\mu)^{-1}$ is a potentially biased estimator of incidence at some point in the past. To see this we can break down the components of I_c. First, $\pi = \int_0^M g(t)\phi(t)dt$, where $\phi(t)$ is the probability that persons infected t years ago will be in the early disease stage at the time of the survey. We note that $\phi(t) = 0$ for $t > M$ and $\mu = \int_0^M \phi(t)dt$. We therefore may define a random variable

S whose probability density is given by $f_s(t) = \mu^{-1}\phi(t)$ for $t \geq 0$. Thus, $\pi = \mu \cdot \mathrm{E}[g(S)]$ and

$$I_c = \frac{1}{1-p}\frac{\pi}{\mu} = \frac{\int_0^M g(t) f_s(t) dt}{1-p} = \frac{\mathrm{E}[g(S)]}{1-p}.$$

If we consider a first order Talyor expansion of $g(t)$ about ψ=E[S] we note that

$$I_c \approx \frac{1}{1-p}\int_0^M \left(g(\psi) + g'(\psi)(t-\psi)\right) f_s(t) dt =$$

$$\frac{g(\psi)}{1-p}\int_0^M f_s(t) dt + \frac{g'(\psi)}{1-p}\int_0^M (t-\psi) f_s(t) dt = \frac{g(\psi)}{1-p}.$$

This approximation is equivalent to asserting $\mathrm{E}[g(S)] \approx g(\mathrm{E}[S])$.

Then I_c is approximately the incidence ψ years ago which we will denote $I(\psi)$. If the current prevalence is approximately equal to the prevalence ψ years ago then the difference between I_c and $I(\psi)$ will be $\int_0^M \left[\int_\psi^t g''(x)(t-x)dx\right] f_s(t)dt$. This comes from the remainder theorem for a Taylor expansion (Beesack, 1966). In cases where $g(t)$ is approximately linear over the past M years this difference will be negligible since $g''(x)$ will be close to 0. We will later explore the reliability of the estimator when $g(t)$ is not approximately linear over the past M years.

We call ψ the shadow since it determines how far our incidence estimate is cast backward into the past. A small shadow implies both that we are estimating recent incidence and that our assumption that the prevalence has not changed significantly in the past ψ years may be reasonable. Since $\psi = \mu^{-1}\int_0^\infty t\phi(t)dt$ we can rewrite it as

$$\frac{\int_0^M t\phi(t)dt}{\int_0^M \phi(t)dt}.$$

so that ψ is a weighted average of the duration of time early disease stage cases have already spent in that stage prior to the study. Each time t in the past when individuals seroconverted is weighted by $\phi(t)$, the probability of remaining in the early disease stage t years after seroconversion. Thus each infected individual gives information about the incidence at the

time of infection and not the current instantaneous hazard rate.

It is important to note that ψ, like μ, is dependent on the specific early disease stage definition. Our estimate of $\phi(t)$ will determine both $\psi = \mu^{-1} \int_0^\infty t\phi(t)dt$ and $\mu = \int_0^M \phi(t)dt$. As an example, if the BED-CEIA is used with a cutoff of 2 ODn, people will necessarily spend longer in the early disease stage than if the BED-CEIA were used with the traditional cutoff of 0.8 ODn. We will later develop methods for selecting cutoffs.

The $\phi(t)$ curve is also crucial for understanding the bias in the cross-sectional incidence estimator in cases where incidence is unlikely to be linear over the past M years. The main difference between I_c and $I(\psi)$ can be written as

$$\frac{\int_0^M \left[\int_\psi^t g''(x)(t-x)dx\right] \phi(t)dt}{\int_0^M \phi(t)dt}.$$

This is again a weighted average where $\phi(t)$ provides the weights. As incidence will be approximately linear over a small enough time period an ideal $\phi(t)$ will approach zero rapidly. This implies the probability of remaining in the early disease stage for long periods must be close to, if not equal to, zero. If this is not the case the cross-sectional incidence estimator will be biased and will not approximate the incidence ψ years in the past.

In the next chapter we explore how to incorporate these concepts into defining the early disease stage so that incidence can be estimated in the recent past.

CHAPTER 2

Preliminary Work

2.1 Defining the Early Disease Stage

We mentioned above that one reason long term infections remain in the early disease stage is immune collapse. Since HIV attacks the immune system a measure of disease progression based solely on the antibody response may fail to distinguish recent infections from particularly advanced infections. One way to mitigate this problem is to include a non-serologic assay in the definition of the early disease stage. The health of the immune system of an infected individual is often measured in terms of the number of CD4 positive cells per mm^3 of blood. The use of both the BED-CEIA and CD4 cell count can thus define an early disease stage which includes individuals with immature immune responses to HIV but excludes those whose immune systems have deteriorated from advanced disease.

From here we can see the advantage of using more than one biomarker. We call a testing algorithm which defines the early disease stage with multiple biomarkers a multi-assay algorithm (MAA). We first introduced this type of testing algorithm in 2013 (Laeyendecker et al., 2012).

This first published MAA used four different biomarkers to define the early disease stage. In addition to only including individuals with a CD4 cell count greater than 200 cells/mm^3, corresponding to clinical AIDS, individuals were required to have an HIV viral load of greater than 400 copies/mL. A viral load below this cutoff implies the individual is virally suppressed and likely to be on antiretroviral therapy – a treatment process most individuals do not start early in infection. This is partially because people do not immediately realize they have been infected with the virus.

The other two biomarkers used were serologic assays specifically developed for estimating HIV incidence. The BED-CEIA, described above, was used with a slightly higher cutoff of 1.0 ODn. The other serologic assay, the BioRad Avidity assay, measures the strength of antibody binding to target antigens (Masciotra et al., 2010). The strength of antibody binding is thought to increase over the course of infection making lower scores indicative of recent infection. The result from the BioRad Avidity assay is an avidity index calculated as the percentage of antigen-binding of chaotropic-treated antibody compared to the antigen-binding of non-treated antibody. Individuals with an avidity index less than 80% were included in the early disease stage. This cutoff, like the one for the BED-CEIA, was chosen to err on the less restrictive side of suggested cutoffs proposed by individuals who have worked closely with these assays. When multiple assays are used no individual assay must exclude all the long-term prevalent cases from the early disease stage.

Figure 2.1 illustrates how the early disease stage is defined by sequentially testing samples from individuals who have seroconverted.

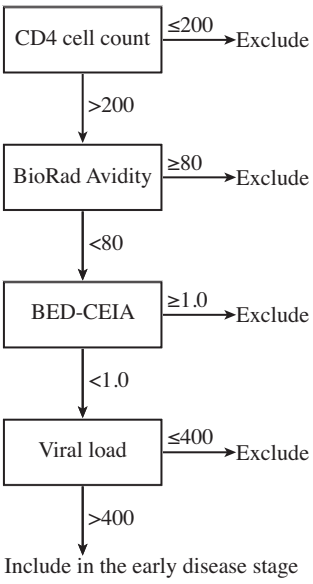

Figure 2.1: The first published MAA. The following units are used for the component assays: BED-CEIA: ODn; BioRad Avidity assay: percentage (avidity index); CD4 cell count: cells/mm^3; viral load: copies/mL. Abbreviations: BED-CEIA: BED IgG capture enzyme immunoassay; BioRad Avidity: BioRad Avidity assay; ODn: normalized optical density units.

We next set about evaluating the performance of this MAA based on considerations described in the previous section.

2.2 Description of the Data

Data are available from subsets of the three cohort studies, the Multicenter AIDS Cohort Study (MACS), the HIV Network for Prevention Trials (HIVNET 001/001.1) vaccine preparedness cohort, and the AIDS Link to Intravenous Experience (ALIVE) cohort (Kaslow et al., 1987; Celum et al., 2001; Vlahov et al., 1990). In total 709 unique individuals with known seroconversion windows provided 1,782 samples. The seroconversion window is defined by the date of the last known negative HIV antibody test, d^-, and the date of the first known positive HIV antibody test, d^+. Some samples were taken when individuals were in the process of seroconverting. We assumed these individuals seroconverted within 28 days of the samples being drawn.

For each individual i continued measurements j of the four biomarkers are taken at times τ_{ij} after d_i^+. We say that the measurements are taken at times t_{ij} following seroconversion where $t_{ij} = \tau_{ij} + s_i$ and s_i is the unknown amount of time between seroconversion and d_i^+. Figure 2.2 illustrates this data setup.

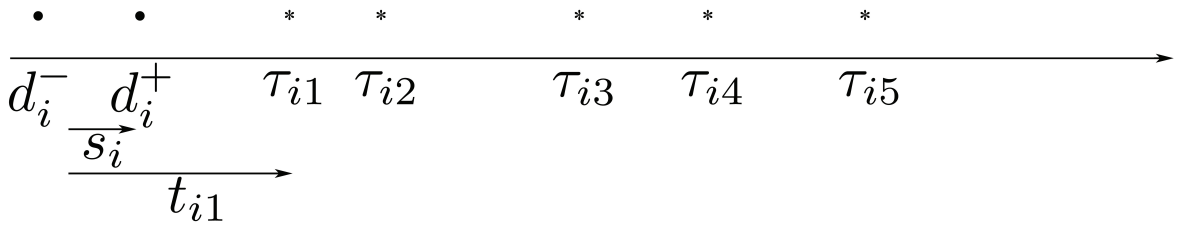

Figure 2.2: Data structure for one individual i with observations $j = 1, 2, 3, 4$, and 5. The • indicates that blood was drawn and an assay was run to test for the presence of antibodies directed against HIV. The * indicates that blood was drawn and evaluated with various assays to measure biomarkers.

Of the 709 individuals, 244 contributed a single observation, 369 contributed 2 observations, and 96 contributed between 3 and 16 observations. The unknown seroconversion intervals ranged from 12 days to 1.5 years with an average seroconversion interval of 5.7

months. Based on the midpoint of the seroconversion windows, people were sampled up to 8.6 years after seroconversion. All four biomarkers were measured on every sample. All the samples are believed to be infected with Clade B virus. More details about the samples collected from these studies can be found in our previous work (Laeyendecker et al., 2012).

An additional 500 samples from the Johns Hopkins Hospital Clinical Cohort (JHHCC) known to be infected more than eight years were used for supplemental analysis (Moore, 1998). For these samples there was only a first known positive HIV antibody test and no last known negative HIV antibody test.

2.3 Statistical Evaluation

Some of the following methodology is updated from the paper where this MAA was first published (Laeyendecker et al., 2012). We will not highlight these minor changes but we note that they lead to small changes in parameter estimates. The conclusions remain the same.

For each of the 1,782 blood samples the MAA shown in Figure 2.1 was applied to create a new variable y_{ij} which is 1 if the sample is in the early disease stage and 0 otherwise. Logistic regression was then used to model the probability that $y_{ij} = 1$ as a function of the time t_{ij} since seroconversion, using a cubic spline with a knot at two years. This model was chosen to try and impose minimal assumptions on the shape of the curve. Specifically, the approach does not require the probability of being in the early disease stage to be one at the time of seroconversion nor does it require the probability to monotonically decrease with the duration of infection. From a biological perspective a small percentage of individuals naturally control the virus and are likely to have a viral load below 400 copies/mL. Thus we would not expect 100% of individuals to be in the early disease stage at seroconversion. Further, since individuals can reenter the early disease stage, monotonicity is not guaranteed.

This model does not account for the correlation within subjects but rather relies on the consistency of the parameter estimates and the large number of individuals sampled. When handling issues related to the variability of the estimates produced by the model, we

account for the correlation within subjects through a clustered bootstrap described shortly. We further discuss incorporating the correlation that exists within subjects in Chapter 6.

Because the time of seroconversion was unknown the model was fit by randomly sampling a seroconversion time from a uniform distribution over the possible dates for each individual. This process was repeated 1,000 times and the fitted curves were averaged. The model is then used to estimate $\phi(t)$ for all t by noting that $\phi(t^*) \approx \mathrm{P}(y^* = 1|t^*)$ where y^* is a new sample from the population from an individual who seroconverted t^* years before the cross-sectional survey. The approximation is due to the fact that only living people are sampled. Thus, $\phi(t^*) = \mathrm{P}(y^* = 1|t^*, \text{survival})\mathrm{P}(\text{survival}|t^*)$. The effect of survival on $\phi(t)$ is minimal when $\phi(t)$ approaches 0 rapidly with increasing time.

We only examined the fit of the $\phi(t)$ curve for $0 \leq t \leq 8$ to avoid extrapolating outside of our data and to preclude possible unruly behavior by the spline curves near the edge of the data. There is no need to estimate $\phi(t)$ past $t = M$ years. An outline of the procedure is detailed below.

---

1. Based on the MAA in Figure 2.1 for each sample create y_{ij} which equals 1 if the sample is in the early disease stage and 0 otherwise
2. Draw s_i from a uniform distribution on the interval $(0, d^+ - d^-)$ for $1 \leq i \leq 709$ and calculate $t_{ij} = \tau_{ij} + s_i$ for each sample
3. Fit the model $\text{logit}[\mathrm{P}(y_{ij} = 1)] = \beta_0^{(k)} + \beta_1^{(k)} t_{ij} + \beta_2^{(k)} t_{ij}^2 + \beta_3^{(k)} t_{ij}^3 + \beta_4^{(k)} (t_{ij} - 2)_+^3$
4. Repeat steps 2 and 3 for $1 \leq k \leq 1000$ and calculate $\sum_k \beta_\ell^{(k)}/1000 = \beta_\ell$ for $\ell = 0$, 1, 2, 3, and 4
5. Let $\phi(t) = \text{expit}\{\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 (t - 2)_+^3\}$ for $0 \leq t \leq 8$

For this MAA we found that no sample infected > 5 years, based on the midpoint of the seroconversion window, was in the early disease stage and that $\phi(t)$ was 0 at 8 years. Further

support that $M < 8$ comes from the fact that none of the 500 samples from the JHHCC (all infected more than 8 years) were in the early disease stage. Figure 2.3 shows the estimated $\phi(t)$ curve.

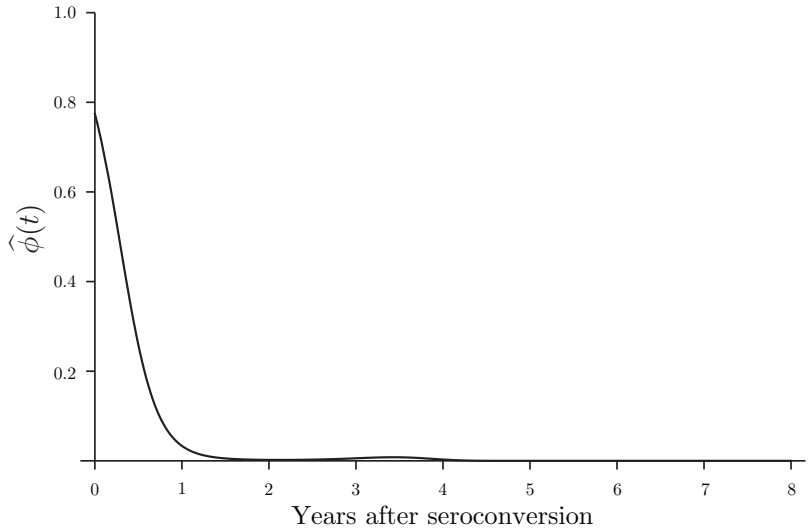

Figure 2.3: Probability of being in the early disease stage by years since seroconversion. Individuals are marked in the early disease stage if they were found to have a CD4 cell count > 200 cells/mm^3, a BioRad Avidity assay result $< 80\%$, a BED-CEIA result < 1.0 ODn, and an HIV viral load > 400 copies/mL.

We calculated μ and ψ based on $\widehat{\phi}(t)$. Our estimate of μ was 0.34 years and our estimate for ψ was 0.40 years. In order to generate a confidence interval (CI) for each of these parameters we performed a clustered bootstrap. The number of individuals per study was fixed in each iteration of the bootstrap and individuals were sampled with all of their observations to preserve their correlation structures (Davison and Hinkley, 1997; Ren et al., 2010; Field and Welsh, 2007). We performed 999 iterations of the bootstrap and generated 95% confidence intervals for μ, (0.27, 0.40) years, and ψ, (0.32, 0.54) years, based on the percentile method (DiCiccio and Romano, 1988). Some of the iterations of the bootstrap did not include any samples which were both collected more than two years after seroconversion and were in the early disease stage. In these iterations the term associated with the knot separates the data perfectly. In these cases the integral was only taken out to 2 years and we assumed $M = 2$.

This MAA has a significantly greater estimated μ than the 22.5 days associated with the originally proposed method. The estimated shadow of 0.40 years means that incidence

is being estimated in the recent past. Furthermore, $\widehat{\phi}(t)$ remains close to zero after 1 year, implying that the method may be applicable even if $g(t)$ is historically non-linear. This also minimizes the effect of not incorporating a survival function.

For comparison we note that the same statistical methods imply that when the BED-CEIA is used by itself, with a cutoff of 0.8 ODn, over 10% of individuals are estimated to remain in the early disease at 8 years past seroconversion. In a best possible scenario, where this percentage somehow drops to 0 after 8 years, we estimate that the shadow would be nearly 3 years. The MAA offers a significant improvement over the BED-CEIA.

We concluded that this MAA could be used to accurately estimate incidence in Clade B epidemics which exist in the United States and parts of Europe. We examine potential improvements to this MAA in the coming sections.

CHAPTER 3

Statistical Considerations in Designing MAAs

3.1 Search Criteria for Promising MAAs

The first published MAA had component assays and cutoffs chosen based on the natural history of HIV infection. In this section we explore how to use statistical considerations to design an MAA.

After we published our first paper focused on these considerations (Brookmeyer et al., 2013) the CDC followed up on the BED-CEIA and released a second serologic assay also based on IgG antibodies called the limiting antigen avidity enzyme immunoassay (LAg Avidity assay). We elected to include this assay in our future work as a replacement for the BED-CEIA (Konikoff et al., 2013).

In order to minimize the variability in the cross-sectional incidence estimator we attempted to find the MAA with the largest μ subject to the following constraints:

1. The estimated probability of being classified in the early disease stage at 8 years after seroconversion needed to be < 0.001.

2. None of the samples infected more than 8 years from the JHHCC could be found to be in the early disease stage.

3. The upper bound of the 95% confidence interval for the shadow needed to be less than 1 year.

4. The point estimate of the shadow, $\hat{\psi}$, needed to be less than 250 days.

The first two criteria are meant to ensure that $M \leq 8$ years so that we can properly estimate $\phi(t)$ from our data. Ideally M would be much smaller based on the considerations described in the previous chapter. The last two criteria are meant to ensure the MAA estimates incidence within the past year. Estimating incidence within one year of a cross-sectional survey has been a stated goal of the National Institutes of Health. The cutoff for the point estimate of the shadow of 250 days was chosen because there is a computational expense in calculating the confidence interval for ψ for each algorithm. When we examined confidence intervals for the shadow associated with MAAs that had point estimates close to 250 days, we found the upper limits were extremely close to the one year cutoff. Thus, the 250 day cutoff can be viewed as a further conservative measure to ensure incidence is being estimated within the year preceding the survey.

We elected to consider algorithms which can be performed sequentially in the form of the MAA presented in Figure 2.1. The advantage of this type of algorithm is that at each step the particular component assay will exclude a proportion of the samples from the early disease stage. This will reduce cost since these samples will not require further testing by the remaining assays. MAAs need not take this form, and other types of MAAs may offer additional unexplored advantages.

In order to find the MAA with the largest μ, subject to the above constraints, we performed a brute-force search. We looked at 22 cutoffs ranging from $0.5 - 3.9$ ODn for the LAg Avidity assay, 12 cutoffs ranging from $30 - 100\%$ (avidity index) for the BioRad Avidity assay, 10 cutoffs ranging from $400 - 10,000$ copies/mL for HIV viral load, and 12 cutoffs ranging from $50 - 1,000$ cells/mm^3 for CD4 cell count. The search covered all combinations of these cutoffs plus the possibility of not using one or more component assays.

For each set of component assays and cutoffs we estimated the $\phi(t)$ curve based on the same procedure described above. Two samples had been depleted from previous testing and could not be tested with the LAg Avidity assay. While these were deleted from the overall search they were used in the calculations below since the other component assays were able to exclude them from the early disease stage for the specific MAAs we discuss further. We denote the MAA with the largest μ subject to the four constraints as MAA2 and refer to

the MAA which we described in Chapter 2 as MAA1.

A potential limitation of this search is the inclusion of CD4 cell count as a biomarker in these algorithms. Unlike the other biomarkers discussed, CD4 cell counts must be measured at the time of sample collection. This may limit the applicability of these MAAs. It will also drive up costs as the CD4 cell counts will likely need to be measured first. The cost of measuring CD4 cell counts is much greater than that of performing the LAg Avidity and BioRad Avidity assays. Thus making CD4 cell count the first component assay reduces some of the gain the sequential testing format offers. Ideally most samples should already be excluded from the early disease stage before expensive assays are applied. These considerations led us to find MAA3, the MAA which has the largest μ, subject to the above constraints, which did not include CD4 cell count.

Lastly, we considered alternative optimality criteria which were meant to reduce the potential for bias in the cross-sectional incidence estimator. We searched for MAA4 which had the highest μ of MAAs which, based on the midpoint of the seroconversion interval, (1) marked no more than 1 in 1000 samples in the early disease stage among the samples in the data set which were collected more than one year past seroconversion and (2) marked 0 of the 1000 samples infected for the longest amount of time in the early disease stage. In Section 3.3.2 we will examine the bias in the cross-sectional incidence estimator in more detail.

3.2 Results

Since both μ and ψ depend on $\phi(t)$ there is an inherent tradeoff between having algorithms with larger μ and those with smaller ψ. The longer people spend in the early disease stage the less variability there is in the cross-sectional incidence estimator. When people spend too long in the early disease stage we no longer can learn about recent incidence, the estimator will be biased, and ψ will be large. The tradeoff can therefore be thought of as the classical one between variance and bias. Figure 3.1 illustrates this tradeoff by showing $\widehat{\mu}$ plotted against $\hat{\psi}$ for those MAAs which have $\widehat{\phi}(t) < 0.001$ at 8 years and had $\widehat{\psi} < 250$ days. The

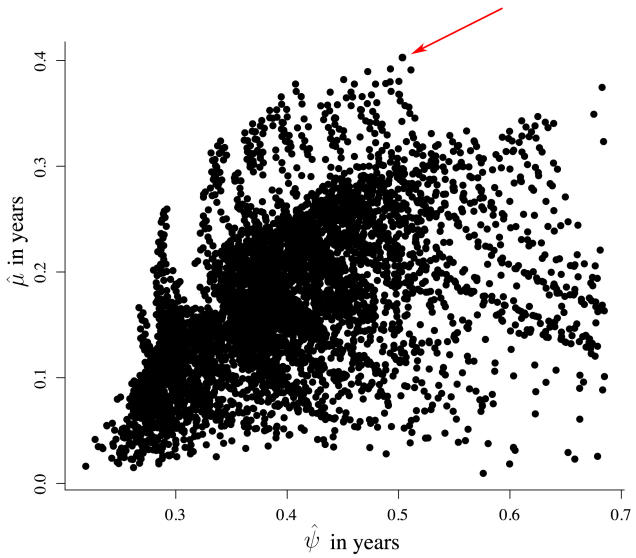

Figure 3.1: Plot of estimated mean duration spent in the early disease stage $\hat{\mu}$ against the estimated shadow $\hat{\psi}$ for MAAs with $\widehat{\phi}(t) < 0.001$ at 8 years and $\widehat{\psi} < 250$ days.

red arrow indicates MAA2.

MAA2 marks individuals in the early disease stage if they are found to have a CD4 cell count > 50 cells/mm^3, a BioRad Avidity assay result $< 85\%$, a LAg Avidity assay result < 2.9 ODn, and an HIV viral load > 400 copies/mL. For MAA2 μ was estimated as 0.40 years (95% CI 0.33, 0.47) and ψ as 0.49 years (95% CI 0.39, 0.63). Figure 3.2 illustrates how this MAA classifies the samples in our data set. Only the subset of the samples which have viral loads > 400 copies/mL are plotted in order to reduce the number of dimensions to three. Those samples inside of the pink rectangular box are marked in the early disease stage. The data are plotted so that the darkness of each point corresponds to the length of time the sample was taken from seroconversion. The midpoint of the seroconversion interval was used for illustration. Panel A plots all 1,391 samples with viral loads > 400 copies/mL. Since it can be hard to see which samples are excluded from the early disease stage, Panel B plots only those individuals who have been marked in the early disease stage. There is a clear association between more recent infections and being marked in the early disease stage. This association is by no means perfect and many samples infected less than one year are excluded from the early disease stage. Similarly, a handful of samples infected greater than

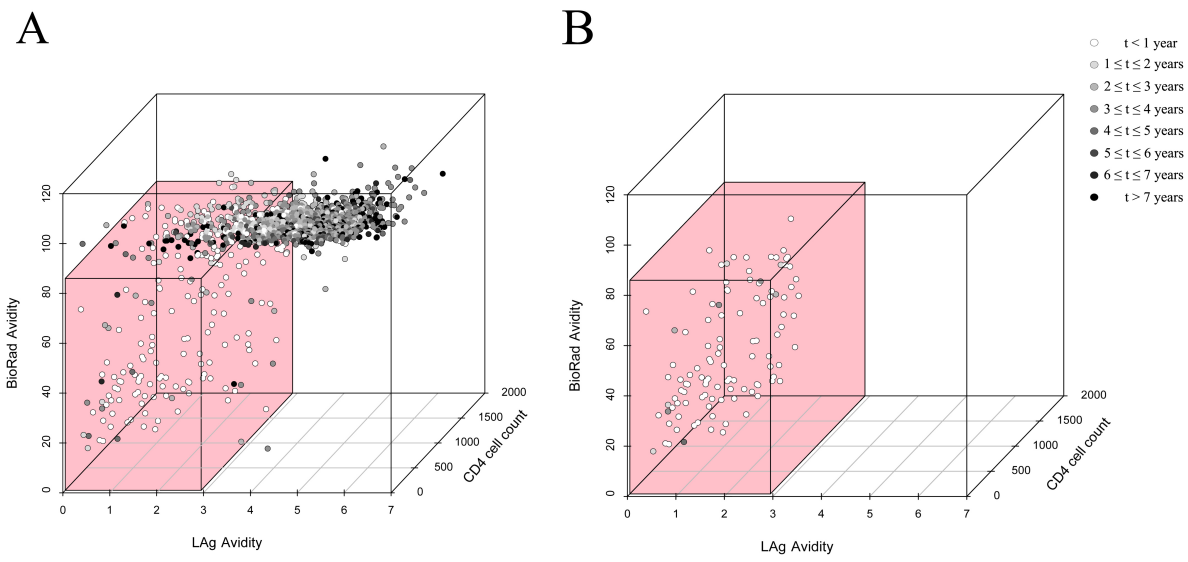

Figure 3.2: Data classification for MAA2. The pink rectangular box indicates the early disease stage. Panel A shows all the samples in the data set with viral load > 400 copies/mL. Panel B shows only those samples marked in the early disease stage.

one year are included.

Somewhat surprisingly MAA3 does not include HIV viral load and marks individuals in the early disease stage solely based on the two serologic assays. Thus the combination of the LAg Avidity and the BioRad Avidity assays together would appear preferable to 3-assay MAAs which also include viral load. MAA3 marks individuals in the early disease stage if they have a BioRad Avidity assay result $< 40\%$ and a LAg Avidity assay result < 2.8 ODn. Without CD4 cell count a much stricter cutoff on the BioRad Avidity assay is needed to exclude individuals who have been infected for long periods of time from the early disease stage.

Figure 3.3 illustrates how this MAA classifies the samples in our data set. The pink rectangular box corresponds to the early disease stage defined by this MAA. When compared to MAA2 this MAA includes in the early disease stage fewer samples infected for less than one year (65 compared to 110) and more samples infected for greater than 1 year (16 compared to 10). This results in both a lower estimate for μ and higher estimate for ψ. For MAA3 these were estimated to be 0.33 years (95% CI 0.26, 0.39) and 0.68 years (95% CI 0.43, 0.97) respectively.

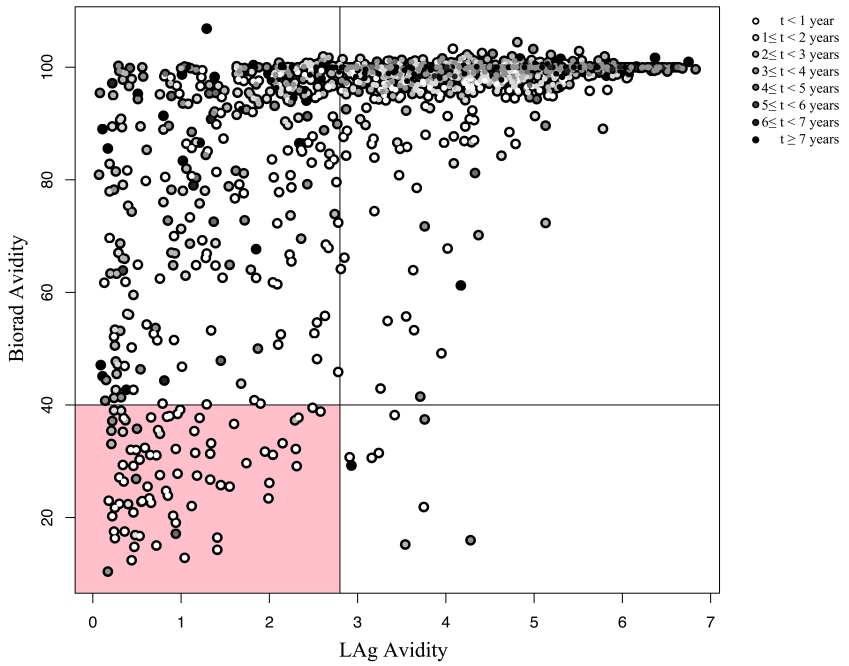

Figure 3.3: Data classification for MAA3. The pink rectangle shows those samples marked in the early disease stage.

MAA4 includes people in the early disease stage if they have a CD4 cell count > 350 cells/mm^3, a BioRad Avidity assay result $< 85\%$, a LAg Avidity assay result < 2.9 ODn, and an HIV viral load > 600 copies/mL. MAA4 has an estimated μ of 0.34 years (95% CI 0.26, 0.41) and ψ of 0.33 years (95% CI 0.29, 0.38). Figure 3.4 illustrates which samples are classified in the early disease stage for this MAA. By design all long term infections have been excluded in an attempt to reduce bias.

The $\phi(t)$ curve for MAA4 was fit using a quadratic spline. The cubic spline with the knot could not be used since none of the samples taken more than two years after seroconversion are in the early disease stage when MAA4 is used. The quadratic spline model was chosen over the cubic spline model based on the Akaike information criterion. This highlights a weakness of this methodology. Since the data being modeled changes with each MAA a different model may be appropriate for each MAA. However, since the brute force search covered over 30,000 different MAAs it is impossible to model each data set separately. A cubic spline with a knot was chosen to model the data which arose from the definition of the early disease stage associated with MAA1. Initial attempts to address this problem focused

on creating specific rules. We delineated situations where the knot at 2 years should not be included or where a quadratic spline would be more appropriate than a cubic spline. In Chapter 6 we present an approach to estimating $\phi(t)$ which avoids this problem entirely by modeling the underlying continuous progression of the various biomarkers.

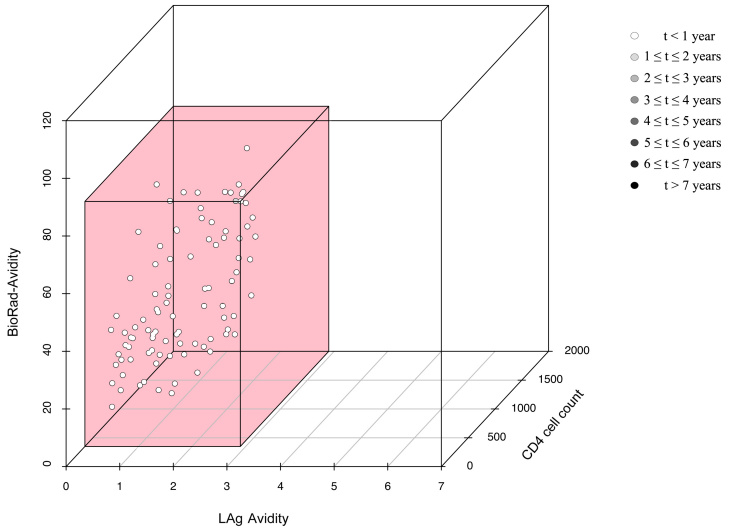

Figure 3.4: Early disease stage cases for MAA4.

Figure 3.5 displays the estimated $\phi(t)$ curves for MAA1, MAA2, MAA3, and MAA4 respectively.

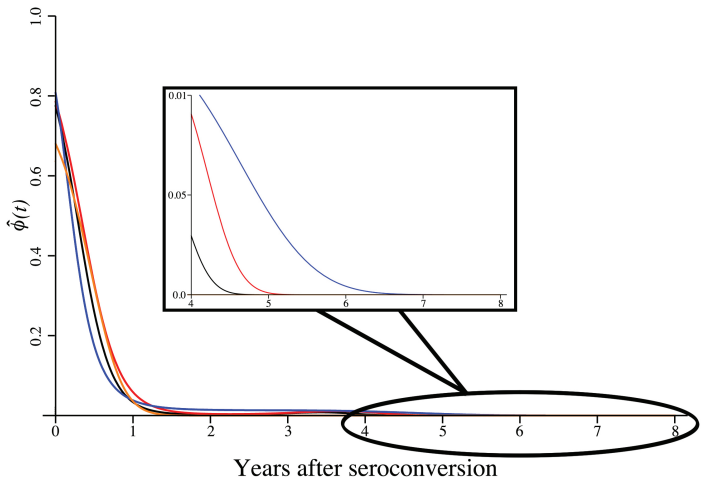

Figure 3.5: Estimated $\phi(t)$ curves for MAA1, in black, MAA2, in red, MAA3, in blue, and MAA4 in orange.

While all four MAAs rapidly decrease toward 0 we see that the MAA3 (in blue), which does

not use CD4 cell count and HIV viral load, has a heavier tail than the other MAAs. This is why MAA3 has the highest shadow of the four MAAs we have discussed. All four MAAs have $M < 7$ years but MAA4 has by far the smallest M with $\phi(t)$ converging to zero by 1.5 years. When $\phi(t)$ approaches zero, and how close it stays to zero for larger values of t, will affect the validity of the assumption that $I_c \approx I(\psi)$ when $g(t)$ is not linear. Even slight differences in the tail of $\phi(t)$ can have a major impact on the validity of the incidence estimator. We included MAA4 for this reason. The four conditions listed in Section 3.1 will allow us to estimate incidence within the past year only if $I_c \approx I(\psi)$. If this assumption holds then MAA2 improves on MAA1 by increasing μ which should thus reduce the variability in the incidence estimator. MAA3 maintains a similar μ to MAA1 while avoiding incorporating CD4 cell count and viral load. This allows MAA3 to be performed on stored blood samples and should make MAA3 cost significantly less than the other MAAs. We may think of MAA2, MAA3 and MAA4 as different tools for different situations. We will focus our future attention on these three MAAs. In the next sections we explore issues of bias, determining sample sizes to control the precession of the estimate, and controlling costs. Future work will need to explore how best to combine these considerations.

3.3 Evaluation

3.3.1 Comparison to Cohort Studies

To compare the results from a cross-sectional survey to those from a longitudinal study we could conduct a cohort study over a single year and at the end of the year take a random sample of the same population and apply the cross-sectional methodology. The two studies would be measuring slightly different quantities but both would serve as approximations to the yearly incidence rate. To make the estimates comparable we might fix the total cost.

While we do not have this type of data, we do have a selection of stored samples from three cohort studies: the HIV Prevention Trials Network (HPTN) 061 study (Koblin et al., 2013) the HPTN 064 study (Hodder et al., 2013) and HIVNET 001 (Celum et al., 2001)

which was also mentioned in Section 2.2. Table 3.1 shows incidence estimates based on MAA2, MAA3, and the standard longitudinal approach for these cohort studies.

	Cohort		
Approach	HPTN 064	HIVNET 001	HPTN 061
Longitudinal	$0.24\%(0.07, 0.62)$	$1.04\%(0.70, 1.55)$	$3.02\%(2.01, 4.37)$
Cross-sectional MAA2	$0.26\%(0.03, 0.95)$	$1.09\%(0.60, 1.84)$	$3.44\%(1.75, 6.20)$
Cross-sectional MAA3	$0.32\%(0.04, 1.17)$	$0.92\%(0.45, 1.73)$	$4.57\%(2.37, 8.24)$

Table 3.1: Comparison of incidence estimates between a longitudinal approach and two MAAs.

More details on the studies can be found in (Konikoff et al., 2013). The confidence intervals from the cross-sectional approaches are based on the exact approach, Formula (4.1), described later in Section 4.2. The point estimates are similar across the different methodologies for HPTN 064 and HIVNET 001. While the point estimates are very different in HPTN 061, the cross-sectional confidence intervals are extremely wide and do cover the longitudinal point estimate. This table is meant to show that the MAAs are in the correct ballpark. It should not be used to compare the two methods. The sampling was done for the cohort studies and not in a manner appropriate for the cross-sectional approach. Due to the time it takes to enroll people, the cohort studies occurred over more than one calendar year even though individuals were measured for at most one year of follow up. The cross-sectional estimates are calculated using uninfected individuals who made it to their last study visit and individuals who seroconverted. These events could occur at drastically different calendar times. There is a further assumption that everyone who seroconverted would have made it to the final study visit. This is certainly not true. Despite these limitations it is promising that the estimates are generally consistent.

3.3.2 Bias in MAAs

Since the mid 1990s estimates of the number of new infections in the United States have remained relatively constant at around 50,000 people per year (Hall et al., 2008). Over this time period the United States' population had increased almost linearly from around

265 million in 1996 to 310 million in 2010 (United States Census Bureau, 2014). Therefore the proportion of the population infected per year had been decreasing in an approximately linear fashion. This makes $g(t)$, which measures time *backward* from the present, a linearly increasing function. Thus it is reasonable to assume that in the United States $I_c \approx I(\psi)$ and our methodology should be applicable.

Yet, if the epidemic were to change course, for example because of a new highly successful prevention effort, we would hope that we could pick up this trend. We showed in Section 1.2 that the main difference between I_c and $I(\psi)$ will be

$$\frac{\int_0^M \left[\int_\psi^t g''(x)(t-x)dx\right] \phi(t)dt}{\int_0^M \phi(t)dt}.$$

As we have already estimated $\phi(t)$ we can calculate the difference between I_c and $I(\psi)$ for any $g(t)$.

Figure 3.6 shows a decreasing epidemic (Epidemic A), a sinuous epidemic where incidence both decreases and increases (Epidemic B), a linearly increasing epidemic (Epidemic C), and an exponentially increasing epidemic (Epidemic D). Figure 3.6 was created by plotting the function $g_{10}(10-\tau)$ against τ for $0 \leq \tau \leq 10$, where $g_{10}(t)$ gives the expected number of new infections per year, divided by the size of the population, t years *before* calendar time $\tau = 10$. The function is defined with the time t moving backward in keeping with the definition in Chapter 1. However, because it is easier to conceptualize time moving forward, Figure 3.6 plots $g_{10}(10-\tau)$ after we have followed each of the epidemics for 10 years. The vertical grey lines are included here for later reference.

Epidemic B could represent a population in which HIV first spreads in a subpopulation where it is remains isolated until it breaks into a secondary part of the population. Epidemics C and D show continually worsening epidemics which may be realistic if there is mixing in the population and interventions are not introduced. Since individuals can remain asymptomatic for many years this could occur early in an epidemic when public health officials have not yet realized the scope of the emerging situation. Epidemic A could represent an epidemic where

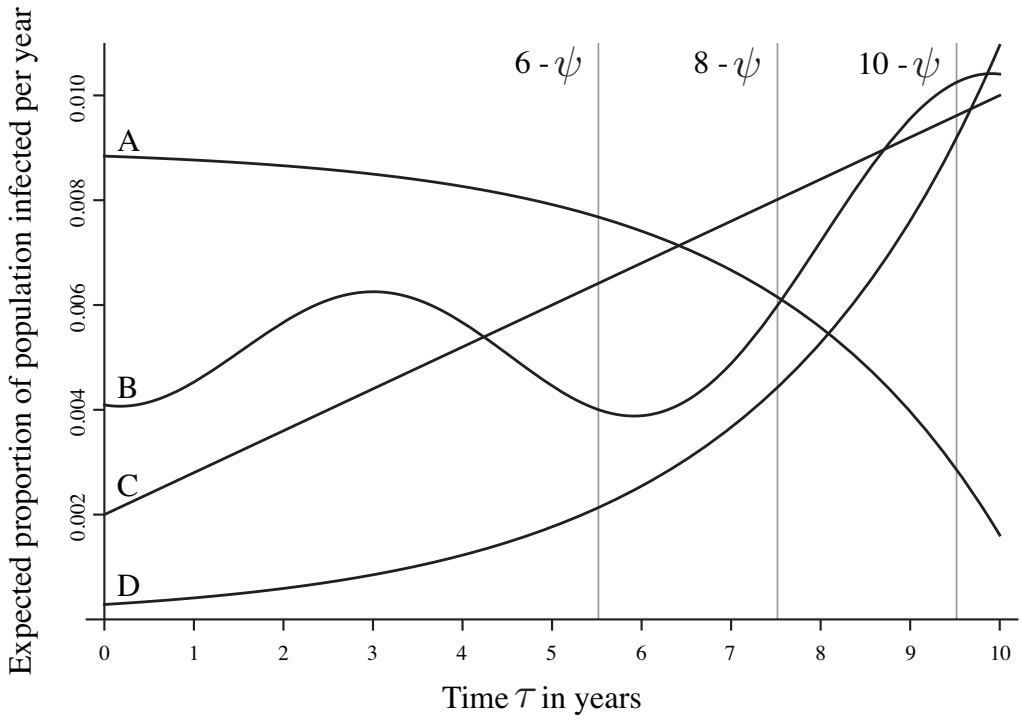

Figure 3.6: Hypothetical underlying epidemic curves for four simulated scenarios

antiretrovirals have successfully reduced the infectivity of the infected part of the population.

For each of these epidemics we calculated the relative error (RE) from using I_c to estimate $I(\psi)$ as

$$\text{RE} = \frac{I(\psi) - I_c}{I(\psi)} \approx \frac{g(\psi) - \text{E}[g(S)]}{g(\psi)}$$

where the approximation arises from the assumption that the prevalence of HIV ψ days ago is the same as the prevalence of HIV at the time of the survey. To visualize how this changes we plotted the approximate RE at each calendar time τ for years 7 to 10. Specifically we plotted

$$\frac{g_{10}(10 - \tau + \psi) - \mu^{-1} \int_0^M g_{10}(10 - \tau + x)\phi(x)dx}{g_{10}(10 - \tau + \psi)}$$

against τ.

Since $M < 7$ for MAA2, MAA3, and MAA4 we must specify $g_{10}(t)$ for at least 7 years before the survey occurs. As Epidemic C is linear the relative error is 0 regardless of the choice of MAA.

Figure 3.7 shows the approximate relative error for Epidemic A which is the declining epidemic. We see that MAA3, which had the $\phi(t)$ curve with heaviest tail, has the largest RE at all times. The RE is highest at $\tau = 10$ years. At this point in time $g(\psi) = 0.33\%$ whereas $\mathrm{E}[g(S)] = 0.29\%$. This MAA has relatively little bias despite the fact that the epidemic is nonlinear. The sooner $\phi(t)$ converges the smaller the bias for all three MAAs. MAA4, in orange, has almost no bias with $g(\psi)$ and $\mathrm{E}[g(S)]$ differing by less than three one thousandth of a percent over the entire time span in Figure 3.7.

Whether or not I_c under or overestimates $I(\psi)$ will depend on the concavity of $g(t)$. If $g(t)$ is concave up (convex) then $g(\psi) \leq \mathrm{E}[g(S)]$ by Jensen's inequality. The inequality is reversed if $g(t)$ is concave down. We thus have I_c underestimating $I(\psi)$ in Figure 3.7 since for Epidemic A $g(t)$ is concave down.

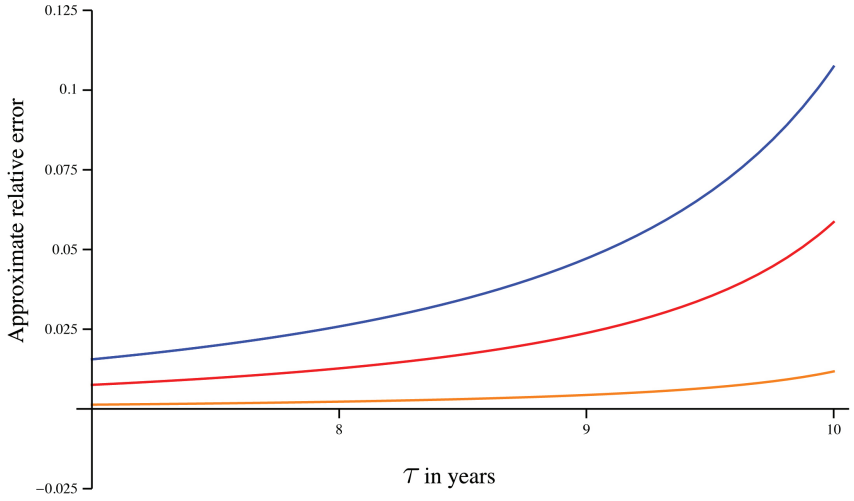

Figure 3.7: Approximate relative error for a cross-sectional survey conducted at $7 \leq \tau \leq 10$ in Epidemic A shown in Figure 3.6. MAA2 appears in red, MAA3 in blue, and MAA4 in orange. Colors match the MAAs shown in Figure 3.5.

We would expect that Epidemic D, which is exponentially increasing, would thus have I_c overestimate $I(\psi)$. We see this reflected in Figure 3.8 on the next page where the RE is always negative. As in Figure 3.7 the amount of bias is completely determined by the tails of the $\phi(t)$ curves. In this case the relative error does not change over the course of the epidemic.

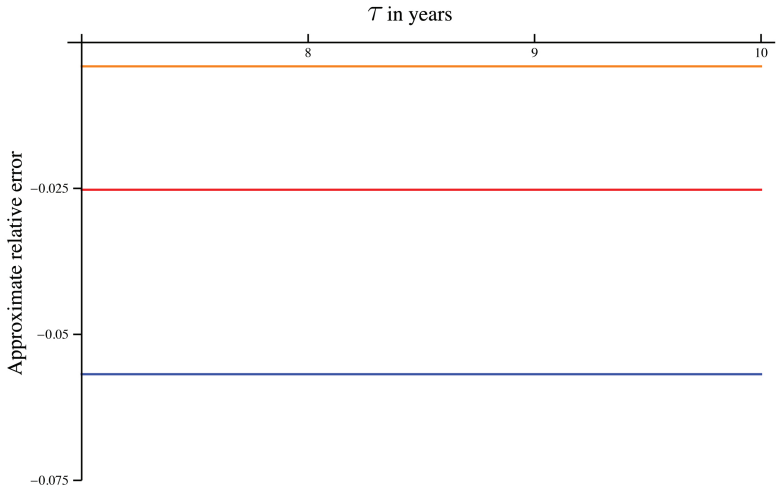

Figure 3.8: Approximate relative error for a cross-sectional survey conducted at $7 \leq \tau \leq 10$ in Epidemic D shown in Figure 3.6. MAA2 appears in red, MAA3 in blue, and MAA4 in orange. Colors match the MAAs shown in Figure 3.5.

In the special case where the epidemic follows an exponential curve of the form $k_1 \exp\{k_2 t\}$, for constants k_1 and k_2, the approximate relative error is independent of the calendar time τ as follows

$$\frac{g_{10}(10-\tau+\psi) - \mu^{-1}\int_0^M g_{10}(10-\tau+x)\phi(x)dx}{g_{10}(10-\tau+\psi)} =$$

$$\frac{k_1\exp\{k_2(10-\tau+\psi)\} - \mu^{-1}\int_0^M k_1\exp\{k_2(10-\tau+x)\}\phi(x)dx}{k_1\exp\{k_2(10-\tau+\psi)\}} =$$

$$\frac{k_1\exp\{k_2(10-\tau)\}\exp\{k_2\psi\} - \mu^{-1}k_1\exp\{k_2(10-\tau)\}\int_0^M \exp\{k_2 x\}\phi(x)dx}{k_1\exp\{k_2(10-\tau)\}\exp\{k_2\psi\}} =$$

$$\frac{\exp\{k_2\psi\} - \mu^{-1}\int_0^M \exp\{k_2 x\}\phi(x)dx}{\exp\{k_2\psi\}}.$$

There is no reason why I_c must only overestimate or only underestimate $I(\psi)$. Figure 3.9 shows how the changing concavity of $g(t)$ associated with Epidemic B in Figure 3.6 leads

to I_c first overestimating and later underestimating $I(\psi)$. In this example the epidemic is decidedly nonlinear except over very short amounts of time. This leads to the highest RE of the three example epidemics we have detailed. However even in this case, MAA2 and MAA3 perform fairly well. As in the previous two examples there is almost no bias associated with MAA4.

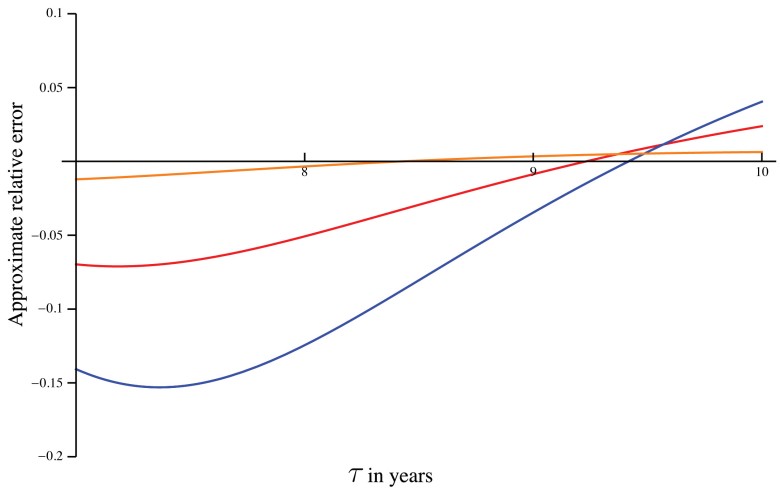

Figure 3.9: Approximate relative error for a cross-sectional survey conducted at $7 \leq \tau \leq 10$ in Epidemic D shown in Figure 3.6. MAA2 appears in red, MAA3 in blue, and MAA4 in orange. Colors match the MAAs shown in Figure 3.5.

We have shown that the bias associated with these MAAs is minimal in a variety of epidemic settings. Further, we can design MAAs, such as MAA4, for which approximating $I(\psi)$ by I_c will be almost exact. The reason MAA4 is not necessarily better than MAA2 or MAA3 relates to issues of cost and precision. We will look at issues of precision through the lens of sample size considerations in Section 4.2. In Chapter 4 we will also see the effect of the bias in the estimator on power calculations.

3.3.3 Limitation of the Brute Force Search

One concern when selecting an MAA based on it having the largest point estimate of μ, subject to certain criteria, is that we may be overestimating μ. One way to asses this potential bias is to repeat the search within each iteration of a bootstrap. In each iteration we find the MAA with the largest point estimate of μ, subject to certain criteria, and then

calculate the distance between that point estimate and the average value of each μ associated with that MAA across the bootstrap iterations. An estimator of the expected actual bias is then the average of these distances over all the bootstrap iterations. Here we are using the fact that the bootstrap is supposed to draw hypothetical samples from the given sample to inform us of the relationship between the true sample and the population. Thus, the average value of μ in the bootstraps should estimate the true population value of μ and the average distance from this value should approximate the bias. The procedure is outlined below.

---

1. For each $\text{MAA}^{(i)}$ and bootstrap iteration j calculate μ_{ij} where each i corresponds to a set of component assays and cutoffs.

2. For each j find which $\text{MAA}^{(i)}$ has the largest μ_{ij} for the specific bootstrap iteration. $\text{MAA}^{(i)}$ must also meet further criteria such as those outlined at the beginning of Section 3.1. The μ associated with this top MAA at iteration j is denoted μ_{i^*j}.

3. Calculate $\sum_j(\mu_{i^*j} - \bar{\mu}_{i\cdot})/B$ where there are B iterations of the bootstrap and for each $\text{MAA}^{(i)}$ we let $\bar{\mu}_{i\cdot} = \sum_j \mu_{ij}/B$

For the search which found MAA2 this procedure estimated the bias to be approximately 3 days. We concluded that this potential bias was not particularly concerning.

3.3.4 Cost of Different MAAs

We have mentioned before the higher cost of measuring CD4 cell count compared to using a serologic assay. In fact, CD4 cell count costs about 2.5 times as much as the LAg Avidity assay or the BioRad Avidity assay. The two serologic assays have similar costs. Measuring HIV viral load is even more expensive than measuring CD4 cell count. It costs about 5 times as much as the serologic assays. Using the 1780 samples for which we had complete data on all four biomarkers we evaluated the relative cost of using MAA3 instead of MAA2. We

applied each MAA to the data allowing for the order of the assays to change. For example, we looked at MAA3 allowing for the BioRad Avidity assay to be either the first or second component assay used. For MAA2 we required the first biomarker to be CD4 cell count since it must be measured at the time of sample collection. The optimal ordering for the assays is shown in Figure 3.10.

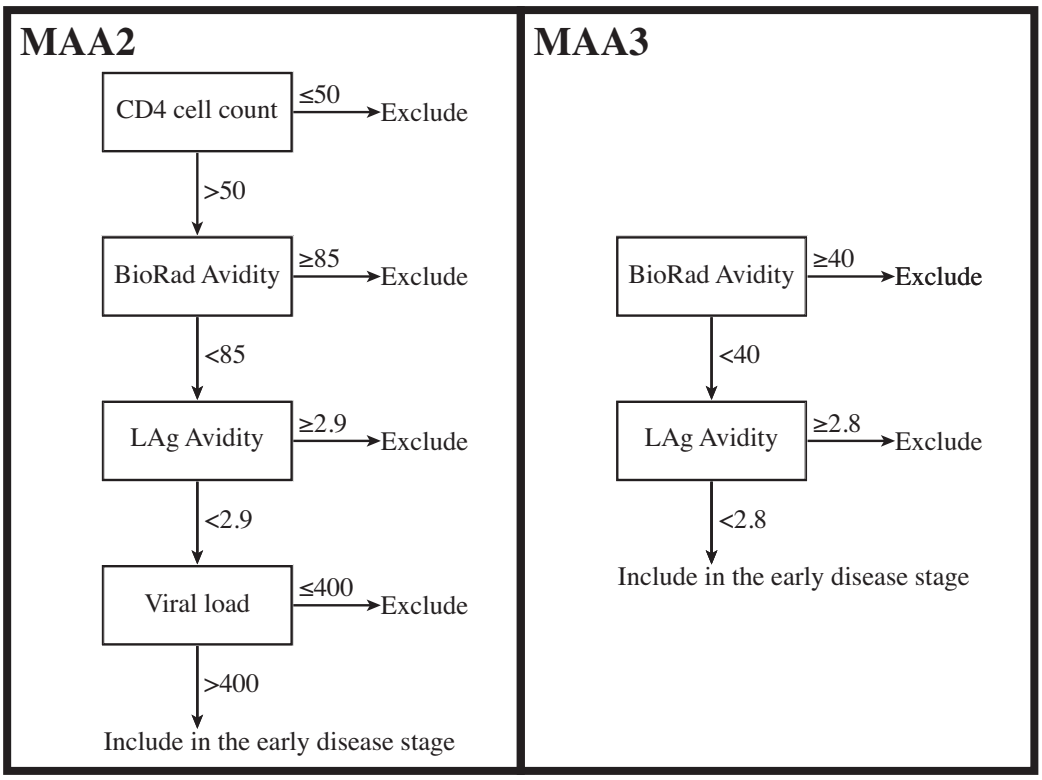

Figure 3.10: Cost efficient ordering of MAA2 and MAA3. The following units are used for the component assays: LAg Avidity: ODn; BioRad Avidity assay: percentage (avidity index); CD4 cell count: cells/mm^3; viral load: copies/mL.

We can see how MAA2 places the most expensive biomarker last so that it must be evaluated for the fewest number of samples. However, since CD4 cell count must be measured first MAA2 will be significantly more expensive than MAA3. We found it would cost 25% less to use MAA3 on the same number of samples as MAA2. This comparison does not account for the fact that fewer samples will be needed when using MAA2 based on the considerations discussed shortly in Chapter 4. Since the underlying incidence affects the needed sample sizes the relative cost of using each MAA will depend on the epidemic. As a rough approximation we can use the fact that the sample sizes are approximately proportional to the ratio of the

μ associated with each MAA (Brookmeyer, 2010b). In this case, since we estimated μ for MAA2 as 0.40 years and μ for MAA3 as 0.33 years, we will need 0.40/0.33 more samples if we use MAA3. This adjustment increases the relative cost of MAA3 slightly so that it is 31% of MAA2. Thus MAA3 is likely to be about three times cheaper than MAA2. Admittedly this analysis only factors in the cost of each assay and not the implementation. However, since MAA3 does not require immediate testing of blood samples it is likely preferable than MAA2 both in the cost of implementation and the cost of running the assays. The tradeoff here is the additional bias in MAA3 compared to MAA2 described in 3.3.2.

3.3.5 Evaluation of Confidence Intervals

Since we rely on the clustered bootstrap to quantify uncertainty in our estimates we may ask whether the method we are using to generate confidence intervals achieves its stated coverage. One way to estimate the coverages of the confidence intervals is to perform a simulation. We assumed that the true population is composed of an infinite number of identical replicates of our sample. We then drew a sample, S_1, from the population which is equivalent to performing a single bootstrap iteration. We next pretended we could only see S_1 and performed the clustered bootstrapping procedure on S_1 to calculate CIs for μ and ψ. Since we knew the entire population we checked whether or not the confidence intervals truly covered μ and ψ. The process was then repeated by drawing samples S_i for $2 \leq i \leq 1,000$. This gave an estimate of the true coverage probabilities of the method. Bootstrapping within the bootstrap is known as double bootstrapping (Davison and Hinkley, 1997).

Unlike the percentile method which only requires a single bootstrap to estimate a confidence interval, other bootstrap confidence intervals depend on performing two layers of bootstrapping. To simulate these methods we performed three levels of bootstrapping. The top level can be viewed as 1,000 simulations. In each round of the simulation we conducted a double bootstrap. The bootstrapping itself followed the same procedure as outlined in Section 2.3 but the number of iterations was increased to 1,999. Bootstrapping, in its derivation, assumes that all possible samples will be drawn. For computational reasons we often sample

with replacement a fixed number of times. The idea is that when the number of iterations is large it will approximate the results from sampling all possible combinations. The larger the number of iterations the less error we introduce through this approximation.

As mentioned in Section 2.3 if no samples are infected more than two years the term associated with the knot separates the data perfectly. Instead of assuming $M < 2$ in these cases, as we did above, we removed the knot for this simulation. The difficulty arises because we are trying to fit a general regression model to $1000 \cdot 1999 \cdot 1999 \approx 4$ billion different samples.

Table 3.2 presents the results from applying the above procedure while using MAA2.

Parameter	**A**	**B**	**C**
μ	94.9% (93.3, 96.2)	95.3% (93.8, 96.5)	94.6% (93.0, 95.9)
ψ	91.1% (89.2, 92.8)	93.5% (91.8, 94.9)	93.1% (91.3, 94.6)
	D	**E**	**F**
μ	94.9% (93.3, 96.2)	94.5% (92.9, 95.8)	95.3% (93.8, 96.5)
ψ	95.8% (94.4, 97.0)	94.6% (93.0, 95.9)	95.7% (94.3, 96.9)

Table 3.2: Percent of confidence intervals which covered the true parameter. **A**: Basic bootstrap confidence interval from a single bootstrap. **B**: Percentile confidence interval from a single bootstrap. **C**: Bias-corrected confidence interval from a single bootstrap. **D**: Bootstrap-t (studentized) confidence interval with variances calculated from a double bootstrap. **E**: Basic bootstrap CI from a double bootstrap. **F**: Percentile confidence interval from a double bootstrap. 95% confidence intervals are presented next to the coverage probabilities.

The different types of bootstrap confidence intervals are all outlined in (Davison and Hinkley, 1997). The 95% confidence intervals for the coverage probability are calculated based on the assumption that the number of intervals covering the true parameters is a binomial draw. Confidence intervals were then based on the normal approximation to the binomial.

We see that the highest coverage comes from the percentile method applied to a double bootstrap. Based on the 95% CIs it appears that methods relying only on a single bootstrap, shown in the top half of the table, are anti-conservative when it comes to the shadow. The techniques in the top half of the table are considered "first-order" accurate whereas the techniques in the bottom half of the table are considered "second-order" accurate (Davison and Hinkley, 1997).

Techniques **C** through **F** are all meant to control for the fact that the basic and percentile bootstrap confidence intervals may be biased. We can see some of this bias by examining the confidence intervals presented for ψ in Chapters 2 and 3. The confidence intervals are not symmetric around our point estimate $\hat{\psi}$. In fact the average estimate of ψ produced by bootstrapping is not equal to $\hat{\psi}$ for these algorithms. The bootstrap is supposed to draw hypothetical samples from the given sample to inform us of the relationship between the true sample and the population. Thus, if there were no bias in the bootstrap we would expect to, on average, return our point estimate for the parameter. Techniques **C** through **F** are different methods for estimating and adjusting for this apparent bias. Unlike the confidence intervals for ψ the confidence intervals for μ in Chapters 2 and 3 are approximately symmetric around $\hat{\mu}$. Since there is little bias in the method we do not see the same gains in the confidence intervals for μ as we did with the confidence intervals for ψ.

CHAPTER 4

Sample Size Methods

Much of the content of this chapter has been accepted for publication by *Biometrics*. The article, Sample Size Methods for Estimating HIV Incidence from Cross-Sectional Surveys, as well as code for implementing these methods in R are available at the *Biometrics* website on Wiley Online Library.

In what follows we will assume that a preliminary estimate of the HIV prevalence, p_0, is available immediately before the cross-sectional survey is conducted. Estimating HIV prevalence is a simpler undertaking than estimating HIV incidence (Brookmeyer, 2010a). We will structure the problem so that researchers can solve for $u = n(1-p)$ where n is the sample size and p is the true HIV prevalence. We call u the required number of uninfected samples since $n(1-p)$ is the expected value of N_u, the number uninfected people sampled. Researchers may solve for u and wait to solve for n until p_0 becomes available. This is particularly helpful when setting the power for consecutive surveys as it allows researchers to wait until immediately before the second survey to supply an estimate of the HIV prevalence at the second time point.

4.1 Accounting for Uncertainty in μ in the Distribution of X

In this section we derive sample size methods for estimating incidence with a desired degree of precision. Recall, the cross-sectional incidence estimator used in the literature is $\widehat{I}_c = X(N_u\mu)^{-1}$, where X is the number of individuals marked in the early disease stage. X is a random variable whose realizations after the survey is conducted is denoted x.

If we conceptualize the survey as splitting the n samples into those which are, and

are not, in the early disease stage, then $X \sim \text{binomial}(n, \pi)$ where, as above, π is the probability of being in the early disease stage at the time of the cross-sectional survey. Since $I_c = \pi\{(1-p)\mu\}^{-1}$ we have that $\pi = (1-p)\mu I_c$. Therefore $X \sim \text{binomial}(n, (1-p)\mu I_c)$. The "success" probability will be small since the vast majority of people in a population will be uninfected and most of the infected individuals will not appear in the early disease stage. We therefore apply a Poisson approximation to the binomial distribution and let $X \sim \text{Poisson}(n(1-p)\mu I_c)$ so that $X \sim \text{Poisson}(u\mu I_c)$.

An appealing feature of this derivation is that it holds regardless of whether individuals enter the early disease stage before seroconverting or after seroconverting. It therefore harmonizes this work with previous work Brookmeyer conducted on calculating confidence intervals for I_c (Brookmeyer, 1997). In that work he had suggested letting $X \sim \text{Poisson}(n_u \mu I_c)$ under a different derivation.

When we estimate the incidence as $x(n_u\mu)^{-1}$ we are implicitly assuming that μ is known. However, μ is never known exactly and must be estimated for each specific MAA. We can incorporate our uncertainty by placing a distribution, called $h(\mu)$ on μ. One suggestion is to let

$$h(\mu) = \frac{\mu^{\gamma-1} \exp^{-\frac{\mu}{\beta}}}{\Gamma(\gamma)\beta^{\gamma}}$$

which is a Gamma distribution with parameters γ and β (Brookmeyer, 1997). Under this assumption we have that

$$\text{P}(X = x) = \frac{\Gamma(\gamma + x)}{\Gamma(\gamma)x!} \left(\frac{\beta u I_c}{\beta u I_c + 1}\right)^x \left(\frac{1}{\beta u I_c + 1}\right)^{\gamma}.$$

This is the result of the fact that a Gamma-Poisson mixture follows a negative binomial distribution (Agresti, 2013). The Gamma distribution thus simplifies the remaining calculations but the methodology can be generalized to any distribution on μ. In particular, in Chapter 6, we will look at a case where we estimate the posterior distribution of μ within a Bayesian framework. The maximum likelihood estimate for I_c is given by $\frac{x}{u\gamma\beta}$. Noting that $\text{E}[\mu]=\gamma\beta$ in the Gamma distribution above and that $\text{E}[N_u]=u$ we see the consistency with the cross-sectional incidence estimator $\widehat{I_c} = X(N_u\mu)^{-1}$. We will shortly examine different

values of γ and β.

4.2 Methods for a Single Cross-Sectional Survey

We now derive the sample size needed to achieve an estimate of incidence with a desired amount of precision. The calculations will depend on a preliminary estimate of incidence, I_0, because the underlying incidence dictates the number of samples which will be classified in the early disease stage. The smaller the underlying incidence the larger the sample size we will need to capture a sufficient number of people in the early disease stage. To make this explicit in our calculations we will focus on controlling the width of the confidence interval divided by the true underlying incidence.

We will use the exact $100(1-\alpha)\%$ confidence interval for the "success" probability of a negative binomial experiment (Casella and Berger, 2002). In the negative binomial distribution above the "success" probability is $(\beta u I_c+1)^{-1}$ and the corresponding $100(1-\alpha)\%$ confidence interval is given by $\{\text{Beta}(\alpha/2, \gamma, x+1), \text{Beta}(1-\alpha/2, \gamma, x)\}$, where $\text{Beta}(q, a, b)$ is the q^{th} quantile of a Beta distribution with parameters a and b. We may transform the endpoints of the confidence interval for $(\beta u I_c + 1)^{-1}$ to get that

$$\left[\left\{\frac{1}{\text{Beta}(1-\alpha/2, \gamma, x)} - 1\right\}\Big/\beta u, \left\{\frac{1}{\text{Beta}(\alpha/2, \gamma, x+1)} - 1\right\}\Big/\beta u\right]$$

is a $100(1-\alpha)\%$ confidence interval for I_c. Since x is unknown before sampling we replace x with $\text{E}[X] \approx u I_0 \gamma \beta$ to get

$$\left[\left\{\frac{1}{\text{Beta}(1-\alpha/2, \gamma, u I_0 \gamma \beta)} - 1\right\}\Big/\beta u, \left\{\frac{1}{\text{Beta}(\alpha/2, \gamma, u I_0 \gamma \beta + 1)} - 1\right\}\Big/\beta u\right] \tag{4.1}$$

Then to achieve a $100(1-\alpha)\%$ confidence interval of width W we solve for u in

$$W = \left[\left\{\frac{1}{\text{Beta}(\alpha/2, \gamma, u I_0 \gamma \beta + 1)} - 1\right\} - \left\{\frac{1}{\text{Beta}(1-\alpha/2, \gamma, u I_0 \gamma \beta)} - 1\right\}\right]\Big/\beta u.$$

We will assume that for any given algorithm γ and β are known constants estimated

from previous work. As an illustrative example we will use MAA2 described in Chapter 3 for which we estimated that $\gamma = 120.77$ and $\beta = 0.003$ by fitting a Gamma distribution to a sampling distribution of μ generated by bootstrapping (Konikoff et al., 2013). Code for determining the required sample size for any W and I_c is available at the Wiley Online Library. Alternatively, given a preliminary estimate of incidence, I_0, and a fixed sample size the anticipated confidence interval can be determined. For any given I_0 a practitioner can set the precision, determine the corresponding sample size, and then solve for the anticipated confidence interval implied by the original choice of W. That is once we have solved for u in

$$W/I_0 = \left[\left\{\frac{1}{\text{Beta}(\alpha/2, \gamma, uI_0\gamma\beta + 1)} - 1\right\} - \left\{\frac{1}{\text{Beta}(1 - \alpha/2, \gamma, uI_0\gamma\beta)} - 1\right\}\right] \Big/ \beta u I_0 \quad (4.2)$$

we can obtain the implied confidence interval by calculating (4.1) for u and I_0. Repeating this process for various levels of precision, that is choices of W/I_0, may provide intuition to a researcher who is comfortable specifying I_0 but unsure how to specify the precision and would rather control the margin of error for the asymmetric confidence interval.

Table 4.1 which gives the required uninfected sample sizes needed to achieve a set W/I_c for a 95% confidence intervals for selected values of W and I_c.

I_c	W/I_c						
	0.50	1	1.5	2	2.5	3	3.5
0.25%	141,895	20,768	9,115	5,338	3,602	2,648	2,060
0.50%	70,948	10,384	4,558	2,669	1,801	1,324	1,030
1.00%	35,474	5,192	2,279	1,335	901	662	515
1.50%	23,650	3,462	1,520	890	601	442	344
2.00%	17,737	2,596	1,140	668	451	331	258
2.50%	14,190	2,077	912	534	361	265	206
3.00%	11,825	1,731	760	445	301	221	172
4.00%	8,869	1,298	570	334	226	166	129
5.00%	7,095	1,039	456	267	181	133	103

Table 4.1: Uninfected samples needed for a desired precision. Results are for MAA2 defined in Chapter 3. I_c is the underlying cross-sectional incidence and W is the width of the 95% confidence interval.

The higher the desired precision the larger the required number of uninfected samples

we will need. For example, if the underlying incidence is 0.25% per year 2,060 uninfected samples are needed to achieve $W/I_c = 3.5$, 20,768 uninfected samples are needed to achieve $W/I_c = 1$, and 141,895 uninfected samples are needed to achieve $W/I_c = 0.5$. Since it is much easier to measure larger incidences only 7,095 uninfected samples are needed to achieve $W/I_c = 0.5$ if the underlying incidence were 5% per year instead of 0.25% per year. In fact, we may note that in (4.2) everywhere the sample size u appears it is multiplied by the incidence I_0. Thus the sample size is inversely proportional to the underlying incidence and, for example, if incidence is halved the sample size will need to be doubled. This point can be seen by comparing the first two rows of Table 4.1. It is clear from Table 4.1 that the number of samples needed can become prohibitive when a high degree of precision is required and the incidence is low.

Since this methodology is applicable to other biomarker-based testing algorithms we examine three other hypothetical situations illustrated in Figure 4.1. Rather than compare MAA2 to MAA3 and MAA4 we keep either the mean or the variance of the gamma distribution fixed and change the other. Since the Gamma distribution is determined by its first two moments we can examine how differences in the distribution of μ affect the required sample sizes through this more general setup.

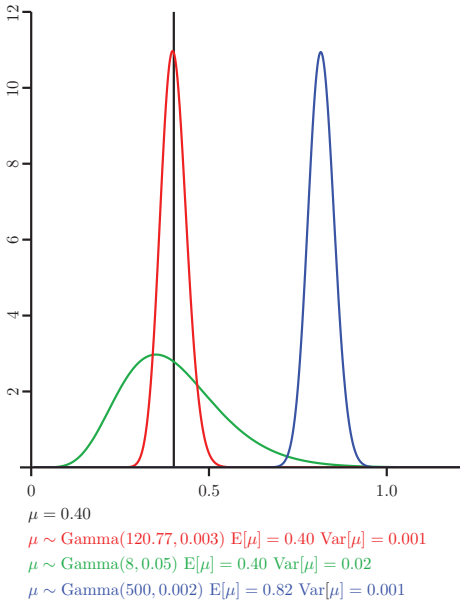

Figure 4.1: Distributions of μ used in sample size calculations.

In the first situation (black line) we ignore uncertainty in μ, in the second (green curve) we increase the variability in μ, and in the third we change the Gamma distribution so that μ is centered at a larger mean (blue curve). In the first two alternative scenarios μ is centered at the same mean as MAA2 from Chapter 3. In the third scenario we extend the amount of time, on average, people spend in the early disease stage.

Figure 4.2 plots W/I_c for 95% confidence intervals against the required number of uninfected samples for the scenarios illustrated in Figure 4.1. The underlying incidence in Figure 4.2 was set at 1% per year which might occur in a high risk subpopulation of the United States.

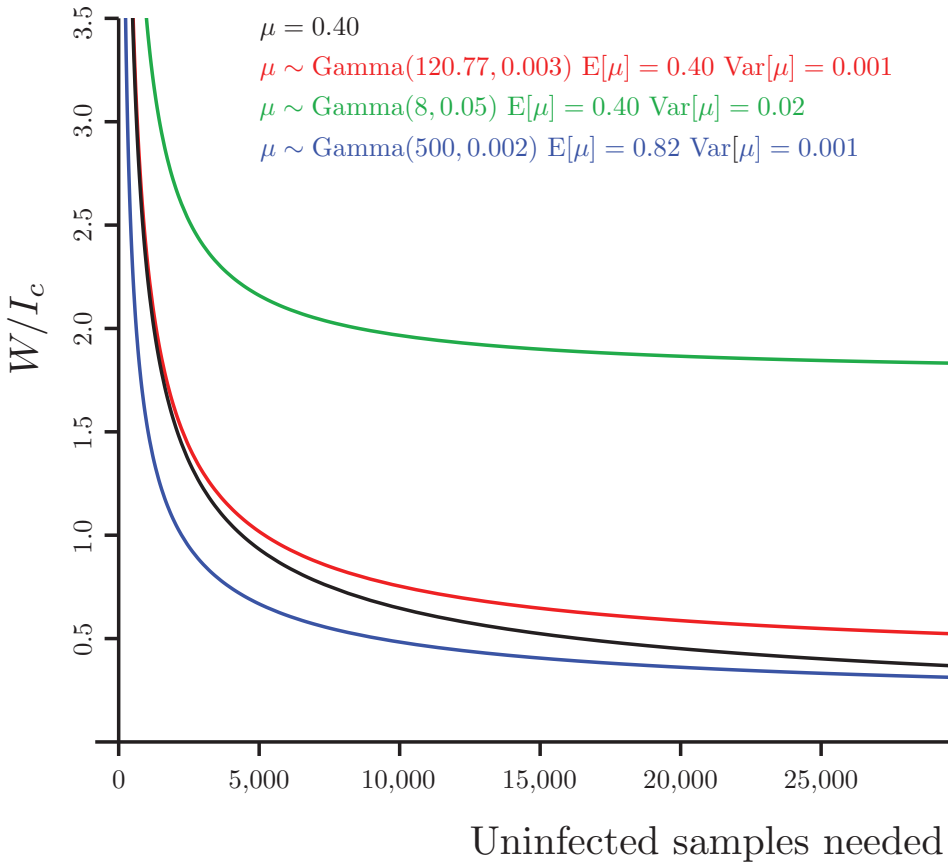

Figure 4.2: W/I_c achieved per number of uninfected samples when $I_c = 1\%$ dependent on distributions of μ in Figure 4.1.

All else equal, smaller sample sizes are needed when individuals spend more time, on average, in the early disease stage. This rule can be seen when comparing the blue and red curves. When we instead fix the mean of the Gamma distributions we can see how

additional samples are needed to account for the increased uncertainty by comparing the black, red, and green curves. The difference between the black curve and the red curve shows the extent of under sampling that would occur if we ignored our uncertainty in μ while using MAA2. While these curves are relatively close we may notice how they separate as the desired precision increases. In general differences in the distributions are magnified when the situation demands larger sample sizes such as when higher precision is required or when the underlying incidence is small. When we compare the green curve, with additional variability, to the red and black curves we see that no sample size can make up for too much uncertainty in μ. It is clear that a $W/I_c < 1.5$, something achievable with under 5,000 uninfected samples in the other cases, is prohibitive in this case. This emphasizes the importance of accurately estimating μ and highlights the significance of properly accounting for our uncertainty.

4.3 Sample Size Methods for Detecting Changes in Incidence

We now move from a single cross-sectional survey to two consecutive cross-sectional surveys where the aim is to determine if incidence has changed. This is an important question both in surveillance of the epidemic and when conducting prevention trials. The surveys are conducted at calendar times t_1 and t_2 and estimate incidences I_1 and I_2. We drop the subscript c for convenience. We will concern ourselves with testing the null hypothesis of no difference in incidence against the alternative hypothesis that incidence has increased by some amount. More formally, we wish to test H_0: $I_2/I_1 = 1$ against an alternative that H_A: $I_2/I_1 = r > 1$. For any specific alternative, r, we will derive the needed sample sizes, n_1 and n_2, to have a high power of detecting this change if it occurred. As above, the sample sizes will be based on the number of uninfected samples, u_1 and u_2, needed at each survey but the sample sizes, n_1 and n_2, should be fixed by using preliminary estimates of prevalence at the two time points. While the calculations below are for a one-sided test, the power calculation for a 2-sided test can be conducted by halving the α-level, as under a given alternative hypothesis the probability of rejecting the null hypothesis in the opposite

direction will be negligible. If we wish to look for a decrease in incidence the roles of I_1 and I_2 can be interchanged.

In order to calculate the sample size needed for each of the two cross-sectional surveys we extend previous work which derived sample size formulas when X follows a Poisson distribution with known mean (Gail, 1974; Brown and Green, 1982). We will do this by rewriting P(Reject $H_0|H_A$) as $\sum_{t=0}^{\infty}$P(Reject $H_0|H_A, T = t)\times$ P($T = t|H_A$) where T is the total number of individuals found in the early disease stage from both surveys.

Specifically, as in the one sample case, we assume that $X_l|\mu \sim$ Poisson$(u_l\mu I_l)$ for $l = 1, 2$. Then since the surveys are independent we have that $T = X_1 + X_2 \sim$ Poisson$[(u_1I_1 + u_2I_2)\mu]$ conditionally on knowledge of μ. This implies that if we further condition on the observed $T = t$, we get

$$X_2|\mu, t \sim \text{Binomial}\left(t, \frac{u_2I_2\mu}{(u_1I_1 + u_2I_2)\mu}\right).$$

We may notice that μ drops out when we condition on t because it affects the two samples equally. Thus,

$$X_2|t \sim \text{Binomial}\left(t, \frac{u_2I_2}{u_1I_1 + u_2I_2}\right).$$

We may now calculate the power for any value $T = t$ given a specific alternative hypothesis r. We do not require the sample sizes at the two time points to be equal. However, one must set the ratio of the number of uninfected samples required at the two time periods, $s = u_2/u_1$, before carrying out the power calculation. Note that under the null hypothesis we have

$$X_2|t \sim \text{Binomial}\left(t, \frac{u_2}{u_1 + u_2}\right)$$

or

$$X_2|t \sim \text{Binomial}\left(t, \frac{s}{1 + s}\right)$$

and under the alternative hypothesis

$$X_2|t \sim \text{Binomial}\left(t, \frac{rs}{1 + rs}\right).$$

The rejection rule under the null hypothesis is to reject H_0 if x_2 is greater than or equal to the smallest cutoff, C, such that

$$\sum_{k=C}^{t} \binom{t}{k} \left(\frac{s}{1+s}\right)^k \left(\frac{1}{1+s}\right)^{t-k} \leq \alpha,$$

where α is the desired size of the one sided test. This cutoff, C, can be found in statistical packages such as R. Next we calculate

$$\sum_{k=C}^{t} \binom{t}{k} \left(\frac{rs}{1+rs}\right)^k \left(\frac{1}{1+rs}\right)^{t-k}$$

which is P(Reject $H_0|H_A, T = t$), the probability of rejecting the null hypothesis when the alternative r is true for a fixed t.

However before the sample is conducted, T is not fixed but is random. Therefore we must also calculate $\mathrm{P}(T = t|H_A)$. Our uncertainty in μ implies that T is a Gamma-Poisson mixture which is, as above, negative binomial. To see this, notice that T follows a Poisson with mean $u_1(I_1 + sI_2)\mu \sim \text{Gamma}(\gamma, \beta u_1(I_1 + sI_2))$. Therefore we have that

$$\mathrm{P}(T = t|H_A) = \frac{\Gamma(\gamma + t)}{\Gamma(\gamma)t!} \left(\frac{\beta u_1 I_1(1+rs)}{\beta u_1 I_1(1+rs)+1}\right)^t \left(\frac{1}{\beta u_1 I_1(1+rs)+1}\right)^\gamma.$$

Then P(Reject $H_0|H_A$) $= \sum_{k=0}^{\infty}$P(Reject $H_0|H_A, T = t)\times$P$(T = t|H_A)$ which we may write as

$$\sum_{t=0}^{\infty} \left[\left\{ \sum_{k=C}^{t} \binom{t}{k} \left(\frac{rs}{1+rs}\right)^k \left(\frac{1}{1+rs}\right)^{t-k} \right\} \frac{\Gamma(\gamma + t)}{\Gamma(\gamma)t!} \left(\frac{\beta u_1 I_1(1+rs)}{\beta u_1 I_1(1+rs)+1}\right)^t \left(\frac{1}{\beta u_1 I_1(1+rs)+1}\right)^\gamma \right]$$

We see that power is a function of the alternative r, the desired ratio of the two samples sizes s, and $u_1 I_1$ (or alternatively $u_2 I_2$). Thus for any specific α level, and desired power, we may solve for $u_1 I_1$ for given r and s.

Table 4.2 shows the necessary $u_1 I_1$ and in parenthesis $(u_1 + u_2)I_1$ to achieve 90% power for various values of r and s. For a fixed value of s the rows correspond to the black, red, green, and blue distributions of μ in Figure 4.1.

	$r = 1.2$	$r = 1.5$	$r = 2$	$r = 3$	$r = 5$
$s = 2$					
i	898.7 (2696.0)	166.4 (499.0)	50.5 (151.5)	17.0 (50.8)	6.2 (18.6)
ii	907.7 (2723.1)	168.2 (504.5)	51.1 (153.2)	17.2 (51.4)	6.3 (18.8)
iii	1049.0 (3147.0)	195.2 (585.6)	59.6 (178.7)	20.2 (60.4)	7.5 (22.3)
iv	442.8 (1328.2)	82.0 (246.0)	24.9 (74.7)	8.4 (25.1)	3.1 (9.2)
$s = 1$					
i	1196.6 (2393.1)	221.2 (442.3)	67.4 (134.8)	22.2 (44.4)	8.2 (16.3)
ii	1208.8 (2417.5)	223.5 (447.0)	68.1 (136.2)	22.5 (44.9)	8.3 (16.5)
iii	1396.2 (2792.3)	259.2 (518.4)	79.0 (158.0)	26.3 (52.6)	9.8 (19.6)
iv	589.6 (1179.1)	109.0 (218.0)	33.2 (66.4)	11.0 (21.9)	4.1 (8.1)
$s = \frac{1}{2}$					
i	1793.3 (2689.9)	331.1 (496.6)	100.4 (150.6)	33.8 (50.6)	12.1 (18.1)
ii	1811.0 (2716.5)	334.3 (501.5)	101.5 (152.2)	34.1 (51.1)	12.2 (18.3)
iii	2090.3 (3135.4)	386.6 (579.8)	117.7 (176.5)	39.3 (59.0)	14.1 (21.1)
iv	883.4 (1325.1)	163.1 (244.7)	49.5 (74.2)	16.7 (25.0)	6.0 (8.9)

Table 4.2: Needed uninfected samples at the first time point multiplied by initial incidence, $u_1 I_1$, and, in parenthesis, total uninfected samples needed multiplied by the incidence at the first time. (i) $\mu = 0.40$, (ii) $\mu \sim$ Gamma (120.77, 0.003), (iii) $\mu \sim$ Gamma (8, 0.05), (iv) $\mu \sim$ Gamma (500, 0.002), r is ratio of the incidence at the second time point to the incidence at the first time point when the alternative hypothesis is true, and s is the ratio of the needed uninfected samples at the second time point to the needed uninfected samples at the first time point.

This table shows the required number of uninfected samples in the first survey multiplied by the incidence at time t_1 as well as the total uninfected samples needed from both surveys multiplied by the incidence at time t_1. For example, if a researcher using MAA2 wants to have 90% power of detection for a doubling of the incidence at time t_2 compared to the incidence at time t_1 (taking $s = 1$ so that $u_1 = u_2$) the researcher would need to sample so that $u_1 I_1$=68.1. If the incidence at time t_1 is 1% per year this would imply that $u_1 = u_2 = 6,810$ uninfected individuals are required at both time points. These numbers can then be adjusted to fix the total sample sizes n_1 and n_2 based on preliminary estimates of prevalence. Code for determining the sample sizes in other situations is available at the Wiley Online Library.

When we compare the rows marked i and ii, we see that as in the one sample case, accounting for the uncertainty in μ when using MAA2 increases the required sample sizes. The effect is most pronounced when the underlying incidence is low. Since Table 4.2 is

explicitly in terms of the incidence this point is more directly understood than in the previous section. We again notice that the sample sizes are inversely proportional to the underlying incidences. If we look at the row marked iii, where μ has a higher variance, we see that much larger sample sizes will be required. We can also see that the smaller the change in incidence we wish to detect (the closer r is to 1) the larger the sample size we will need. Table 4.2 combines these two effects to illustrate the tremendous cost associated with measuring small changes in incidence when the underlying incidence is low. A further similarity to the single survey case can be seen by looking at the rows marked iv. We see that larger values of μ again imply smaller sample sizes.

4.4 Evaluation

In order to test our methodology in a variety of underlying epidemics we created a simulation. The initial phase of each epidemic is the same for the first 730 days. We let $50e^{(d/300)}$ infections occur each day for $1 \leq d \leq 730$ days. The number of infections is rounded to the nearest whole number whenever necessary. At $d = 730$ we let there be 1,000,000 uninfected individuals in the population. From here infections occur according to the four unique epidemic curves shown in Figure 3.6. We first looked at these epidemics in the context of potential bias in the cross-sectional incidence estimator. If we let $\tau = 0$ represent the beginning of these unique epidemic curves we examined changes in incidence from $\tau = 6 - \psi$ years to $\tau = 10 - \psi$ years and from $\tau = 8 - \psi$ years to $\tau = 10 - \psi$ years. Vertical grey lines denote these time points in Figure 3.6. The infections before $\tau = 0$ do not affect the number of individuals in the early disease stage at the three time periods we sample the population – at years 6, 8, and 10. In Chapter 3 we estimated that the probability an individual is found in the early disease converges to 0 by 5 years after infection for MAA2. That is we estimated $\phi(t) = 0$ for $t > 5$ (see Figure 3.5). The initial phase of the epidemic is relevant to the prevalence of the disease at the sample dates.

An overview of how the simulations were conducted can be seen in Figure 4.3. Births occur at a daily rate of 0.004 percent of the uninfected part of the population. Infections

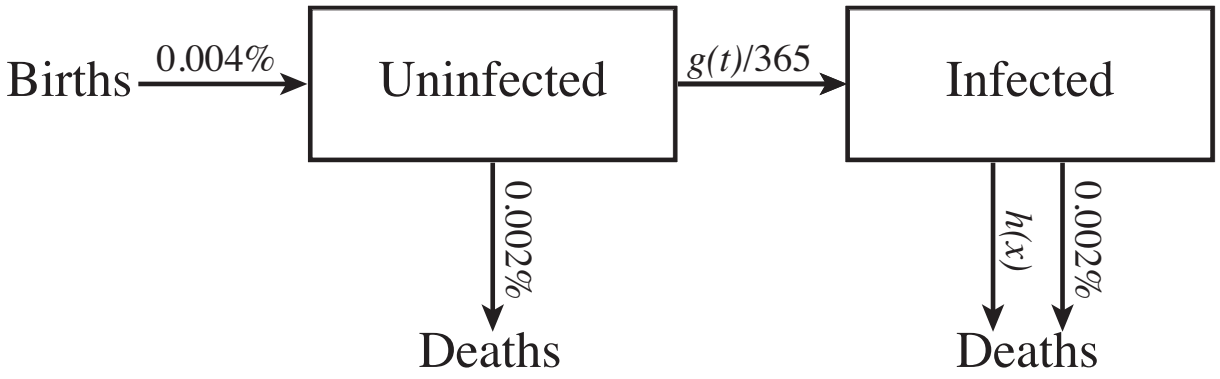

Figure 4.3: Daily epidemic changes. Births occur at a rate of 0.004 percent of the uninfected part of the population, deaths from causes other than HIV occur at a rate of 0.002 percent of the population, deaths specifically from HIV occur according to Weibull distribution with shape parameter 2.5 and scale parameter 14 so that $h(x)$ is $0.00000013x^{1.5}$ for x is in days, and the number of new infections in the population is determined by multiplying the size of the population by $g(t)/365$.

occur according to the curves in Figure 3.6. Upon infection each individual is assigned a death date (specifically caused by the virus) based on a draw from a Weibull distribution with shape parameter 2.5 and scale parameter 14. This sets the median survival at approximately 12 years past infection. All individuals are assigned a death date (independent of infection) based on a random draw from an exponential distribution with rate 0.00002 so that 0.002 percent of the population would die of other causes each day. Infected individuals thus receive two death dates upon infection and the earlier date is used.

We used the simulation to measure the power when we compare incidences at $6-\psi$ years and $10-\psi$ years and when we compare incidences at $8-\psi$ years and $10-\psi$ years. We fixed the alternative hypothesis r based on the true incidences ψ years before the sample dates and we let $s = 1$ for simplicity. The expected power for these studies was 0.90. We used the true underlying incidence at $6-\psi$ years (for the first comparison) and the true underlying incidence at $8-\psi$ years (for the second comparison) to determine the needed uninfected sample sizes based on the methods developed in Section 4.3. These numbers are shown in Table 4.3. We fixed the total number of samples at each time point based on the true prevalence at the time of the survey.

When we sampled the population we separated the uninfected individuals from the in-

fected individuals. The infected individuals were further "tested" with our biomarker-based testing algorithm by taking a binomial draw where the probability of being in the early disease stage was determined by the estimated $\phi(t)$ function corresponding to MAA2 from Chapter 3. We adjusted the estimate by the probability of surviving to the time of the survey once infected with the virus. This had a minimal effect changing, for example, $\widehat{\psi}$ from 0.49 to 0.48 years.

We then calculated our estimates of the incidences at the two time points as well as whether or not we would have rejected the null hypothesis. We repeated this procedure 1,000 times. To account for uncertainty in μ we bootstrapped the sample we used to estimate the $\phi(t)$ function 999 times so that each time the true value of μ changed (see Chapter 1). We calculated the power as the number of times we rejected H_0 over the 999,000 total runs. The calculated power for all four epidemics is displayed in Table 4.3. In each comparison we also calculated the anticipated 95% confidence intervals for both incidences being measured based on Formula (4.1) in Section 4.2.

	True Incidence	ASI	CI coverage	Power	u
(1)					
Epidemic A	0.80% and 0.30%	0.80% and 0.29%	0.94 and 0.92	0.92	9,984
Epidemic B	0.41% and 1.08%	0.42% and 1.05%	0.95 and 0.95	0.87	7,595
Epidemic C	0.66% and 1.01%	0.65% and 1.01%	0.95 and 0.95	0.90	29,557
Epidemic D	0.21% and 0.94%	0.22% and 0.96%	0.91 and 0.96	0.90	4,848
(2)					
Epidemic A	0.65% and 0.30%	0.64% and 0.29%	0.95 and 0.92	0.92	17,920
Epidemic B	0.62% and 1.08%	0.65% and 1.05%	0.96 and 0.94	0.82	18,383
Epidemic C	0.83% and 1.01%	0.83% and 1.01%	0.95 and 0.95	0.90	123,545
Epidemic D	0.45% and 0.94%	0.45% and 0.96%	0.95 and 0.96	0.90	12,960

Table 4.3: Comparison of incidences at years 6-ψ and 10-ψ **(1)** and years 8-ψ and 10-ψ **(2)** for four epidemic scenarios. ψ is estimated to be 0.48 years. ASI: Average simulated incidence over all bootstraps and trials. Epidemics A, B, C, and D are the epidemic scenarios labeled in Figure 3.6. CI coverage is the percentage of the 999,000 incidences covered by the confidence interval created according to the Formula (4.1) in Section 4.2. Power is calculated according to the formula in Section 4.3 with equal sample sizes. u is the needed uninfected samples.

If we look at the first two columns of Table 4.3 we see that even after 999,000 iterations the average of the estimated incidences do not perfectly match the true underlying incidences. We would expect this based on two assumptions the cross-sectional methodology makes. These assumptions are explained in the Chapter 1. The more important of these assumptions is that $g(t)$ must be approximately linear in the few years preceding sampling. If $g(t)$ is convex then we will tend to overestimate the true incidence and if $g(t)$ is concave we will tend to underestimate the true incidence. The expected 90% power is achieved in the linear, decreasing, and exponential growth epidemics where the bias in the incidence estimates at the first and second time points is in the same direction. In the epidemic where incidence both decreases and increases the true quantities we are comparing are closer together than the true incidences would lead us to believe. If we look at Table 4.2 we can see that if r is truly closer to 1 than was assumed we may wind up drastically under sampling. This can happen even with a small change. Still, in the two scenarios given, the power is greater than 80% which may be acceptable.

An additional complication at play is that ψ is also defined by $\phi(t)$ (see Chapter 1). Therefore changing $\phi(t)$ in each bootstrap changes ψ slightly and thus each bootstrap would need its own sample size calculation. We fix the sample sizes across all the bootstraps to mirror reality where we must estimate ψ as $\widehat{\psi}$. Since this curve is never known exactly we are forced to base our alternative hypothesis on the incidence $\widehat{\psi}$ years before the survey.

Although the sample sizes were chosen for the power calculations we can still use the simulation to check the validity of the methods developed in Section 4.2. Using the determined samples sizes and the underlying incidences we calculated the percentage of the 999,000 estimated incidences which were covered by the anticipated 95% confidence interval used to define W. We see that the expected 95% coverage is achieved in most of the scenarios presented in the third column of Table 4.3. The confidence intervals seem to be anti-conservative when the underlying incidence is $\leq 0.3\%$ per year. This appears to be a consequence of the fact that the number of individuals marked in the early disease stage must be a whole number so that the distribution of the simulated incidence values increases approximately in steps between 0, $\frac{1}{\mu u}$, $\frac{2}{\mu u}$, $\frac{3}{\mu u}$, etc. There is some variability around each

step as our actual estimates of incidence use n_u, the number of uninfected individuals sampled, rather than u, which is the expected value of N_u. When the underlying incidence is particularly close to 0 the mass is centered around a small number of these jumps and it becomes hard to get proper coverage. For example, for Epidemic D in Table 4.3 when the true incidence is 0.21%, 1.6% of the simulated values are exactly 0, none lie between 0 and 0.0515%, and 6.5% are between 0.0515%, and 0.0520%. Thus even a small shift in the bound of the confidence interval will exclude all 6.5% instead of including this mass.

4.5 Further Extensions

We note that this methodology could be used to compare the incidence of two different subpopulations if the underlying distribution for μ is the same in both groups. This may be the case in two different states with similar demographics.

If the distribution of the average duration spent in the early disease stage is not the same in the two groups then we can generalize the above methodology. This could be the case if we were comparing incidence in two different countries where there are differences in the specific clade of virus infecting people and in the use of antiretrovirals. To generalize the methodology we will need to specify the joint distribution, h(μ_1,μ_2). We assume independence so that $h(\mu_1, \mu_2) = e(\mu_1)f(\mu_2)$. We then have that the power P(Reject $H_0|H_A$) can be written as

$$\int_0^\infty \int_0^\infty \sum_{t=0}^{\infty} \mathrm{P}(\mathrm{Reject} H_0|H_A, T = t, \mu_1, \mu_2)\mathrm{P}(T = t|H_A, \mu_1, \mu_2)e(\mu_1)f(\mu_2)d\mu_1 d\mu_2.$$

To simplify the calculations we will let $s = 1$ so that $u_1 = u_2$. Then under H_0 we have

$$X_2|t, \mu_1, \mu_2 \sim \text{Binomial}\left(t, \frac{\mu_2}{\mu_1 + \mu_2}\right)$$

and under H_A we get

$$X_2|t, \mu_1, \mu_2 \sim \text{Binomial}\left(t, \frac{\mu_2 r}{\mu_1 + \mu_2 r}\right).$$

Therefore the power conditional on t, μ_1, and μ_2 can be calculated by finding the smallest

cutoff C such that

$$\sum_{k=C}^{t} \binom{t}{k} \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^k \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{t-k}$$

is less than or equal to the desired α level and then evaluating

$$\sum_{k=C}^{t} \binom{t}{k} \left(\frac{\mu_2 r}{\mu_1+\mu_2 r}\right)^k \left(\frac{\mu_1}{\mu_1+\mu_2 r}\right)^{t-k}.$$

The next term, $\mathrm{P}(T = t|H_A, \mu_1, \mu_2)$, is simply the probability density function of a Poisson random variable which has mean $u_1 I_1(\mu_1 + r\mu_2)$. Lastly we assume that, as in the case above, $e(\mu_1)$ and $f(\mu_2)$ are known distributions estimated before the cross-sectional survey. Then the power $\mathrm{P}(\mathrm{Reject}H_0|H_A)$ is

$$\int_0^\infty \int_0^\infty \sum_{t=0}^{\infty} \left[\sum_{k=C}^{t} \binom{t}{k} \left(\frac{\mu_2 r}{\mu_1+\mu_2 r}\right)^k \left(\frac{\mu_1}{\mu_1+\mu_2 r}\right)^{t-k}\right] \frac{(u_1 I_1(\mu_1 + r\mu_2))^t}{e^{(u_1 I_1(\mu_1+r\mu_2))} t!} e(\mu_1) f(\mu_2) d\mu_1 d\mu_2.$$

We have solved for $u_1 I_1$ in Mathematica for the specific scenario where $e(\mu_1)$ follows the Gamma distribution shown by the red curve in Figure 4.1 and $f(\mu_2)$ follows the Gamma distribution shown by the green curve in Figure 4.1. We omit those results in the interest of leaving this section as general as possible. The computations require numerical integration and are much more computationally intensive than the cases we examined in Section 4.3 which were solved rapidly in R.

This extension allows for further flexibility if one wishes to use the methods developed in this paper in other contexts. Our methods apply to any situation where $X \sim \mathrm{Poisson}(C \times \lambda)$ where X is observable, uncertainty in C motivates specification of a prior distribution, and inference is focused on λ. This situation arises if λ is the event rate of any condition and X is the observed number of events over an uncertain "length" C. For any disease in steady-state (Freeman and Hutchison, 1980) this emerges from the relationship prevalence = (incidence $\times$ average duration) when the average duration of infection is not directly observable. Outside of the arena of incidence estimation we may readily find applications for this methodology. For example, in occupational epidemiology we may be interested in comparing the rate of

accidents at two different locations. Traditionally a comparative Poisson trial would be used to compare the rate of accidents. However, uncertainty may exist in the person hours contributed in each location. The methods developed in this paper allow for the traditional comparative Poisson trial to proceed by accounting for this uncertainty by placing a prior distribution on the number of person hours.

4.6 Discussion

In this chapter we derived a methodology for finding the needed sample sizes for both single and consecutive cross-sectional surveys designed to estimate incidence. The methodology gives the required number of uninfected samples to conduct a single survey with a desired amount of precision or successive surveys to detect trends in incidence. These numbers can be adjusted to find the required sample sizes. Since, in practice, cross-sectional surveys occur over a period of several weeks preliminary estimates used for this calculation could be updated as sampling occurs. Further work could explore how best to make these adjustments.

We have detailed situations where one must account for uncertainty in μ to avoid under sampling. When we account for this uncertainty by placing a Gamma distribution on μ we showed how the shape of the distribution affects the required sample sizes. The Gamma distribution is defined by its mean and variance. Larger samples sizes are required when the variance increases or the mean decreases. In general, larger sample sizes are required the more mass the distribution puts on smaller values of μ. Researchers should keep these concepts in mind when considering a potential biomarker-based testing algorithm to use in a cross-sectional survey. The effects are magnified when the situation already demands large sample sizes such as when the underlying incidence is small, a large amount of precision is required, or we wish to detect a small change in incidence over time. In the United States where incidence is significantly less than 1% ignoring the uncertainty in μ could lead to a sizable loss of precision and reduction in power.

We tested our methodology in four different underlying epidemics. We found that the formula developed for the single survey generally had good coverage. When we examined

successive surveys we found that the bias in the cross-sectional estimator can decrease the power when the underlying epidemic is fluctuating. However, we achieved the desired power in a linearly increasing, an exponentially increasing, and a waning epidemic. The samples size calculations therefore seem to be accurate in settings where the underlying methodology behind cross-sectional HIV incidence estimation is not biased. One solution to the possibility of under powering the studies would be to place a prior on the initial estimate of incidence at ψ years before the first time point. This would allow for the reality that the initial estimate will have uncertainty both because we are trying to estimate current incidence and because we will not know the underlying shape of the epidemic.

CHAPTER 5

Missing Biomarker Data

We have seen in our sample size calculations that there are two different sources of information which enter into our final estimate of incidence. Missing data present a challenge in both the cross-sectional survey itself and in our estimation of $\phi(t)$ which occurs before the survey is conducted. In this chapter we present techniques designed to address specific challenges presented by missing biomarker assay results.

5.1 Missing Data in the Cross-Sectional Survey

CD4 cell count must be measured at the time of sample collection and cannot be determined from a stored specimen. If for some reason a CD4 cell count is not measured at the time of the cross-sectional survey it is still possible to test the stored sample for viral load, the LAg Avidity assay, and the BioRad Avidity assay. If we were to use MAA2 shown in Figure 3.10 it may be possible to exclude certain samples from the early disease stage based on these biomarkers. However it is never possible to include someone in the early disease stage without all four biomarkers. One way to treat all samples which have missing CD4 cell counts in a uniform manner is to use multiple MAAs. As an example MAA2 could be used when CD4 cell count is available and MAA3 could be used when CD4 cell count is missing. Then, if we let

- $\phi_{2\cup3}(t) = \text{P(Marked in an early disease stage by either MAA}|t)$
- $\phi_2(t) = \text{P(MAA2 positive}|\text{MAA2 used},t)$
- $\phi_3(t) = \text{P(MAA3 positive}|\text{MAA3 used},t)$

We have, by the law of total probability,

$$\phi_{2\cup3}(t) = \phi_2(t)\mathrm{P}(\text{MAA2 used}|t) + \phi_3(t)\mathrm{P}(\text{MAA3 used}|t).$$

In this scenario, where missing data arises from testing difficulties, it is reasonable to assume that which MAA is used is not dependent on how long someone has been infected. Therefore

$$\phi_{2\cup3}(t) = \phi_2(t)\mathrm{P}(\text{MAA2 used}) + \phi_3(t)\mathrm{P}(\text{MAA3 used})$$

so that

$$\mu = \int_0^M \phi_{2\cup3}(t)dt = \int_0^M \{\phi_2(t)\mathrm{P}(\text{MAA2 used}) + \phi_3(t)\mathrm{P}(\text{MAA3 used})\}dt =$$

$$\mathrm{P}(\text{MAA2 used}) \int_0^M \phi_2(t)dt + \mathrm{P}(\text{MAA3 used}) \int_0^M \phi_3(t)dt =$$

$$\mathrm{P}(\text{MAA2 used})\mu_2 + \mathrm{P}(\text{MAA3 used})\mu_3.$$

where μ_2 and μ_3 are the average duration spent in the early disease stage associated with MAA2 and MAA3, respectively. Thus we are able to maintain a larger μ by using MAA2 and MAA3 together than if we simply used MAA3. Then the average duration individuals spend in either early disease stage is simply the weighted average of the mean duration in the early disease stage associated with each MAA. In this scenario P(MAA2 used) = 1 - P(MAA3 used) can be determined for sample size calculations based on the expected failure rate of testing for CD4 cell count in the field. After the survey, when we estimate the incidence, the actual percentages where CD4 cell count was used could be applied as sample estimates of the weights. This method generalizes to any number of MAAs so that

$$\mu = \sum_i \mathrm{P}(\text{MAA}^{(i)} \text{ used})\mu_i.$$

The shadow for a biomarker-based testing algorithm based on several MAAs can similarly be determined.

$$\psi = \frac{1}{\mu}\int_0^M t\phi(t)dt = \frac{1}{\mu}\int_0^M t\left\{\sum_i \phi_i(t)\mathrm{P}(\mathrm{MAA}^{(i)}\text{ used})\right\}dt =$$

$$\sum_i \left\{\frac{\mathrm{P}(\mathrm{MAA}^{(i)}\text{ used})}{\mu}\int_0^M t\phi_i(t)dt\right\} =$$

$$\sum_i \left\{\frac{\mathrm{P}(\mathrm{MAA}^{(i)}\text{ used})\mu_i}{\mu}\psi_i dt\right\}$$

Where $\phi(t)$ is the probability, at time t, of being marked in the early disease stage by any $\mathrm{MAA}^{(i)}$, and $\phi_i(t)$ is the probability, at time t, of being marked in the early disease stage by $\mathrm{MAA}^{(i)}$ specifically. Since $\mu = \sum_i \mathrm{P}(\mathrm{MAA}^{(i)}\text{ used})\mu_i$ this is again a weighted average.

Biomarker-based testing algorithms composed of several MAAs have applications beyond missing data applications. It is known that disease progression differs between the various clades of HIV. Thus, for example, the true values of μ and ψ will be different if the same MAA is used in a Clade A epidemic and in a Clade D epidemic. The epidemics in some countries, including Uganda and Kenya, are mixed clade epidemics. The above procedure can be applied to estimate incidence in these countries. Specifically we may estimate

$$I(\psi) = I\left(\sum_i \frac{\mathrm{P}(\mathrm{MAA}^{(i)}\text{ used})\mu_i}{\mu}\psi_i dt\right)$$

with the estimator

$$\widehat{I}_c = \frac{X}{N_u\mu} = \frac{X}{N_u\sum_i \mathrm{P}(\mathrm{MAA}^{(i)}\text{ used})\mu_i}$$

where X is the number of people marked in the early disease stage by any clade-specific MAA and N_u is the number of uninfected people sampled.

5.2 Missing Data in the External Data Set: Estimating $\phi(t)$

New biomarkers for estimating HIV incidence are actively being sought by the National Institutes of Health. While biomarker assays are under development standardized machinery will not exist for testing samples. This greatly increases the cost of running these assays as well as the manpower and laboratory time needed to test samples. Since MAAs have shown desirable characteristics we may care less about the individual performance of a new assay and more about the performance of a hypothetical MAA which includes this assay as one of its components. If the new assay is run last, then the sequential nature of the MAA will limit the number of samples which need to be tested with the additional assay when estimating $\phi(t)$. Only those samples which have not already been excluded from the early disease stage by the previous component assays will need to be tested.

Missing data which arises in this framework cannot be considered missing completely at random since the missingness is related to the observed values of the original assays. One of the reasons why missing data may occur is sample depletion. The quantity of stored specimens is limited and the specimens eventually will not contain enough biological material for further evaluation. In these cases the missingness would be considered missing at random because the hypothetical new assay result for a sample is unrelated to whether or not a specific sample is tested with the new assay.

In order to explain how to properly handle this missing data we define the following

- W = the event that a sample would be marked in the early disease stage for an MAA which includes both the new and original assays
- Z = the event that a sample would be marked in the early disease stage for an MAA which includes only the original assays
- MA = the event that a sample is missing an assay result for the new assay
- A = the event that a sample has an assay result for the new assay

To evaluate an MAA which includes both the new and original assays we must model the associated $\phi(t)$ curve. Now

$$\phi(t) = \mathrm{P}(W|t) = \mathrm{P}(W \cap Z|t) = \mathrm{P}(W|Z,t)\mathrm{P}(Z|t)$$

since Z is contained in W. Next

$$\mathrm{P}(W|Z,t) = \mathrm{P}(W|Z,t,MA)\mathrm{P}(MA|t,Z) + \mathrm{P}(W|Z,t,A)\mathrm{P}(A|t,Z).$$

If we assume that the missingness only depends on the original assays then

$$\mathrm{P}(W|Z,t,MA) = \mathrm{P}(W|Z,t,A)$$

and thus

$$\mathrm{P}(W|Z,t,MA)\mathrm{P}(MA|t,Z)+\mathrm{P}(W|Z,t,A)\mathrm{P}(A|t,Z) = \mathrm{P}(W|Z,t,A)[\mathrm{P}(MA|t,Z)+\mathrm{P}(A|t,Z)]$$

$$= \mathrm{P}(W|Z,t,A).$$

Therefore

$$\phi(t) = \mathrm{P}(W|t) = \mathrm{P}(W|Z,t,A)\mathrm{P}(Z|t).$$

This allows us to use the full data to estimate $\mathrm{P}(Z|t)$ and then use only the observed data on the new assay to estimate $\mathrm{P}(W|Z,t,A)$. The estimation can occur completely separately because maximizing the likelihood is equivalent to maximizing the log of the likelihood. Since the log breaks the two parts into a sum, the maximum likelihood will occur when both individual parts are maximized.

5.2.1 Application: HIV Diversity as a Biomarker

5.2.1.1 Sequence Ambiguity as a Biomarker

The genetic diversity of HIV is thought to increase within an individual over the course of infection (Shankarappa et al., 1999). One way to measure the diversity is to measure the percentage of ambiguous bases in the POL region of the HIV virus. The technology for the HIV genotyping system was developed for HIV drug resistance testing (ViroSeq HIV-1 Genotyping System, Celera, Alameda, CA, USA). This means that many laboratories in the United States already are equipped to run the commercially available ViroSeq system. The process is extremely costly and labor intensive. We will call this biomarker *sequence ambiguity*. Since the process for implementation is widely available but the costs are high, sequence ambiguity is a promising biomarker for measuring HIV incidence but only within the structure of an MAA where it need not be run on all samples.

A cutoff of 0.5% for the sequence ambiguity assay was used in an initial report which evaluated the biomarker by itself (Kouyos et al., 2011). We elected to see if at this cutoff sequence ambiguity could replace CD4 cell count in MAA2. Since sequence ambiguity can be run as the last component assay, unlike the expensive CD4 cell count which must be run as the first component assay, this could potentially greatly reduce the cost. The proposed MAA is displayed in Figure 5.1.

Of the 1,782 samples in the data set only 119 samples have BioRad Avidity results $< 85\%$, LAg Avidity results < 2.9 ODn and HIV viral loads > 400 copies/mL. We need to evaluate only these samples with the ViroSeq system to measure the sequence ambiguity since the remaining samples will be excluded from the early disease stage regardless of the sequence ambiguity score. However, 6 samples have been depleted and can no longer be evaluated. Further, the sequence ambiguity assay failed to produce a result for 5 more samples. To simplify the analysis we will assume that the missingness of all 11 samples is not informative conditional on the BioRad Avidity assay, LAg Avidity assay, and HIV viral load results.

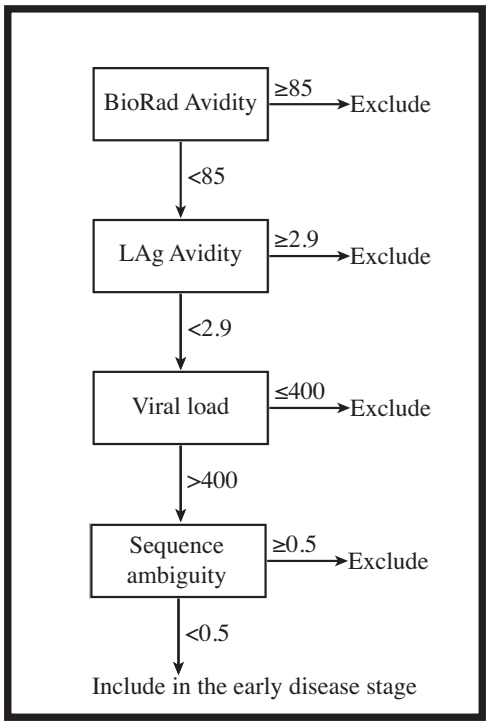

Figure 5.1: Proposed MAA with sequence ambiguity. The following units are used for the component assays: BioRad Avidity assay: percentage (avidity index); LAg Avidity assay: optical density units (ODn); viral load: copies/mL; sequence ambiguity: percentage.

Below we outline how to estimate $\phi(t)$ for the MAA shown in Figure 5.1 based on the framework outlined in Section 5.2.

---

1. For each sample create z_{ij} which equals 1 if the BioRad Avidity assay result $< 85\%$, the LAg Avidity assay result < 2.9 ODn, and the HIV viral load > 400 copies/mL. Let $z_{ij} = 0$ otherwise.

2. For each sample where $z_{ij} = 1$ create $w_{ij} = 1$ if the sequence ambiguity $< 0.5\%$ and 0 otherwise.

3. Draw s_i from a uniform distribution on the interval $(0, d^+ - d^-)$ for $1 \leq i \leq 709$ and calculate $t_{ij} = \tau_{ij} + s_i$ for each sample.

4. Fit the model $\text{logit}[\text{P}(z_{ij} = 1)] = \beta_0^{(k)} + \beta_1^{(k)} t_{ij} + \beta_2^{(k)} t_{ij}^2 + \beta_3^{(k)} t_{ij}^3 + \beta_4^{(k)} (t_{ij} - 2)_+^3$

5. Fit the model $\text{logit}[\text{P}(w_{ij} = 1)] = \eta_0^{(k)} + \eta_1^{(k)} t_{ij}$

6. Repeat steps 3, 4, and 5 for $1 \leq k \leq 1000$ and calculate $\sum_k \beta_\ell^{(k)}/1000 = \beta_\ell$ for $\ell = 0$, 1, 2, 3, and 4 and $\sum_k \eta_\ell^{(k)}/1000 = \eta_\ell$ for $\ell = 0$ and 1

7. Calculate $\phi(t) = \phi_{W|Z}(t)\phi_Z(t)$ where $\phi_Z(t) = \text{expit}\{\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 (t-2)_+^3\}$ for $0 \leq t \leq 8$ and $\phi_{W|Z}(t) = \text{expit}\{\eta_0 + \eta_1 t\}$ for $0 \leq t \leq 8$

One limitation of this method is that we have increased the number of parameters we need to estimate. Further, only 108 samples are used to estimate $\phi_{W|Z}(t) = \text{expit}\{\eta_0 + \eta_1 t\}$. The small amount of data is why we chose to use a linear function instead of the more flexible cubic spline. Figure 5.2 shows the estimates of $\phi(t)$, $\phi_Z(t)$, and $\phi_{W|Z}(t)$ in black, red and blue respectively.

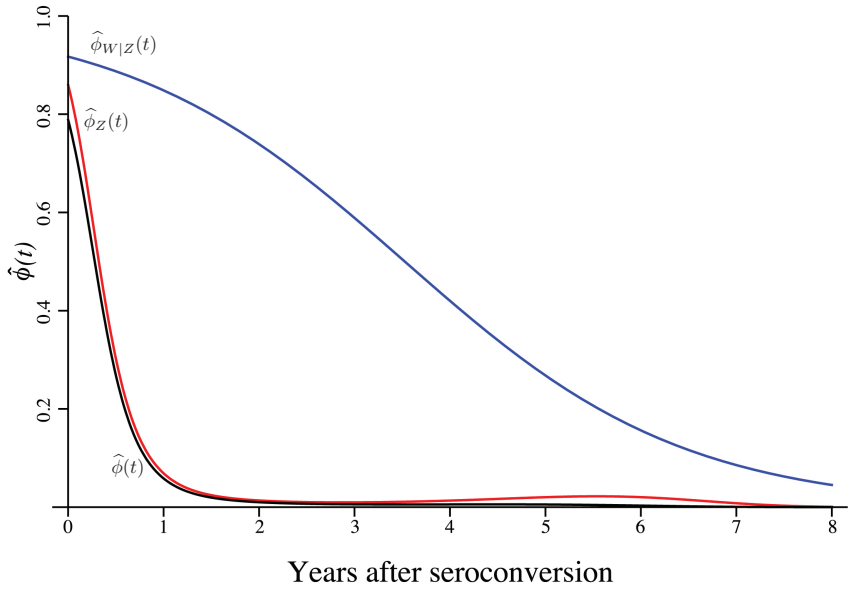

Figure 5.2: Various probability curves calculated for estimating $\phi(t)$ in the presence of missing data. The black curve is the estimated $\phi(t)$ curve for the MAA shown in Figure 5.1. The red curve is the estimated $\phi_Z(t)$ curve with an early disease stage defined by the MAA shown in Figure 5.1 but without the sequence ambiguity component assay. The blue curve gives the effect of adding in the sequence ambiguity biomarkers, so that the red curve multiplied by the blue curve is the black curve.

We see that adding the sequence ambiguity assay appears to trim the tail of the curve compared to an MAA which defines the early disease stage to be a BioRad Avidity assay result $< 85\%$, a LAg Avidity assay result < 2.9 ODn, and an HIV viral load > 400 copies/mL and does not use the sequence ambiguity assay. The point estimates for μ and ψ are 0.38 and 0.61 years for the MAA with sequence ambiguity compared to 0.48 and 1.11 years for the MAA which does not include sequence ambiguity. Adding sequence ambiguity decreases the average length of time spent in the early disease stage. Its benefit appears to be the large reduction in the shadow. While the reduction in the shadow is promising, these results must be interpreted cautiously.

Figure 5.3 shows the sequence ambiguity scores for the 119 samples with a BioRad Avidity assay $< 85\%$, a LAg Avidity assay < 2.9 ODn, and an HIV viral load > 400 copies/mL. The cutoff of 0.5% is shown as a red line. Samples below this line are marked in the early disease stage denoted in pink. The 11 samples without sequence ambiguity results are plotted at the cutoff as red Xs.

Importantly, there are no observed data beyond 5 years. Figure 5.2 shows that the most important effect of adding sequence ambiguity is the trimming of the tail of the red curve. This is based entirely on model extrapolation with the observed drop in the blue curve continuing based on the earlier data. Fortunately, the confidence intervals for μ and ψ properly capture this lack of certainty. They were estimated as (0.30,0.46) years and (0.36, 1.17) years respectively. Thus, we are fairly certain that adding sequence ambiguity will reduce μ (estimated as 0.48 years without sequence ambiguity), but the apparent reduction in the estimate of the shadow is not conclusive (estimated as 1.11 years without sequence ambiguity). Adding sequence ambiguity may have a negligible effect on the shadow. It almost certainly has a negative effect on the average amount of time spent in the early disease stage. Further, the fact that the upper bound of the confidence interval for the shadow is greater than 1 year implies that incidence might not be measured within one year of the cross-sectional survey. It is possible that with more data on sequence ambiguity we could estimate narrower confidence intervals and that this MAA could be used in practice. For now the data does not conclusively show that it can replace CD4 cell count.

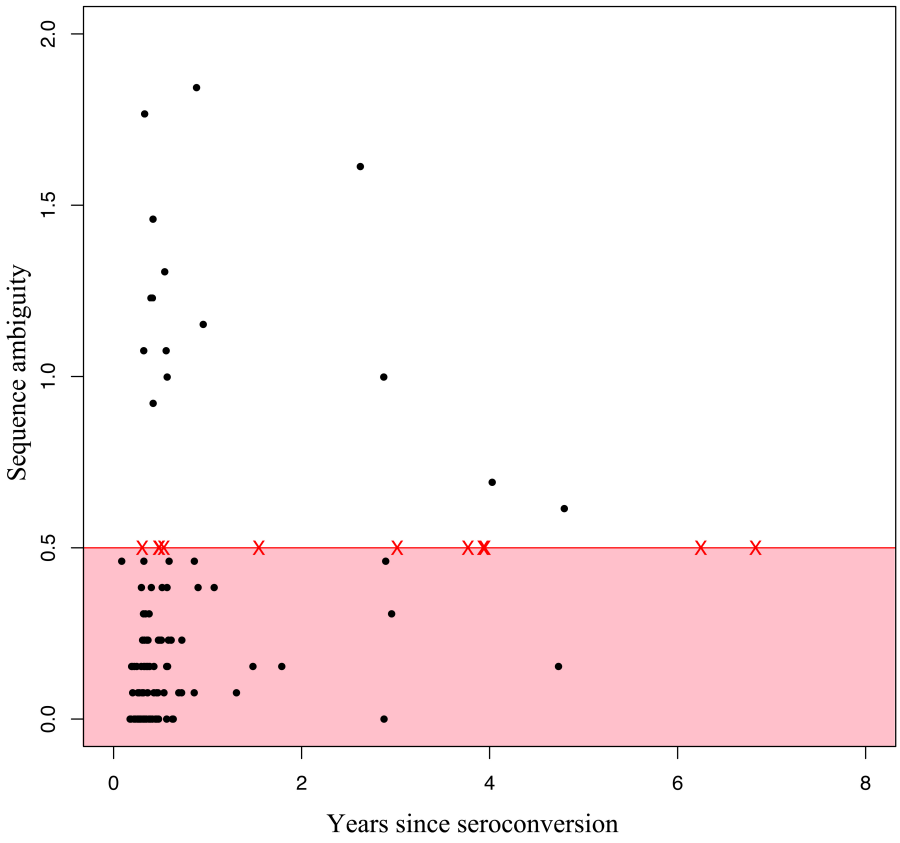

Figure 5.3: Plot of sequence ambiguity scores for 119 samples where the BioRad Avidity assay $< 85\%$, the LAg Avidity assay < 2.9 ODn, and the HIV viral load > 400 copies/mL. Samples below the cutoff line of 0.5%, in the pink region of the graph, are in the early disease stage. The 11 red Xs at the cutoff line represent missing data points.

5.2.1.2 High-Resolution Melting as a Biomarker

Recently a new biomarker was developed to try to quantify the genetic diversity of the virus through high-resolution melting (HRM) technology (Cousins et al., 2011; Cousins, Swan, et al., 2012; Cousins et al., 2013; Towler et al., 2010). The amount of genetic diversity in a region of the virus is measured by the width of the DNA melting peak. Higher HRM scores are thought to indicate longer infection times. HRM scores have been shown to be highly correlated with diversity measures obtained from next-generation sequencing data (Cousins, Ou, et al., 2012). HRM is a less labor intensive and a more cost efficient biomarker than sequence ambiguity. It is, however, not commercially available and only a single lab is equipped to test samples with this biomarker.

A cutoff of 4.5 for the HRM diversity assay on the first part of the ENV region (ENV1) of the virus has been suggested (Cousins, Konikoff, Laeyendecker, et al., 2014; Cousins, Konikoff, Sabin, et al., 2014). For comparison to the sequence ambiguity biomarker we repeated the above analysis for the HRM biomarker with a cutoff of 4.5. The MAA we examined is shown in Figure 5.4.

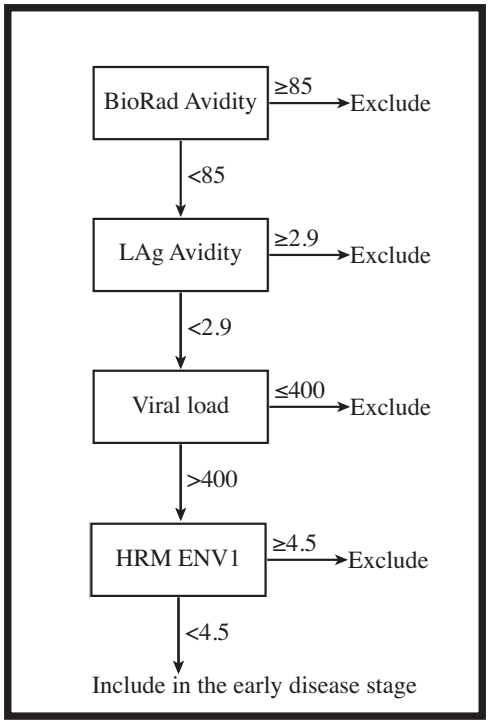

Figure 5.4: Proposed MAA with HRM using the first part of the ENV region of the virus (ENV1). The following units are used for the component assays: BioRad Avidity assay: percentage (avidity index); LAg Avidity assay: optical density units (ODn); viral load: copies/mL; HRM ENV1: HRM score.

Of the 119 samples, which have a BioRad Avidity assay result $< 85\%$, a LAg Avidity assay result < 2.9 ODn, and an HIV viral load > 400 copies/mL, 111 were successfully tested with the HRM assay. The six depleted samples could not be tested and two additional samples failed testing. Figure 5.5 shows the HRM scores for the ENV1 region for the 119 samples with a BioRad Avidity assay result $< 85\%$, a LAg Avidity assay result < 2.9 ODn, and an HIV viral load > 400 copies/mL. The cutoff of 4.5 is shown as a red line. Samples below this line are marked in the early disease stage. The 8 samples without sequence ambiguity results are plotted at the cutoff as red Xs. We see that there is still limited data on how

this diversity marker performs on individuals infected for longer periods of time. HRM does seem to do a better job than sequence ambiguity at classifying recently infected individuals as in the early disease stage. The estimates for μ and ψ for this MAA were 0.41 (0.33,0.48) years and 0.55 (0.39, 0.91) years respectively. Thus, based on our limited data, it would appear that this MAA can be used to estimate incidence within the past year and further exploration of this biomarker is warranted.

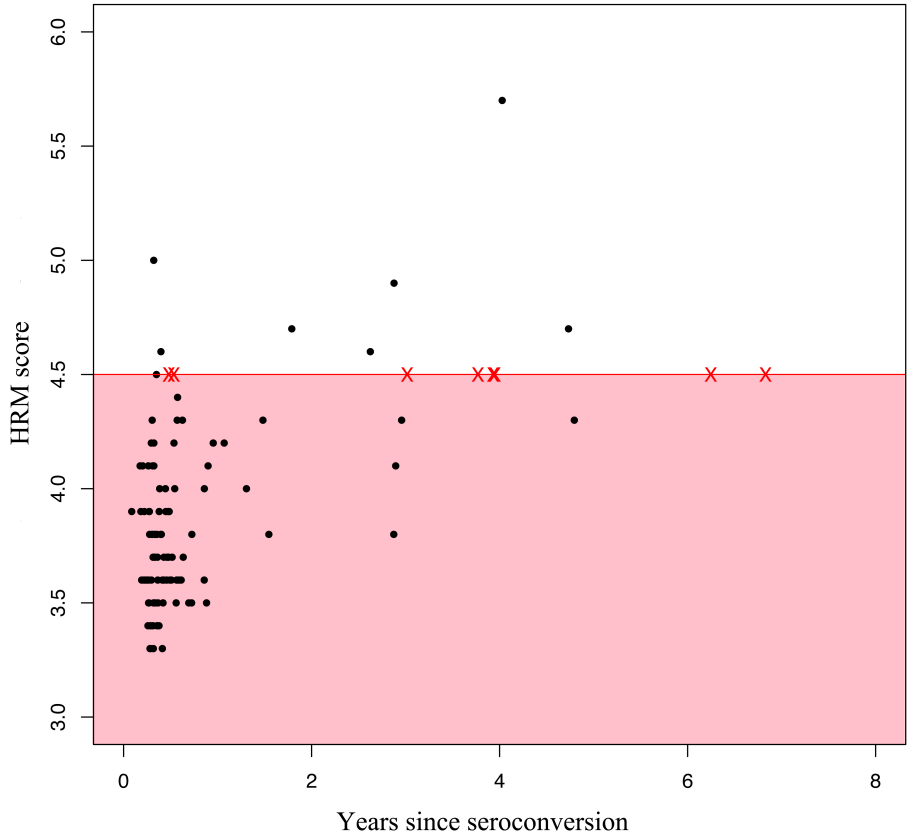

Figure 5.5: Plot of HRM scores for 119 samples where the BioRad Avidity assay $< 85\%$, the LAg Avidity assay < 2.9 ODn, and the HIV viral load > 400 copies/mL. Samples below the cutoff line of 4.5, in the pink region of the graph, are in the early disease stage. The 8 red Xs at the cutoff line represent missing data points.

CHAPTER 6

Multivariate Modeling of Biomarkers

In Chapters 2 and 3 we presented one approach for estimating $\phi(t)$. In this chapter we develop an alternative approach for estimating this probability curve by modeling the longitudinal biomarker measurements rather than a binary variable indicating presence or absence in the early disease stage. To distinguish the two methods we will call the methodology introduced in Chapter 2 the *binary classification approach* (BCA) and the one presented in this chapter the *biomarker progression approach* (BPA).

We will specifically develop the BPA for a 2-assay MAA structured in the form of MAA3, shown in Figure 3.10, with the LAg Avidity and BioRad Avidity assays as component assays. We show how to estimate $\phi(t)$ from a joint bivariate model of how the normalized optical density units (ODn) measured by the LAg Avidity assay, and the avidity index (AI) measured by the Biorad Avidity assay, change over the course of infection. Extensions for additional biomarkers are discussed in Section 6.5. We note that results found throughout this document do not depend on how $\phi(t)$ is estimated and can easily be adapted to this new methodology. The BPA will necessitate one small change. In the sample size calculations, instead of using the bootstrap to generate a distribution for μ, the BPA explicitly treats μ as having a distribution within a Bayesian framework. Thus the posterior distribution for μ can be used in the calculations directly or a Gamma distribution can be fit to the posterior distribution by matching the first two moments.

6.1 The Biomarker Progression Approach

Both the LAg Avidity assay and BioRad Avidity assay responses grow following seroconversion (Duong et al., 2012; Kassanjee et al., 2014). Figure 6.1 illustrates the growth pattern by plotting the log LAg Avidity and log BioRad Avidity assay results from the data set described in Chapter 2.

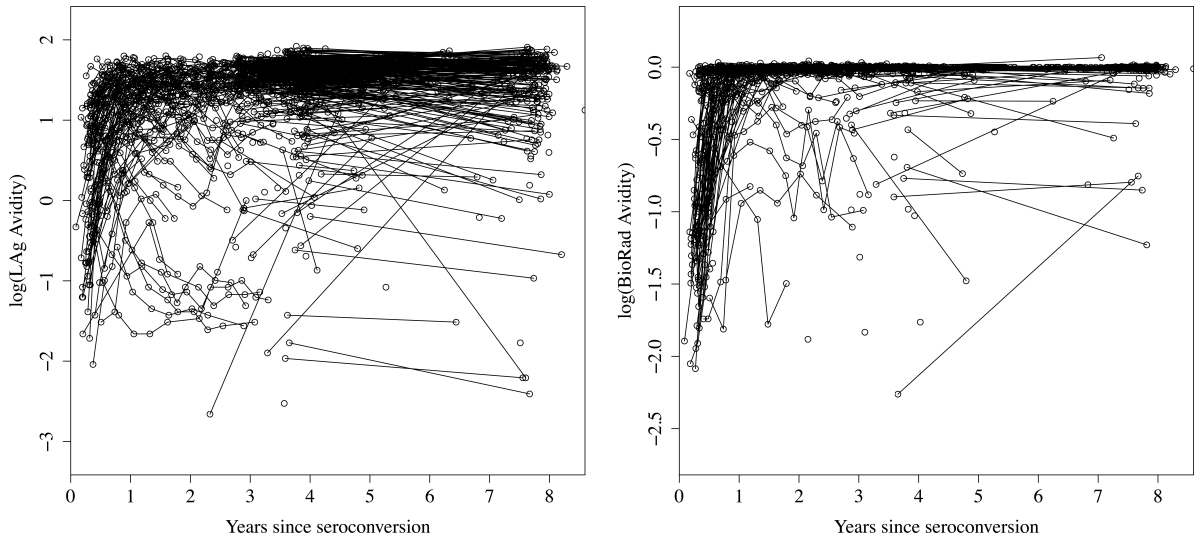

Figure 6.1: Profile plots illustrating the growth of the log LAg Avidity and log BioRad Avidity assay results (1,782 measurements from 709 people). Lines connect samples taken from the same person.

There is a fair amount of noise in these plots but also a distinctive growth pattern for both assays.

Figure 6.2 illustrates typical data from an individual which might be available to model this bivariate immune response. Here we use the same notation described in Section 2.2 and illustrated in Figure 2.2. The Lag Avidity assay and the BioRad Avidity assay measurements collected at each time τ_{ij} are denoted ℓ_{ij}, and b_{ij} respectively. The pink rectangular box indicates inclusion in the early disease stage for the MAA3. We will use this MAA, first described in Chapter 3, as a motivating example and explore different cutoffs shortly. Thus by the measurement at time τ_{i2} the person is excluded from the early disease stage since his or her BioRad Avidity assay result is greater than 40%.

In Figure 6.1 we see that marginally both assay responses grow rapidly before leveling

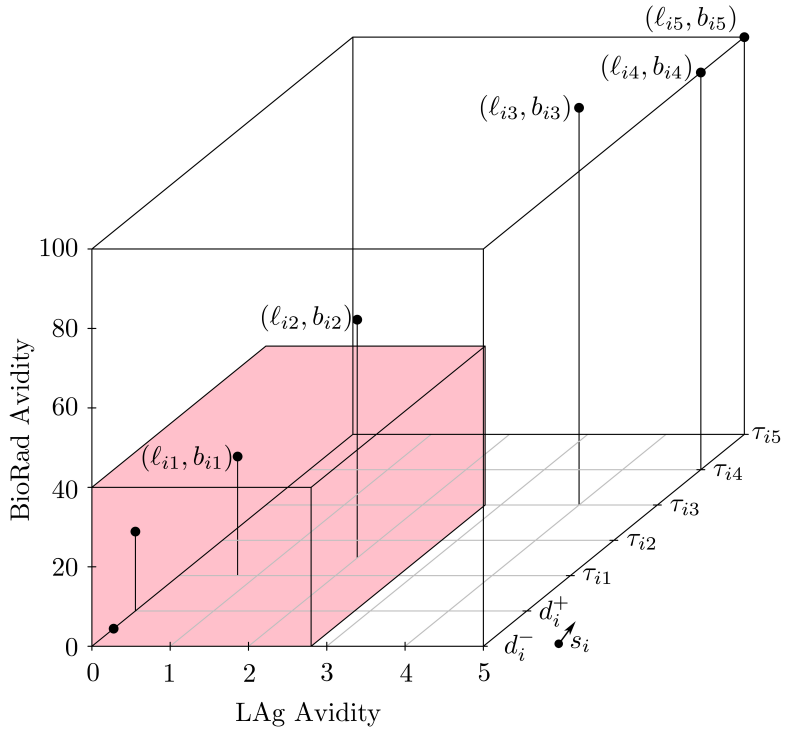

Figure 6.2: Typical bivariate biomarker measurements from one individual i with observations $j = 1$, 2, 3, 4, and 5. The LAg Avidity, ℓ_{ij}, and BioRad Avidity, b_{ij}, measurements are taken at times τ_{ij} following the first positive test, d_i^+. Seroconversion occurred between the last negative test, d_i^-, and the first positive one. The total time since seroconversion is $t_{ij} = \tau_{ij} + s_i$.

off to an asymptote. This motivates a bivariate nonlinear mixed effect model

$$\begin{aligned} \log(\ell_{ij}) &= \beta_0 + (\beta_1 - \beta_0)\exp(-\exp(\eta_{1i})t_{ij}) + \delta_{1ij} \\ \log(b_{ij}) &= \beta_2 + (\beta_3 - \beta_2)\exp(-\exp(\eta_{2i})t_{ij}) + \delta_{2ij} \end{aligned} \tag{6.1}$$

The log Lag Avidity assay result, $\log(\ell_{ij})$, and the log BioRad Avidity assay result, $\log(b_{ij})$, are modeled as monotonically increasing nonlinear functions of the time from seroconversion. The parameters η_{1i} and η_{2i} are specified as random effects and modeled using a bivariate normal distribution $(\eta_{1i}, \eta_{2i})^T \sim \mathrm{N}_2((m_0, m_1)^T, R)$. The random effects represent the growth rate for each individual i with population averages m_0 and m_1. Possible correlation between the two growth rates is determined by the off diagonal term in the covariance

matrix R. At $t_{ij} = 0$ the intercepts for each growth curve are given by fixed effects β_1 and β_3. As the time from seroconversion increases the antibody responses are expected to level off. The remaining fixed effects, β_0 and β_2, represent asymptotes approached by the log Lag Avidity assay result, $\log(\ell_{ij})$, and the log BioRad Avidity assay result, $\log(b_{ij})$, respectively. The measurements between the two assays are further correlated through residual error specified with a bivariate normal distribution $(\delta_{1ij}, \delta_{2ij})^T \sim \mathrm{N}_2((0,0)^T, D)$.

Equation (6.1) depends on the unknown s_i because $t_{ij} = \tau_{ij} + s_i$. We *a priori* modeled

$$s_i \sim \mathrm{Uniform}(0, d^+ - d^-). \tag{6.2}$$

We fit (6.1) and (6.2) in a Bayesian framework using Markov Chain Monte Carlo (MCMC) methods. For m_0, m_1, β_1, β_2, β_3, and β_4 Gaussian priors are specified. Covariance matrices R and D are modeled *a priori* by letting R^{-1} and D^{-1} follow Wishart distributions. The probability density function for a random variable Q, distributed Wishart(Ψ, ν) with scale matrix Ψ and degrees of freedom ν, is given by

$$\frac{|\Psi|^{\nu/2}}{2^{(\nu p)/2}\Gamma_p(\nu/2)}|Q|^{(\nu-p-1)/2}\exp(-\mathrm{tr}(\Psi Q/2))$$

where tr() is the trace function, Γ_p is a p-variate Gamma function, and Ψ is a $p \times p$ positive definite matrix. The degrees of freedom are often set to $p + 1$ since this places a uniform prior on the correlation parameters (Gelman and Hill, 2007). Thus for both R^{-1} and D^{-1} we will set the degrees of freedom to 3. Lower degrees of freedom on the Wishart distributions also minimize the impact of the priors. We use the fact that the prior mean of Q is $\nu\Psi^{-1}$ to specify the prior distribution on R^{-1} and D^{-1} as follows. Let $R^{-1} \sim$ Wishart$(3\Sigma, 3)$ then the prior mean of the precision matrix for $(\eta_{1i}, \eta_{2i})^T$ is Σ^{-1} implying that Σ can be specified as a prior guess at the covariance matrix for $(\eta_{1i}, \eta_{2i})^T$. Similarly if $D^{-1} \sim$ Wishart$(3S, 3)$ then S can be specified as a prior guess at the covariance matrix for $(\delta_{1ij}, \delta_{2ij})^T$.

The motivation for the univariate marginal models was discussed by Sweeting et al. (2010) who first proposed using a nonlinear mixed effect model for the growth of a different

biomarker developed for identifying recent infections (Suligoi et al., 2003). Sweeting et al. however included a random effect for the intercept term in their model. We use fixed effects for β_1 and β_3, as the unknown seroconversion times can explain differences between assays among individuals at seroconversion. A further difference is that our model removes the possibility of a negative antibody assay result, which cannot occur in practice, by modeling the log transformations of ℓ_{ij} and b_{ij} instead of directly modeling the untransformed biomarker results.

Once the bivariate model has been fit we can estimate the probability of inclusion in the early disease stage by both assays for any specific set of cutoffs and time following seroconversion. The model implies that $\mathrm{P}\{(\ell_{ij} \leq C_\ell) \cap (b_{ij} \leq C_b)|t_{ij}\} = \mathrm{P}\{[\beta_0 + (\beta_1 - \beta_0)\exp(-\exp(\eta_{1i})t_{ij}) + \delta_{1ij} \leq \log(C_\ell)] \cap [\beta_2 + (\beta_3 - \beta_2)\exp(-\exp(\eta_{2i})t_{ij}) + \delta_{2ij} \leq \log(C_b)]\}$ where C_ℓ is the cutoff for the LAg Avidity assay and C_b is the cutoff for the BioRad Avidity assay. In MAA3 $C_\ell = 2.8$ ODn and $C_b = 40\%$. This probability can be estimated by sampling from the posterior distributions of β_0, β_1, β_2, β_3, m_0, m_1, R, and D and then simulating measurements from new individuals by drawing observations for $(\eta_{1NEW}, \eta_{2NEW})^T$ from $\mathrm{N}_2((m_0, m_1)^T, R)$ and for $(\delta_{1\mathrm{NEW}}, \delta_{2\mathrm{NEW}})^T$ from $\mathrm{N}_2((0,0)^T, D)$ at time t_{NEW}. The percentage of new observations where both $\ell_{\mathrm{NEW}} \leq C_\ell$ and $b_{\mathrm{NEW}} \leq C_b$ at time t_{NEW} gives the probability.

6.2 Comparison of Methodologies

The BCA estimates $\phi(t)$ directly and provides consistent inference under the minimal assumption that the probability of classification in the early disease stage has been correctly specified as a function of time. The ideal data set for the BCA is comprised of independent biomarker measurements taken at known times since seroconversion. One might imagine that in an era of increased testing it may be possible to collect a large sample from the population with known seroconversion windows. If these individuals have all been infected for different amounts of time, $\phi(t)$ could then be directly estimated by using the BCA.

In comparison the BPA requires repeated measurements on individuals in order to es-

timate $\phi(t)$. An ideal data set for the BPA is comprised of many individuals, with known seroconversion dates, contributing several observations each. The BPA is designed to be used for data arising from longitudinal studies, like the MACS, ALIVE, and HIVNET cohort studies. Both methodologies could be used with samples collected for reasons other than estimating $\phi(t)$, provided we believe that the biomarker profile in our samples is reflective of the larger population.

In the case of the data set explored in Chapter 2 we had many observations on some individuals but only one observation on several hundred people. By assuming that the longitudinal observations are independent the BCA approach sacrifices efficiency but does maintain consistency. The probability that a specific individual is classified in the early disease stage at one time is likely correlated with whether or not the person was classified in the early disease stage at earlier times. In particular, most individuals will not return to the early disease stage once they leave this state. A Markov model where individuals transition between states may be more appropriate for modeling the binary indicator of classification in the early disease stage. Alternatively, a Generalized Estimating Equation (Liang and Zeger, 1986; Zeger and Liang, 1986; Prentice, 1988) approach could be used to account for the correlation while only modeling the population level probability of classification in the early disease stage.

In contrast to these approaches the BPA intuitively models the correlation within subjects by using the biological process which determines a sample's classification as in or out of the early disease stage. The tradeoff here is making more complicated assumptions about the structure of the data. However, it must be noted that the BPA works with richer data than the BCA since it models the original biomarker measurements and not the compressed binary indicator of presence in the early disease stage. Further, while the BCA avoids modeling the correlation structure, it relies on the bootstrap in order to calculate the confidence intervals for μ and ψ. Bootstrapping techniques for hierarchical data lack precise theoretical assessments of performance except in very specific cases (Field and Welsh, 2007).

An additional advantage of the BPA is that we only need to model the data once. The data does not change the way it does in the BCA where the binary response variable depends

on the choice of cutoffs. In Chapter 3 when we searched for the MAA with the largest μ, subject to various constraints, we assumed that a cubic spline was appropriate for the many thousands of different data sets we created. The cubic spline was chosen to model the data that arose from the definition of the early disease stage associated with MAA1. Above we detailed some cases where the cubic spline poorly modeled the data. While specific rules for when to use a knot and when to include the cubic term can help, the BCA works best if $\phi(t)$ is modeled for each MAA. Individually modeling thousands of data sets is not feasible. With the BPA the cutoffs do not influence the data allowing for thousands of cutoffs to be evaluated after fitting one single model to the data.

Another attractive feature of the BPA is that the model is able to learn when individuals likely seroconverted based on their biomarker results. From a biological perspective we would expect the LAg Avidity and BioRad Avidity assay results to convey information on when the person most likely seroconverted. The posterior distribution of the s_i, the unknown times from seroconversion until the first positive test, conveys how much the model learned from the biomarker measurements. A Bayesian approach, with a uniform prior on s_i, could be incorporated into the BCA. However, in this case the implication is that the model can learn something about the time from seroconversion based on presence or absence in the early disease stage. Therefore the choice of cutoffs for defining the early disease stage will influence the posterior distribution of the times. In light of the fact that we are trying to model the probability of being in the early disease stage as a function of the time from seroconversion, and that we will select the cutoffs based on the performance of each MAA, the actual biomarker measurements are likely to contain more meaningful information on the date of seroconversion.

It may also be possible to use external information on the epidemic to place a more informative prior on the s_i. For example, Brookmeyer and Goedert (1989) estimated the cumulative distribution function of seroconversion times in hemophilia treatment centers and Bacchetti and Moss (1989) estimated the distribution of seroconversion times in a clinical cohort study for San Francisco General Hospital. The latter estimate was used to specify a prior distribution in a Bayesian hierarchical model of the progression of HIV infection based

on longitudinal CD4 cell count measures (Lange et al., 1992).

6.3 Statistical Analysis

Non-informative Gaussian priors, centered at 0 with variance 10^4, were set for the means of the random effects m_0 and m_1. For β_0, β_1, β_2, and β_3, prior information was elicited from Dr. Oliver Laeyendecker who was asked to specify the prior mean for each parameter as well as a 95% prior credible interval. Dr. Oliver Laeyendecker manages a laboratory at Johns Hopkins University where he has worked extensively with these biomarkers. The variances for the normal priors were derived from the 95% prior intervals by assuming that four times the standard deviation would cover this interval. The diagonal terms for Σ and S were derived similarly by specifying plausible ranges based on plots of data from Kassanjee et al. (2014) and Duong et al. (2012). *A priori* we assumed that the two random effects had a small positive correlation and that the error terms were uncorrelated. The specified priors were $\beta_0 \sim \mathrm{N}(0, 0.01)$, $\beta_1 \sim \mathrm{N}(-1.2, 1)$, $\beta_2 \sim \mathrm{N}(1.6, .3)$, and $\beta_3 \sim \mathrm{N}(-1.2, 1.25)$. The entries for Σ and S were $\Sigma_{11} = 1.75$, $\Sigma_{12} = \Sigma_{21} = .1$, $\Sigma_{22} = 1.63$, $S_{11} = .12$, $S_{12} = S_{21} = 0$, and $S_{22} = .04$. Where M_{ij} is the entry in the i^{th} row and j^{th} column of matrix M.

Inference was based on 3 parallel chains fit simultaneously in JAGS version 3.3. Each chain consisted of 300,000 iterations with the first third discarded as burn-in. The chains were thinned so that every 20th iteration was kept. This allowed for the $\phi(t)$ curve for an MAA to be calculated in under 24 hours on a single machine. The Brooks and Gelman multivariate extension of the Gelman-Rubin diagnostic (Brooks and Gelman, 1998) was equal to 1 rounded at two decimal places indicating good convergence properties. The individual Gelman-Rubin diagnostic (Gelman and Rubin, 1992) for all parameters were between 1 and 1.01.

At each iteration we simulated 1,000 random draws for $(\eta_{1i}, \eta_{2i})^T$, and $(\delta_{1ij}, \delta_{2ij})^T$ from $\mathrm{N}_2((m_0, m_1)^T, R)$, and $\mathrm{N}_2((0, 0)^T, D)$ respectively. We note that the error terms δ_{1ij} and δ_{2ij} were included in this analysis because the incidence is determined from the number of people marked in the early disease stage regardless of whether measurement error is

responsible for the classification. We repeated the search described in Chapter 3 calculating $\widehat{\phi}(t)$ for MAAs defined by all possible combinations of the cutoffs for the LAg Avidity assay, C_ℓ, and BioRad Avidity assay, C_b, explored above. The first step of the search was to evaluate $\widehat{\phi}(8)$ for each set of cutoffs, eliminating all MAAs where $\widehat{\phi}(8) \geq 0.001$. For those curves where $\widehat{\phi}(8) < 0.001$ we calculated $\widehat{\phi}(t)$ at each iteration for 1,101 unique times t for $t \in [0, 8]$ by using the simulated values at the specific iteration. The percentage of draws where both $\log(\ell_{ij}) < \log(C_\ell)$ ODn and $\log(b_{ij}) < \log(C_b)$ AI determined the probability curve at the specific iteration. We then calculated a posterior distribution for $\mu = \int_0^\infty \phi(t)dt$ and $\psi = \mu^{-1} \int_0^\infty t\phi(t)dt$ over all of the iterations for each MAA. Confidence intervals for μ and ψ were generated using the percentile method.

6.4 Results

6.4.1 Model Estimates

The left hand side of Figure 6.3 shows the estimated 2.5%, 25%, 75%, and 97.5% quantiles for the modeled growth of the log Lag Avidity assay result (Panel A) and the log of the BioRad Avidity assay result (Panel C). The slightly darker outer curves represent pointwise 95% credible intervals at specific times. The growth curves reflect the patterns seen in Figure 6.1 with the Lag Avidity assay having more variability in its measurements than the BioRad Avidity assay. The right hand side of Figure 6.3 shows violin plots for the log Lag Avidity assay result (Panel B) and the log BioRad Avidity assay result (Panel D) at times $t =$ 0, 1, 2, 4, 6, and 8. The violin plots are formed by rotating and mirroring the posterior distribution of the assay results at the specific times. These plots show the implausibility of ever excluding 100% of long term infections from the early disease stage. The viability of the cross-sectional incidence estimator to measure recent incidence is most sensitive to the values that $\phi(t)$ takes at larger times t. For example if we look at the formula for the time in the past when we are measuring incidence $\psi = \mu^{-1} \int_0^\infty t\phi(t)dt$ we see that each value of $\phi(t)$ is multiplied by $\mu^{-1}t$ making larger times more influential.

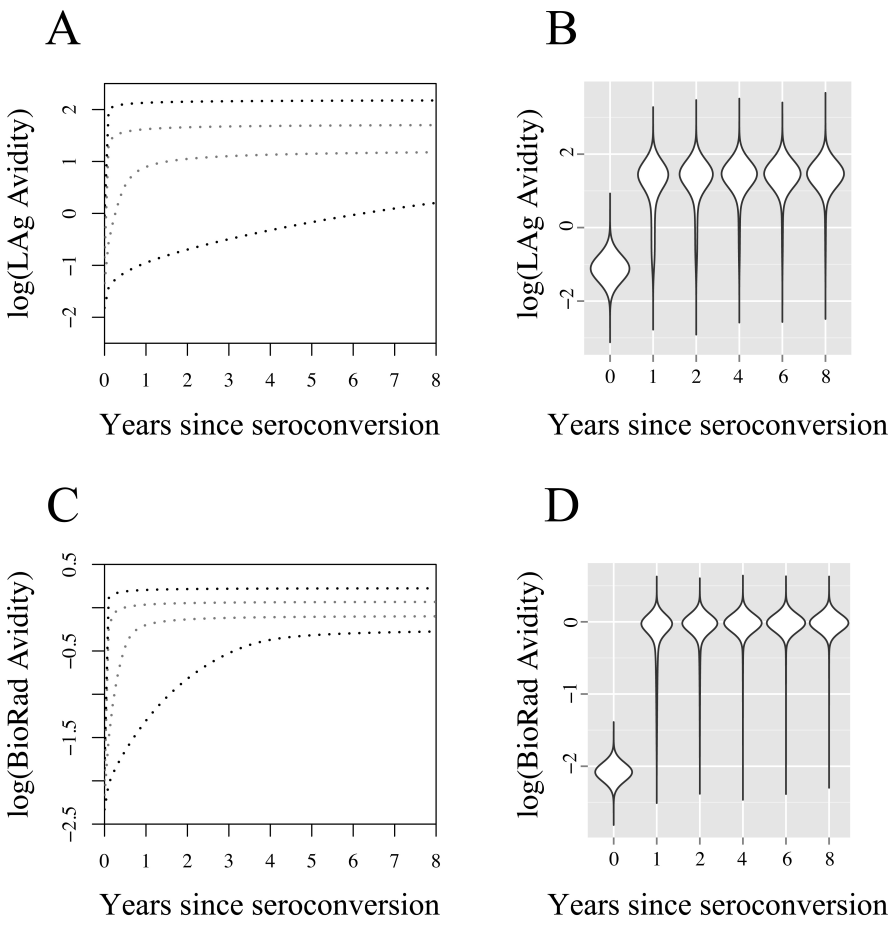

Figure 6.3: Panel A: Estimated 2.5%, 25%, 75%, and 97.5% quantiles for the modeled growth of the log Lag Avidity assay result. Panel B: Violin plot for the log Lag Avidity assay result at selected times. Panel C: Estimated 2.5%, 25%, 75%, and 97.5% quantiles for the modeled growth of the log BioRad Avidity assay result. Panel D: Violin plot for the log BioRad Avidity assay result at selected times.

For each of the 709 individuals we plotted the posterior distribution of s_i, the unknown time from seroconversion until the first positive test for HIV. The posterior distribution for six individuals is shown in figure 6.4.

Figure 6.4 shows that the there are varying degrees of learning in the model. The posterior distribution of s_i for some individuals, like the one portrayed in Panel C is essentially unchanged from the uniform prior. For other individuals, like the one portrayed in Panel F, the model strongly suggests that seroconversion occurred at a specific time. When there are multiple observations on an individual the exact shape of the posterior distribution may be

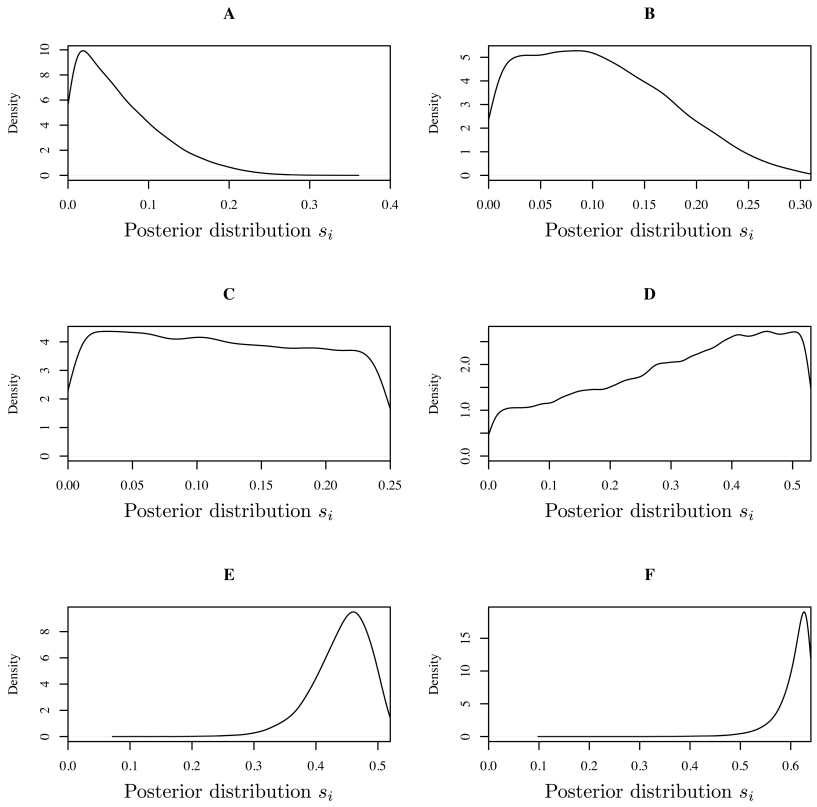

Figure 6.4: Example posterior distributions for s_i, the unknown time from seroconversion until the first positive test for HIV.

difficult to understand. However, for some individuals the logic of the model is evident. For example, Panel D corresponds to an individual with the following 5 data points.

LAg Avidity	BioRad Avidity	τ_{ij}
2.8 ODn	64%	0.15
3.6 ODn	86%	0.33
4.3 ODn	92%	0.49
4.1 ODn	98%	1.23
4.3 ODn	100%	1.48

For this person s_i is somewhere between 0 and 0.53 years. This person's biomarker measurements increase rapidly and level off in the general pattern expected by the model. The posterior mean for β_1 and β_3, corresponding to the intercepts, was 0.3 ODn for the LAg Avidity assay and 12% for the BioRad Avidity assay. The posterior mean for β_0 and β_2, corresponding to the asymptotes, was 4.39 ODn for the LAg Avidity assay and 99% for the BioRad Avidity assay. Thus it makes sense that in Panel D we see the model favoring larger

values of s_i implying, for example, that the first measurement is closer to 0.68 years after seroconversion than 0.15 years.

6.4.2 Comparison of Search Results

For the BPA the MAA with the largest estimated μ, which meets the four criteria detailed on page 14, uses the BioRad Avidity assay with a cutoff of 30%. This single assay algorithm had a larger estimated μ than all 2-assay MAAs considered which met the necessary criteria. We will call this MAA5. When we used the BCA to estimate $\phi(t)$ for this MAA the point estimate of the shadow was greater than 250 days and the upper bound of the confidence interval for the shadow was greater than 1 year. Further $\widehat{\phi}(8)$ was found to be 0.004 implying that it may be inappropriate to assume that $\phi(t) = 0$ for $t > 8$ years. Since it failed these criteria this MAA was rejected in the search conducted with the BCA. Interestingly MAA3, which for the BCA had the largest estimated μ, subject to the search criteria, was rejected in the search conducted with the BPA. Figure 6.5 plots the estimated $\phi(t)$ curves for each approach for MAA5 and the estimated $\phi(t)$ curves for each approach for MAA3.

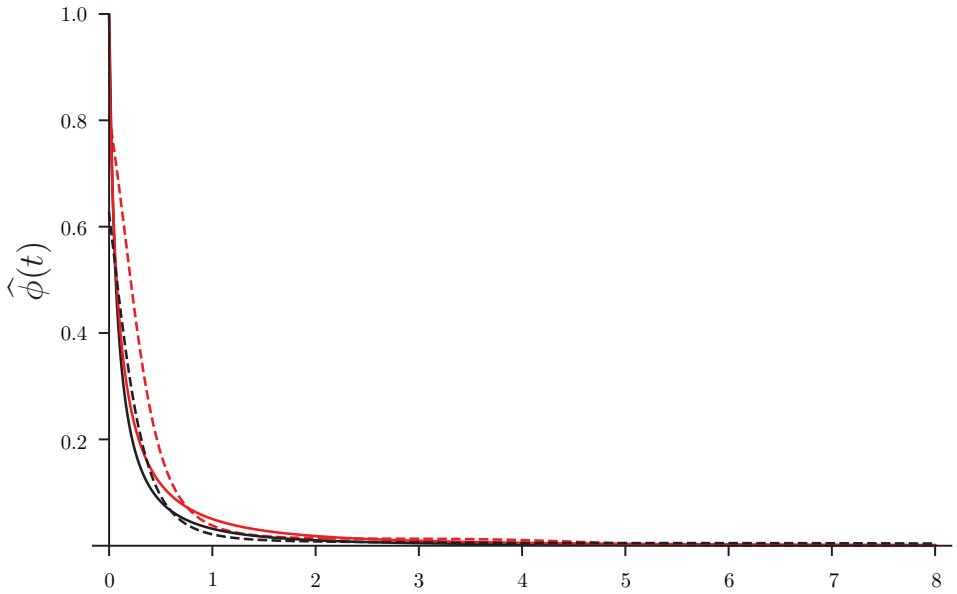

Figure 6.5: Estimated $\phi(t)$ curves for MAA5 with the BCA (dashed black line), MAA5 with the BPA (solid black line), MAA3 with the BCA (dashed red line), MAA3 with the BPA (solid red line).

MAA3 is shown in red and MAA5 is shown in black. Dashed lines are used to indicate the estimated $\phi(t)$ curves for the BCA and solid lines are used to indicated the estimated $\phi(t)$ curves for the BPA.

Since it can be difficult to see subtle differences in the curves, Table 6.1 compares each approach's estimate of $\phi(t)$ for MAA5 and estimate of $\phi(t)$ for MAA3 at times 0, 1, 2, 4, 6, and 8 years after seroconversion. Table 6.1 also shows the estimated values for μ and ψ, along with 95% confidence intervals for both MAAs and approaches.

	$\widehat{\mu}$	$\widehat{\psi}$	$\widehat{\phi}(0)$	$\widehat{\phi}(1)$	$\widehat{\phi}(2)$	$\widehat{\phi}(4)$	$\widehat{\phi}(6)$	$\widehat{\phi}(8)$
MAA5								
BCA	.21 (.15, .27)	.98 (.39, 1.61)	.625	.021	.008	.005	.005	.004
BPA	.19 (.16, .23)	.65 (.44, .91)	1.000	.032	.010	.003	.001	<0.001
MAA3								
BCA	.33 (.26, .39)	.68 (.43, .97)	.807	.038	.014	.010	<0.001	<0.001
BPA	.25 (.20, .30)	.86 (.62, 1.15)	1.000	.050	.018	.005	.003	.001

Table 6.1: Comparison of estimates for the BCA and BPA

The unknown seroconversion times account for some of the differences in the estimates. For example if we were to sample the seroconversion times in the BCA from their posterior distribution created using the BPA our point estimate of μ for MAA3 would be .29 years instead of .33 years. How we handle the unknown seroconversion times clearly influences our parameter estimates.

In the third column of Table 6.1 we see that the two methods differ in their estimates of $\widehat{\phi}(0)$. If we look back at Figure 2.2 we see that the biomarkers, at least in this data set, are only measured after seroconversion is known to have occurred. Samples taken after infection but before seroconversion are especially interesting for a variety of scientific questions and can be difficult to obtain. The value of $\widehat{\phi}(0)$ is thus based on model extrapolation when the BCA is used. The BPA also incorporates prior information into estimating $\widehat{\phi}(0)$. This prior information results in the 100% estimated inclusion in the early disease stage at the point of seroconversion. This is believed to reflect the biological reality. The differences at the moment of seroconversion are not particularly influential in the analysis. For example, for

MAA3 the two estimated curves cross less than 7 days after seroconversion. The difference in those first 7 days between the curves affects the estimate of μ by approximately 1 day.

For MAA5 one important difference is that the tail of the $\phi(t)$ curve is slightly higher when estimated with the BCA compared to the BPA. This small difference is responsible for much of the disparity we see in the estimate of the shadow. There are only 343 samples in the data set which have been collected more than 5 years after seroconversion. For the BCA this means that a single sample which is in the early disease stage past 5 years can prevent convergence to 0. In the data set there is an individual with two measurements at approximately 4 and 8 years past seroconversion. At 4 years this person's BioRad Avidity was measured as 50% and at 8 years it was measured as 29%. Since the BPA estimates the trajectory of the biomarker growth for this individual the sample at 8 years does not prevent the convergence of $\phi(t)$. The model assumes, based on other people's trajectories and the earlier measurement, that on average this person would produce a BioRad Avidity measurement above 30% if sampled repeatedly at this time.

This shows a further potential advantage of the BPA. When we find a sample taken 8 years after seroconversion which has a BioRad Avidity assay result of 29% we might imagine that there are other individuals in the population who have been infected for 8 years with BioRad Avidity assay results in the neighborhood of 29%. Since the biomarker is continuous and since there is always measurement error this is almost certainly true. The BPA determines what these similar values are by using other people's trajectories and other measurements on the same person. When we simulate values for the BioRad Avidity assay from the posterior distribution of this individual we get this range of probable values. This creates a situation where samples very close to the cutoffs do not give the same amount of evidence for or against classification in the early disease stage as those samples which are not close to the cutoffs. While the cross-sectional survey itself dichotomizes each sample, there is no reason why the external data set used to estimate $\phi(t)$ cannot learn more by using the underlying continuous biomarker measurements.

To further illustrate this point imagine that in the data set we use to estimate $\phi(t)$ we have 1,000 samples taken t years after seroconversion. If all 1,000 samples have measurements

between 95% and 100% we might feel comfortable asserting that in the cross-sectional survey 0% of samples taken t years after seroconversion will be marked in the early disease stage for MAA5, that is have an assay result below 30%. If however the 1,000 samples have measurements between 30% and 35% we naturally might think that it is more likely we will find a sample taken t years after seroconversion which is below 30%. The BCA treats the external data set the same regardless of whether the measurements are near the cutoff or not. The BPA, on the other hand, includes how close the data is to the cutoff in its estimate of $\phi(t)$.

For MAA3 the same concept makes the estimate of $\phi(t)$ fail to converge at 8 years. The individual with a BioRad Avidity assay result of 29% at 8 years has a LAg Avidity assay result of 2.9 ODn. For the BCA none of the individuals infected more than 5 years are in the early disease stage for MAA3 (LAg Avidity < 2.8 and BioRad Avidity $< 40\%$) so $\phi(t)$ converges to 0. For the BPA the measurement of 2.9 ODn indicates that a LAg Avidity result below 2.8 ODn is still very possible at 8 years. Thus the tail of the $\phi(t)$ curve for the BPA is slightly higher.

6.5 Extensions to More Biomarkers

A CD4 cell count below 50 cells/mm^3 is indicative of immune collapse, a process which may result in low BioRad Avidity and Lag Avidity assay results late in infection. An HIV viral load below 400 copies/mL indicates viral suppression which is often due to treatment with antiretroviral drugs. Treatment with antiretroviral drugs has been shown to alter the progression of serologic biomarkers (Kassanjee et al., 2014). MAA2 which incorporated these concepts defined the early disease stage as any sample which had a CD4 cell count > 50 cells/mm^3, an HIV viral load of > 400 copies/mL, a Lag Avidity assay result < 2.9 ODn, and a BioRad Avidity assay result $< 85\%$ AI. The serologic assays have higher cutoffs in MAA2 then MAA3 and MAA5 since there is likely to be a more defined growth pattern for the biomarkers once samples with CD4 cell count ≤ 50 cells/mm^3 and HIV viral load of ≤ 400 copies/mL are removed. This is illustrated in the profile plots of the data shown in

Figure 6.6 which may be compared to the similar plots shown in Figure 6.1. We see that many of the lower measurements at later times have been removed.

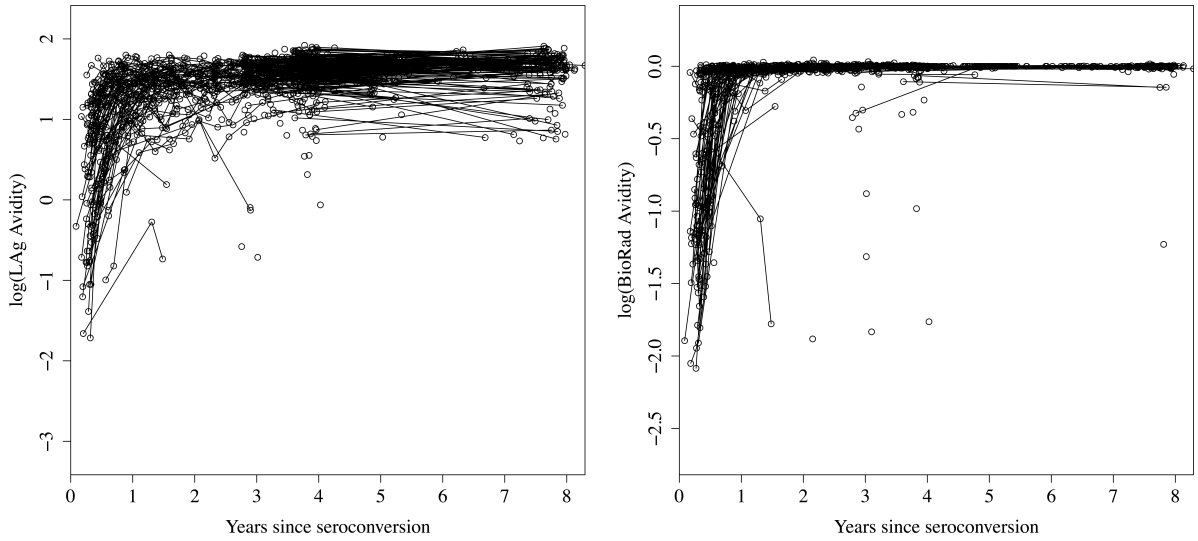

Figure 6.6: Profile plots illustrating the growth of the log LAg Avidity and log BioRad Avidity assay results for samples with CD4 cell counts > 50 cells/mm^3 and HIV viral loads > 400 copies/mL (1,338 samples from 623 people).

To model the probability curve for this 4-assay MAA we write $\phi(t_{ij}) = \mathrm{P}\{(v_{ij} > 400) \cap (c_{ij} > 50) \cap (\ell_{ij} < 2.9) \cap (b_{ij} < 85)|t_{ij}\} = \mathrm{P}\{v_{ij} > 400|t_{ij}\}\mathrm{P}\{c_{ij} > 50|v_{ij} > 400, t_{ij}\}\mathrm{P}\{(\ell_{ij} < 2.9) \cap (b_{ij} < 85)|c_{ij} > 50, v_{ij} > 400, t_{ij}\}$ where c_{ij} and v_{ij} are the measurements of CD4 cell count and HIV viral load. We separate the model into three parts since the expected progression of the BioRad Avidity and Lag Avidity assay results is conditional on a lack of immune collapse and viral suppression. Similarly CD4 cell count follows a somewhat predictable pattern only in the absence of antiretroviral drugs which can suppress the virus and allow the immune system to partially recover.

The last term in the expression can be modeled as described in Section 6.1 in the subset of the data where the samples have CD4 cell count measures above 50 cells/mm^3 and HIV viral load measures above 400 copies/mL. The decline of the CD4 cell count in individuals who are not on antiretrovirals has been modeled extensively (Taylor et al., 1994; Taylor and Law, 1998; Lange et al., 1992; Degruttola et al., 1991; Munoz et al., 1988) and these models could be used to find $\mathrm{P}\{c_{ij} > 50|v_{ij} > 400, t_{ij}\}$ based on the same procedure described in Section 6.1. The remaining term, $\mathrm{P}\{v_{ij} > 400|t_{ij}\}$, representing the probability of viral suppression

as a function of time since seroconversion cannot be modeled directly from cohort studies as individuals in these studies are tested regularly for HIV. Thus individuals in cohort studies are more likely to be aware of their HIV status and linked to care than people in the general population. This implies that the probability of viral suppression for participants in a cohort study is likely not representative of the general population. Because of this complication we leave the full adaptation of the BPA to this type of MAA to future work.

6.6 Discussion

In this chapter we presented a second approach for estimating the probability that a sample, drawn from a new person t years after seroconversion, is marked in the early disease stage. We have detailed some apparent advantages of this approach over the previous methodology. The advantages stem from modeling the continuous longitudinal data instead of condensing all this information into a binary outcome variable denoting presence or absence in the early disease stage. The added information in the biomarker measurements more readily allows for modeling the within-subject correlation between measurements. Further, the model updates the likely seroconversion times based on the observed biomarker measurements. This approach also allows for the data to be modeled a single time and not depend on the MAA under consideration. The previous approach, detailed in Chapters 2 and 3, offers consistent estimation under minimal assumptions. The tradeoff with the new approach is the additional data modeling which is necessary for its implementation.

The two approaches produce different optimal MAAs. MAA3, which was considered optimal in the search conducted in Chapter 3, is eliminated in the search conducted in this chapter because it may estimate incidence more than one year in the past. Similarly, MAA5 which is considered optimal in the search conducted in this chapter, was eliminated in Chapter 3. A validation data set could help determine which estimates are likely more accurate. The point estimates of the shadow for both MAAs and both approaches remain less than one year. Public health practitioners may still be willing to use MAA3 and MAA5 in practice, despite the fact that the confidence intervals for the shadow stretch beyond

one year. Further progress could be made by including non-serologic biomarkers such as CD4 cell count and HIV viral load. However these assays are expensive and further work will be needed to model the probability of viral suppression as a function of the time since seroconversion.

CHAPTER 7

Conclusion

In this dissertation we explored various methodological questions which arose from the practical public health necessity of accurately estimating the rate at which new HIV infections occur in populations. Measuring HIV incidence is crucial for tracking the epidemic and conducting prevention trials aimed at curbing its spread.

In Chapter 1 we introduced cross-sectional incidence estimation and explained some of its potential advantages over the more traditional approach of estimating incidence through longitudinal studies. The cross-sectional approach is anchored in the idea that duration of infection will dictate the biomarker profile in a person's blood. This profile is then used to group people into disease stages associated with either recent or long term infections. In the cross-sectional survey the number of people found in a biomarker-defined early disease stage, relative to the size of the susceptible part of the population, gives an estimate of how the epidemic is progressing. By deriving the cross-sectional incidence estimator we showed that the estimator depends on two population parameters which must be estimated before the cross-sectional survey is conducted. The average amount of time spent in the early disease stage, μ, must be large to control the variance of the estimator, while ψ, the time in the past when the estimator is approximating incidence, must be temporally close to the time of the survey in order to learn about recent infections. These two parameters are both functions of $\phi(t)$, the probability that a person will remain in the early disease stage t years after seroconversion.

In Chapter 2 we introduced the first published multi-assay algorithm which defines the early disease stage based on multiple biomarkers. Previous approaches for estimating HIV incidence focused on using a single biomarker, along with an HIV antibody test, to define

the early disease stage. We presented an approach for estimating $\phi(t)$ based on modeling a binary variable, indicating presence or absence in the early disease stage, as a function of time. Our point estimates of μ and ψ showed that current single biomarker approaches are impractical. By using multiple biomarkers we were able to define the early disease stage so as to increase μ, reducing the variability, and decrease, ψ, controlling the bias in the estimator.

In Chapter 3 we discussed various statistical approaches for defining the early disease stage which produced multi-assay algorithms which could be considered optimal in one sense or another. We compared the estimates of incidence from the cross-sectional approach with previous estimates from longitudinal studies. We concluded that the cross-sectional approach appeared to generate estimates in the correct ballpark. We then explored potential issues of bias, both in the underlying methodology and in how we selected our optimal multi-assay algorithms. We found that the bias in our search for optimal algorithms was likely minimal. The bias in the cross-sectional incidence estimator was shown to be very sensitive to the values $\phi(t)$ takes at later times. Viable multi-assay algorithms need the probability that a person will remain in the early disease stage for more than one year after seroconversion to be close to zero. We also looked at the optimal ordering of the component assays in the multi-assay algorithms. We showed that since CD4 cell count must be used as the first component assay, the cost of including CD4 cell count is high. Multi-assay algorithms which rely on only serologic component assays are therefore less costly.

We then turned to sample size methodology. In Chapter 4 we developed methods for determining the sample sizes needed for both single and consecutive cross-sectional surveys. In the single survey case we showed how to choose a sample size to control the expected precision with which we will estimate incidence. When we have successive surveys we showed how to set the sample sizes to achieve a desired level of power to detect differences in incidence at two time points or in two populations. In these calculations we found that the underlying incidence is inversely proportional to the needed sample sizes. We also showed that one must account for the fact that μ is estimated and not known. We can account for this uncertainty by placing a distribution on μ. We showed that larger sample sizes in the cross-sectional survey itself cannot compensate for large amounts of uncertainty in μ. We also showed that

there are limits to the amount of precision which can be achieved based on the underlying incidence and the chosen multi-assay algorithm. Multi-assay algorithms where individuals spend longer durations on average in the early disease stage require smaller samples sizes. Similarly, for a specific power, the sample sizes can be prohibitive if μ is too small, the underlying incidence is too low, or the difference between the two incidences we are trying to detect is not sufficiently large. We tested our methodology in four different simulated underlying epidemics and found that the methods performed well.

We next explored issues of missing data in Chapter 5. Here we explained how to use multiple multi-assay algorithms to handle missing biomarker data in the field while conducting the cross-sectional survey. This allows us to use multi-assay algorithms with larger estimated μ even if they are likely to have a certain amount of missing data. We also explored how to estimate $\phi(t)$ when there are missing biomarker measurements in the training data set. This method allows for the potential benefits of newly developed biomarkers to be evaluated, within the context of a multi-assay algorithm, even if some of the samples in the training data set have been depleted and cannot be tested with new biomarkers.

Lastly, we presented a second approach for estimating $\phi(t)$ based on a multivariate model of the growth of the continuous underlying biomarkers. While this approach requires stronger assumptions about the structure of the data than the approach we presented in Chapter 2, it offers several potential advantages. By modeling the growth of the underlying biomarkers we are able to naturally incorporate ideas like within-subject correlation and the fact that the biomarker measurements contain information on the unknown amount of time from seroconversion until the first positive HIV antibody result. Further, since this approach models the original data, a new model does not need to be fit for each definition of the early disease stage. We repeated our search for an optimal 2-assay multi-assay algorithm with this new approach and found that the two methods produce different results. We were able to identify reasons for the disagreements between the two approaches. Further work is needed to compare the two methods and to expand the new methodology to include additional biomarkers.

References

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, New Jersey: John Wiley and Sons.

Bacchetti, P., and Moss, A. R. (1989). Incubation period of AIDS in San Francisco. *Nature*, *338*(6212), 251–253.

Beesack, P. R. (1966). A general form of the remainder in Taylor's theorem. *The American Mathematical Monthly*, *73*(1), 64–67.

Brookmeyer, R. (1997). Accounting for follow-up bias in estimation of human immunodeficiency virus incidence rates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(1), 127–140. doi: 10.1111/1467-985X.00049

Brookmeyer, R. (2010a). Measuring the HIV/AIDS epidemic: Approaches and challenges. *Epidemiologic Reviews*, *32*(1), 26–37. doi: 10.1093/epirev/mxq002

Brookmeyer, R. (2010b). On the statistical accuracy of biomarker assays for HIV incidence. *Journal of Acquired Immune Deficiency Syndromes*, *54*(4), 406–414. doi: 10.1097/QAI.0b013e3181dc6d2c

Brookmeyer, R., and Goedert, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated AIDS. *Biometrics*, *45*(1), 325–335.

Brookmeyer, R., Konikoff, J., Laeyendecker, O., and Eshleman, S. H. (2013). Estimation of HIV incidence using multiple biomarkers. *American Journal of Epidemiology*, *177*(3), 264–272.

Brookmeyer, R., and Quinn, T. C. (1995). Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American Journal of Epidemiology*, *141*(2), 166–172.

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. doi: 10.1080/10618600.1998.10474787

Brown, C. C., and Green, S. B. (1982). Additional power computations for designing comparative Poisson trials. *American Journal of Epidemiology*, *115*(5), 752–758.

Casella, G., and Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Wadsworth.

Celum, C. L., Buchbinder, S. P., Donnell, D., Douglas, J. M., Mayer, K., Koblin, B., Marmor, M., Bozeman, S., Grant, R. M., Flores, J., and HW, S. (2001). Early human immunodeficiency virus (HIV) infection in the HIV Network for Prevention Trials Vaccine Preparedness Cohort: Risk behaviors, symptoms, and early plasma and genital tract virus load. *Journal of Infectious Diseases*, *183*(1), 23–35.

Cousins, M. M., Donnell, D., and Eshleman, S. H. (2013). Impact of mutation type and amplicon characteristics on genetic diversity measures generated using a high-resolution melting diversity assay. *The Journal of Molecular Diagnostics*, *15*(1), 130–137. doi: 10.1016/j.jmoldx.2012.08.008

Cousins, M. M., Konikoff, J., Laeyendecker, O., Celum, C., Buchbinder, S. P., Seage, G. R., Kirk, G. D., Moore, R. D., Mehta, S. H., Margolick, J. B., Brown, J., Mayer, K. H., Koblin, B. A., Wheeler, D., Justman, J. E., Hodder, S. L., Quinn, T. C., Brookmeyer, R., and Eshleman, S. H. (2014). Hiv diversity as a biomarker for HIV incidence estimation: Including a high-resolution melting diversity assay in a multiassay algorithm. *Journal of Clinical Microbiology*, *52*(1), 115–121. doi: 10.1128/JCM.02040-13

Cousins, M. M., Konikoff, J., Sabin, D., Khaki, L., Longosz, A. F., Laeyendecker, O., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Moore, R. D., Mehta, S. H., Margolick, J. B., Brown, J., Mayer, K. H., Kobin, B. A., Wheeler, D., Justman, J. E., Hodder, S. L., Quinn, T. C., Brookmeyer, R., and Eshleman, S. H. (2014). A comparison of two measures of HIV diversity in multi-assay algorithms for HIV incidence estimation. *PloS One*, *9*(6), e101043. doi: 10.1371/journal.pone.0101043

Cousins, M. M., Laeyendecker, O., Beauchamp, G., Brookmeyer, R., Towler, W. I., Hudelson, S. E., Khaki, L., Koblin, B., Chesney, M., Moore, R. D., Kelen, G. D., Coates, T., Celum, C., Buchbinder, S. P., Seage, G. R., III, Quinn, T. C., Donnell, D., and H, E. S. (2011). Use of a high resolution melting (HRM) assay to compare gag, pol, and env diversity in adults with different stages of HIV infection. *PLoS One*, *6*(11), e27211. doi: 10.1371/journal.pone.0027211

Cousins, M. M., Ou, S.-S., Wawer, M. J., Munshaw, S., Swan, D., Magaret, C. A., Mullis, C. E., Serwadda, D., Porcella, S. F., Gray, R. H., Quinn, T. C., Donnell, D., Eshleman, S. H., and Redd, A. D. (2012). Comparison of a high resolution melting (HRM) assay to next generation sequencing for analysis of HIV diversity. *Journal of Clinical Microbiology*, *50*(9), 3054-3059. doi: 10.1128/JCM.01460-12

Cousins, M. M., Swan, D., Magaret, C. A., Hoover, D. R., and Eshleman, S. H. (2012). Analysis of HIV using a high resolution melting (HRM) diversity assay: Automation of HRM data analysis enhances the utility of the assay for analysis of HIV incidence. *PloS One*, *7*(12), e51359. doi: 10.1371/journal.pone.0051359

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap methods and their application* (1st ed.). Cambridge, United Kingdom: Cambridge University Press.

Degruttola, V., Lange, N., and Dafni, U. (1991). Modeling the progression of HIV infection. *Journal of the American Statistical Association*, *86*(415), 569–577.

DiCiccio, T. J., and Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, *50*(3), 338–354.

Dobbs, T., Liu, X., Anderson, R., Nkengasong, J., and Parekh, B. S. (2011). A comprehensive evaluation of the proficiency testing program for the HIV-1 BED incidence assay. *Journal of Clinical Microbiology*, *49*(10), 3470-3473. doi: 10.1128/JCM.01122-11

Duong, Y. T., Qiu, M., De, A. K., Jackson, K., Dobbs, T., Kim, A. A., Nkengasong, J. N., and Parekh, B. S. (2012). Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: Potential for HIV-1 incidence estimates and avidity maturation studies. *PLoS One*, *7*(3), e33328. doi: 10.1371/journal.pone.0033328

Field, C. A., and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390.

Freeman, J., and Hutchison, G. B. (1980). Prevalence, incidence and duration. *American Journal of Epidemiology*, *112*(5), 707–723.

Gail, M. (1974). Power computations for designing comparative Poisson trials. *Biometrics*, *30*(2), 231–237.

Gelman, A., and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical*

models. Cambridge, United Kingdom: Cambridge University Press.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Hall, H. I., Song, R., Rhodes, P., Prejean, J., An, Q., Lee, L. M., Karon, J., Brookmeyer, R., Kaplan, E. H., McKenna, M. T., and Janssen, R. S. (2008). Estimation of HIV incidence in the United States. *JAMA*, *300*(5), 520–529. doi: 10.1001/jama.300.5.520

Hargrove, J., Eastwood, H., Mahiane, G., and van Schalkwyk, C. (2012). How should we best estimate the mean recency duration for the BED method? *PloS One*, *7*(11), e49661. doi: 10.1371/journal.pone.0049661

Hodder, S. L., Justman, J., Hughes, J. P., Wang, J., Haley, D. F., Adimora, A. A., Del Rio, C., Golin, C. E., Kuo, I., Rompalo, A., Soto-Torres, L., Mannheimer, S. B., Johnson-Lewis, L., Eshleman, S. H., and El-Sadr, W. M. (2013). HIV acquisition among women from selected areas of the United States: A cohort study. *Annals of Internal Medicine*, *158*(1), 10–18. doi: 10.7326/0003-4819-158-1-201301010-00004

Kaplan, E. H., and Brookmeyer, R. (1999). Snapshot estimators of recent HIV incidence rates. *Operations Research*, *47*(1), 29–37.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo Jr., C. R. (1987). The Multicenter AIDS Cohort Study: Rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, *126*(2), 310–318.

Kassanjee, R., Pilcher, C. D., Keating, S. M., Facente, S. N., McKinney, E., Price, M. A., Martin, J. N., Little, S., Hecht, F. M., Kallas, E. G., Welte, A., Busch, M. P., and Murphy, G. (2014). Independent assessment of candidate HIV incidence assays on specimens in the CEPHIA repository. *AIDS*, *28*(16), 2439-2449. doi: 10.1097/QAD.0000000000000429

Koblin, B. A., Mayer, K. H., Eshleman, S. H., Wang, L., Mannheimer, S., Del Rio, C., Shoptaw, S., Magnus, M., Buchbinder, S., Wilton, L., Liu, T.-Y., Cummings, V., Piwowar-Manning, E., Fields, S. D., Griffith, S., Elharrar, V., and Wheeler, D. (2013). Correlates of HIV acquisition in a cohort of Black men who have sex with men in the

United States: HIV Prevention Trials Network (HPTN) 061. *PLoS One*, *8*(7), e70413. doi: 10.1371/journal.pone.0070413

Konikoff, J., and Brookmeyer, R. (Accepted for publication April 2015). Sample size methods for estimating HIV incidence from cross-sectional surveys. *Biometrics*.

Konikoff, J., Brookmeyer, R., Longosz, A. F., Cousins, M. M., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Moore, R. D., Mehta, S. H., Margolick, J. B., Brown, J., Mayer, K. H., Koblin, B. A., Justman, J. E., Hodder, S. L., Quinn, T. C., Eshleman, S. H., and Laeyendecker, O. (2013). Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of HIV incidence in the United States. *PloS One*, *8*(12), e82772. doi: 10.1371/journal.pone.0082772

Kouyos, R. D., von Wyl, V., Yerly, S., Böni, J., Rieder, P., Joos, B., Taffé, P., Shah, C., Bürgisser, P., Klimkait, T., Weber, R., Hirschel, B., Cavassini, M., Rauch, A., Battegay, M., Vernazza, P. L., Bernasconi, E., Ledergerber, B., Bonhoeffer, S., and Günthard, H. F. (2011). Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clinical Infectious Diseases*, *52*(4), 532–539. doi: 10.1093/cid/ciq164

Laeyendecker, O., Brookmeyer, R., Cousins, M. M., Mullis, C. E., Konikoff, J., Donnell, D., Celum, C., Buchbinder, S. P., Seage, G. R., III, Kirk, G. D., Mehta, S. H., Astemborski, J., Jacobson, L. P., Margolick, J. B., Brown, J., Quinn, T. C., and Eshleman, S. H. (2012). HIV incidence determination in the United States: A multi-assay approach. *Journal of Infectious Diseases*, *207*(2), 232-239. doi: 10.1093/infdis/jis659

Lange, N., Carlin, B. P., and Gelfand, A. E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *Journal of the American Statistical Association*, *87*(419), 615–626.

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Masciotra, S., Dobbs, T., Candal, D., Hanson, D., Delaney, K., Rudoplh, D., Charurat, M., Harrigan, R., McDougal, S., and Owen, M. (2010, February). Antibody avidity-based assay for identifying recent HIV-1 infections based on Genetic Systems TM 1/2

plus O EIA. In *Abstract 937 at the 17th conference on retroviruses and opportunistic infections.* San Francisco, CA.

Moore, R. D. (1998). Understanding the clinical and economic outcomes of HIV therapy: The Johns Hopkins HIV clinical practice cohort. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, *17*, S38–S41.

Munoz, A., Carey, V., Saah, A. J., Phair, J. P., Kingsley, L. A., Fahey, J. L., Ginzburg, H. M., and frank Polk, B. F. (1988). Predictors of decline in CD4 lymphocytes in a cohort of homosexual men infected with human immunodeficiency virus. *Journal of Acquired Immune Deficiency Syndromes*, *1*(4), 396–404.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, *44*(4), 1033–1048.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., and Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, *37*(9), 1487–1498. doi: 10.1080/02664760903046102

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X.-L., and Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology*, *73*(12), 10489–10502.

Suligoi, B., Massi, M., Galli, C., Sciandra, M., Di Sora, F., Pezzotti, P., Recchia, O., Montella, F., Sinicco, A., and Rezza, G. (2003). Identifying recent HIV infections using the avidity index and an automated enzyme immunoassay. *Journal of Acquired Immune Deficiency Syndromes*, *32*(4), 424–428.

Sweeting, M. J., De Angelis, D., Parry, J., and Suligoi, B. (2010). Estimating the distribution of the window period for recent HIV infections: A comparison of statistical methods. *Statistics in Medicine*, *29*(30), 3194–3202. doi: 10.1002/sim.3941

Taylor, J. M. G., Cumberland, W. G., and Sy, J. P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association*, *89*(427), 727–736.

Taylor, J. M. G., and Law, N. (1998). Does the covariance structure matter in longitudinal

modelling for the prediction of future CD4 counts? *Statistics in Medicine*, *17*(20), 2381–2394.

Towler, W. I., James, M. M., Ray, S. C., Wang, L., Donnell, D., Mwatha, A., Guay, L., Nakabiito, C., Musoke, P., Jackson, J. B., and Eshleman, S. H. (2010). Analysis of HIV diversity using a high-resolution melting assay. *AIDS Research and Human Retroviruses*, *26*(8), 913–918. doi: 10.1089/aid.2009.0259

UNAIDS Epidemiology Reference Group Secretariat. (2005). *UNAIDS Reference Group on Estimates, Modelling and Projections' statement on the use of the BED-assay for the estimation of HIV-1 incidence for surveillance or epidemic monitoring.* Retrieved from http://www.epidem.org/sites/default/files/content/resources/attachments/BED%20statement.pdf.

United States Census Bureau. (2014). *Population Clock.* Retrieved from http://www.census.gov/popclock/.

Vlahov, D., Anthony, J. C., Munoz, A., Margolick, J., Nelson, K. E., Celentano, D. D., Solomon, L., and Polk, B. F. (1990). The ALIVE study, a longitudinal study of HIV-1 infection in intravenous drug users: Description of methods and characteristics of participants. *NIDA research monograph*, *109*, 75–100.

Zeger, S. L., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121–130.