LIKELIHOOD BASED STATISTICAL METHODS FOR ESTIMATING HIV INCIDENCE RATE

By

Lesego Gabaitiri

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
UNIVERSITY OF KWAZULU-NATAL
PIETERMARITZBURG, SOUTH AFRICA
18 MARCH 2013

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

Author: ______________________________
Lesego Gabaitiri

Research Supervisor: ______________________________
Professor Henry G. Mwambi

UNIVERSITY OF KWAZULU-NATAL

Date: **18 March 2013**

Author: **Lesego Gabaitiri**

Title: **Likelihood based statistical methods for estimating HIV incidence rate**

School: **Mathematics, Statistics and Computer Science**

Signature of Author

DECLARATION

I, Lesego Gabaitiri, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other person's data, pictures, graphs, or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other person's writing, unless specifically acknowledged as being sourced from other researchers. Where other sources have been quoted, then:
   a. Their words have been re-written, but the general information attributed to them has been referenced.
   b. Where their exact words have been used, their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics, or tables copied and pasted from the internet, unless specifically acknowledged and the source being detailed in the thesis and in the references section.

Signature of Author

PUBLICATIONS

1. Gabaitiri, L. and Mwambi, H. G. (2012). Estimating HIV Incidence with adjustment for Sensitivity and Specificity. *Presented at SUB-SAHARAN AFRICA NETWORK (SUSAN) OF THE INTERNATIONAL BIOMETRIC SOCIETY (IBS) CONFERENCE, Gaborone, July 2011*
2. Gabaitiri, L., Mwambi, H. G., Lagakos, S. W., and Pagano, M. (2013). A likelihood estimation of HIV incidence incorporating information on past prevalence. *South African Statistical Journal, **47**, page 15 to 31*
3. Gabaitiri, L., Mwambi, H. G., Lagakos, S. W., and Pagano, M. (2012b). Estimation of HIV incidence under Linear incidence density function. *Abstract Accepted for the 53rd ANNUAL CONFERENCE OF THE SOUTH AFRICAN STATISTICAL ASSOCIATION (SASA), 2011, Pretoria, South Africa*
4. Gabaitiri, L., Mwambi, H. G., Lagakos, S. W., and Pagano, M. (Submitted). Estimation of HIV incidence under Linear incidence density function. *Submitted to Statistics in Medicine (Submission Number: SIM-12-0715)*
5. Gabaitiri, L. and Mwambi, H. G. (In preparation). Estimation of HIV Incidence from a Cross-Sectional Sample with Missing Data
6. Gabaitiri, L. and Mwambi, H. G. (In preparation). Incorporating the effect of covariates in the estimation of HIV incidence

To Prof Stephen W. Lagakos and Mr Gabaitiri Modise

Table of Contents

Abstract

Estimation of current levels of human immunodeficiency virus (HIV) incidence is essential for monitoring the impact of an epidemic, determining public health priorities, assessing the impact of interventions and for planning purposes. However, there is often insufficient data on incidence as compared to prevalence. A direct approach is to estimate incidence from longitudinal cohort studies. Although this approach can provide direct and unbiased measure of incidence for settings where the study is conducted, it is often too expensive and time consuming. An alternative approach is to estimate incidence from cross sectional survey using biomarkers that distinguish between recent and non-recent/longstanding infections. The original biomarker based approach proposes the detection of HIV-1 p24 antigen in the pre-seroconversion period to identify persons with acute infection for estimating HIV incidence. However, this approach requires large sample sizes in order to obtain reliable estimates of HIV incidence because the duration of antigenemia before antibody detection is short, about 22.5 days. Subsequently, another method that involves dual antibody testing system was developed. In stage one, a sensitive test is used to diagnose HIV infection and a less sensitive test such is used in the second stage to distinguish between long standing infections and recent infections among those who tested positive for HIV in stage one. The question is: how do we combine this data with other relevant information, such as the period an individual takes from being undetectable by a less sensitive test to being detectable, to estimate incidence?

The main objective of this thesis is therefore to develop likelihood based methods that can be used to estimate HIV incidence when data is derived from cross sectional surveys and the disease classification is achieved by combining two biomarker or assay tests. The thesis builds on the dual antibody testing approach and extends the statistical framework that uses the multinomial distribution to derive the maximum likelihood estimators of HIV incidence for different settings.

In order to improve incidence estimation, we develop a model for estimating HIV incidence that incorporate information on the previous or past prevalence and derive maximum likelihood estimators of incidence assuming incidence density is constant over a specified period. Later, we extend the method to settings where a proportion of subjects remain non-reactive to a less sensitive test long after seroconversion.

Diagnostic tests used to determine recent infections are prone to errors. To address this problem, we considered a method that simultaneously makes adjustment for sensitivity and specificity. In addition, we also showed that sensitivity is similar to the proportion of subjects who eventually transit the “recent infection” state.

We also relax the assumption of constant incidence density by proposing linear incidence density to accommodate settings where incidence might be declining or increasing.

We extend the standard adjusted model for estimating incidence to settings where some subjects who tested positive for HIV antibodies were not tested by a less sensitive test resulting in missing outcome data. Models for the risk factors (covariates) of HIV incidence are considered in the last but one chapter. We used data from Botswana AIDS Impact (BAIS) III of 2008 to illustrate the proposed methods. The general conclusion and recommendations for future work are provided in the final chapter.

KEY WORDS: HIV incidence; Cohort studies; Cross sectional surveys; Maximum likelihood; previous prevalence; incidence density; sensitivity and specificity; risk factors.

Acknowledgements

This PhD thesis is a special dedication to my Mentor Professor Stephen W. Lagakos who died from a car accident in October 10, 2009 and my Father Mr Gabaitiri Modise who died on the May 16, 2009. They will always be in my memory.

My sincerest thanks go to my supervisor, Prof Henry G. Mwambi, for his guidance, suggestions and constant support throughout this research. I am also thankful to the late Prof Stephen W. Lagakos for his guidance at the early stages of this research. The project was initiated jointly by the late Prof Stephen W. Lagakos and Prof Henry G. Mwambi during the latter's sabbatical visit to the Biostatistics Department, Johns Hopkins University in 2008.

I would like to thank Prof Marcello Pagano and Dr Rui Wang from Harvard School of Public Health for their contributions and support in the current work.

I would also like to thank the Government of Botswana through the Central Statistics Office (CSO) and National AIDS Coordinating Agency (NACA) for providing the data for this research.

Of course, I am grateful to my mother and my family as a whole for there patience and love. Without there support, this work would never have come into existence.

Finally, I wish to thank the following: Jost Sinvula Isaac, Denman Molemi Pule, Onalenna Nkipa Silas, Thapelo Thapelo, Dr Peter Sebina, Dr Ahmed A Khidir, Dr Keamogetse Setlhare and Prof Keoagile Thaga and for all their support and encouragement.

Pietermaritzburg, South Africa Lesego Gabaitiri
March 18, 2013

Chapter 1

General Introduction

1.1 Introduction

Knowledge of current levels of human immunodeficiency virus (HIV) prevalence and incidence is essential for monitoring the impact of HIV, determining public health priorities, assessing the impact of interventions, identifying high-risk populations for vaccine and other HIV preventions trials and for planning purposes (UNAIDS, 2006; Brookmeyer et al., 1995; Brookmeyer and Quinn, 1995; Janssen et al., 1998; Balasubramanian and Lagakos, 2010; Wang and Lagakos, 2009, 2010; Kaplan and Brookmeyer, 1999). Prevalence, a proportion of the population with a condition at a point in time, can be estimated from cross sectional surveys such as Antenatal Sentinel and Demographic and Health Surveys. Most Antenatal Sentinel Prevalence Surveys focus on the estimation of HIV prevalence among pregnant women of age 15-49 years which are not representative of the target population. Hence the results cannot be generalized to the entire population. Demographic and Health Surveys, which are based on a representative sample, are commonly used to estimate HIV prevalence.

These surveys therefore provide a better estimate of prevalence compared to that from Antenatal Sentinel Prevalence Surveys.

Unlike HIV incidence rate (or HIV incidence), the number of new cases of HIV disease condition divided by the total time experienced by all subjects followed over an interval of time, prevalence alone is not a good measure that can be used to assess the impact of interventions and track the growth of new infections since it only measure the relative burden of HIV disease. In particular, one of the major public health objectives is to reduce incidence of HIV (Guy et al., 2009). This is because reduction in HIV incidence will lead to a decline in HIV prevalence since incidence is part of prevalence. Therefore monitoring the current levels of incidence will be essential for establishing the need for prevention programmes and their effectiveness. However, there is often insufficient data on the current levels of HIV incidence as compared to HIV prevalence. This is so because there are challenges in determining the best strategy for measuring HIV incidence. A good strategy will help us to obtain reliable estimates of HIV incidence.

Several approaches have been used to estimate HIV incidence which include: cohort studies, mathematical models, back-calculation methods, serial prevalence surveys and cross-sectional surveys of biomarkers.

Regarded as a gold standard, longitudinal cohort studies have been used to estimate HIV incidence. The idea is to follow a representative sample of HIV free individuals for a specified period and record new cases of HIV infection. HIV incidence is then computed as a ratio of the new cases of HIV detected over a given period to the number of person years (or any other units of time used) of exposure (Kaplan and

Brookmeyer, 1999). Although this approach provides the ideal method of estimating HIV incidence, there are a number of drawbacks associated with it. First, estimation of incidence rate from cohort studies is hampered by cost and low follow up rates. This is a major problem in poor resource settings. Low follow-up rates could lead to follow up bias which arises if the incidence rates among the subgroups of individuals who return for follow-up are different from incidence rates among those who do not return for follow-up (Brookmeyer and Quinn, 1995; Brookmeyer et al., 1995). The difference could be a result of repeated exposure to counseling, better treatment adherence, and other health education programs and prevention messages amongst those who return for follow-up visits (Brookmeyer, 2010a; Brookmeyer et al., 1995). More generally, the enrollment of persons into a cohort study often leads to behaviour changes that result in a lower observed HIV incidence than in the broader population of interest. Second, we have the selection bias. This type of bias arises when individual who agree to participate and return for follow-up visits are not representative of the target population (Brookmeyer, 2010a). Hence, generalizing the estimates of incidence rate from the cohort to the broader population could be misleading. Lastly, the utility of the cohort approach relies on the assumption that the sample in the longitudinal cohort study is an unbiased subset of the population of interest. If this assumption does not hold then the HIV incidence estimated from the sample may not be a true reflection of the underlying rate. Consequently, generalizability will be not be achievable.

HIV incidence have also been estimated from serial prevalence data because of the availability of such data from surveys such as Demographic and Health Surveys which are often carried out every 4-5 years in some countries (Brookmeyer, 2010a). As noted by Brookmeyer (2010a), the basic idea is to infer absolute HIV incidence from changes in absolute HIV prevalence rather from changes in their respective proportions using a balancing equation. However, this approach is affected by both migration changes and mortality.

Mathematical models of serial cross sectional data on HIV prevalence have also been used to estimate HIV incidence (Lagakos and Gable, 2008). These models have been used particulary by UNAIDS and World Health Organization to provide estimates of population level trends in HIV epidemic (Lagakos and Gable, 2008; Walker et al., 2003). Like any other models for estimating HIV incidence, these models have strengths and weaknesses. The strength of this method is that beside being cheap and quick, it can be useful in tracking changes in HIV epidemic at a population level (Lagakos and Gable, 2008). However, its major drawback is that there is often insufficient information on the input parameters required to estimate HIV incidence.

Back-calculation methods provides an indirect approach to estimating incidence. Back-calculation methods have been applied to reconstruct historical infection rates using AIDS incidence data and the probability distribution of the incubation period from HIV infection to AIDS diagnosis (Brookmeyer, 1991; Bacchetti et al., 1993). Although these methods are less costly, they could not provide timely data on current transmission rates. Furthermore, changes in the AIDS case definition and the introduction of effective treatment that slow disease progression to AIDS rendered back-calculation methods invalid (Hall et al., 2009).

Other methods for estimating HIV incidence utilizes age-specific prevalence data in stable endemic conditions (Gregson et al., 1998). For example, one of the methods is to model HIV prevalence as the cumulative incidence of new infections at each proceeding age. Another method, models incidence as the difference between the observed prevalence levels at two successive age intervals. Although these are likelihood based methods, they do not use biomarker assays to identify new infections but rather they rely on the assumed survival distribution after HIV infection which could be a problem since survival is affected by the use of anti-retroviral therapy.

An alternative approach is to estimate HIV incidence from a cross sectional survey using biomarkers to identify persons acutely or recently infected (Brookmeyer and Quinn, 1995; Janssen et al., 1998). Both Brookmeyer and Quinn (1995) and Brookmeyer et al. (1995) proposed the detection of HIV-1 p24 antigen in the pre-seroconversion period for identifying persons with acute infection for estimating incidence. The idea is to determine the prevalence of p24 antigenemia among persons who have not seroconverted (Brookmeyer and Quinn, 1995). Incidence is then estimated using standard epidemiological relationships between incidence, prevalence and mean duration in which prevalence is a product of incidence rate and mean duration (Freeman and Hutchison, 1980). That is, prevalence is a product of incidence rate and mean duration. In this setting, prevalence refers to the proportion of recently infected persons among those at risk, and the mean duration refers to the mean duration of the "window" period (Brookmeyer and Quinn, 1995; Karon et al., 2008). Mean duration is often referred to as the mean window period and it is the average time it takes newly infected individuals to pass from "recent" infection to "non recent" infection according to the biomarkers (Lagakos and Gable, 2008). However, this approach requires large sample sizes in order to obtain reliable estimates of incidence because the duration of antigenemia before antibody detection is short, about 22.5

days (Brookmeyer and Quinn, 1995; Janssen et al., 1998). Subsequently, another assay that is associated with a longer period from being undetectable to detectable was developed (Janssen et al., 1998). That is, the indirect immunoassay (EIA) is modified or "detuned" or made "less sensitive" by increasing the specimen dilution (Janssen et al., 1998; Parekh et al., 2002). The detuned assay is then used to identify recent infection by detection of differential HIV-1 antibody titer. The assay have been used to estimate incidence in many US populations due to its availability in those settings (Janssen et al., 1998; Parekh et al., 2002). In general, identification of new infections is done in two stages. In the first stage, one takes a cross-sectional random sample of size n from an asymptomatic population, and tests each person using a sensitive test (typically ELISA). Individuals testing negative are assumed to be uninfected at that time. In the second stage, individuals testing positive on a more sensitive test are tested again using a less-sensitive test. Subjects that are found to be positive on a sensitive test but negative on the less sensitive test for HIV infection are considered to be recent infections. This method is known as the serologic testing algorithm for recent HIV seroconversion (STARHS), (Janssen et al., 1998). However, limited availability in other parts of the world and some other limitations of the assay (that include requirements of special equipments, subtype dependent performance and significant variability of the window periods) hampered its wide usage and subsequently led to the evolution of IgG-capture BED-EIA (commonly referred to as BED assay) for detecting recent HIV-1 infections (Parekh et al., 2002; Parekh and McDougal, 2005; WHO, 2009). The BED assay has been used worldwide in five continents, both in the general population and in high risk groups (Barnighausen et al., 2010).

Generally conditions such as HIV have a natural history that initially someone is born free from HIV and the stage is called the “disease free” stage. Then comes a stage where he/she is infected with HIV during which the virus cannot be detected by the standard antibody tests. This stage is often referred to as the “pre-seroconversion period”. Later the virus is then detected by a standard antibody tests and this stage is called the “sero-conversion period”. Finally there is a stage where the virus is “established” and the stage is called the “non-recent” stage.

Using this natural history of HIV, Balasubramanian and Lagakos (2010) developed a theoretical framework using the likelihood approach through a multinomial distribution and derived corresponding likelihood estimators of HIV incidence under different settings, including settings where more than two diagnostic tests are used and settings where subjects are tested at different ages. The method allows incorporation of covariates. The idea is very simple, when we select a random sample of subjects of size n from a population and test each person using a sensitive antibody test such as ELISA, n^{+} subject will test positive for HIV antibody, while n^{-} will test negative such that $n = n^{+} + n^{-}$. The n^{+} are tested again using a less sensitive test. n^{++} will test positive again while n^{+-} will test negative. The n^{+-} are regarded as the recent infection since they have not transferred to the non-recent infection state while the other n^{++} are regarded as the established infections (Janssen et al., 1998). n^{-} subjects are tested again to detect the HIV-1 p24 antigens in the pre-seroconversion stage. The n^{-+} individuals who test positive for HIV-1 p24 antigens are regarded as acute infections while n^{--} subjects testing negative for HIV-1 p24 antigens are assumed to be negative for HIV at the time of testing. Balasubramanian and Lagakos (2010) assumes that the distribution of (n^{--}, n^{-+}, n^{+-}) follows a multinomial distribution with probabilities (π_0, π_1, π_2) respectively. That is,

$$P(n^{--}, n^{-+}, n^{+-}, n^{++} | \pi_0, \pi_1, \pi_2, \pi_3) \sim \text{trinomial}(n; \pi_0, \pi_1, \pi_2).$$

where $\pi_3 = 1 - \pi_0 - \pi_1 - \pi_2$ and $n = n^{--} + n^{-+} + n^{+-} + n^{++}$.

1.2 Data Description

1.2.1 Introduction

The thesis utilizes data from the Botswana AIDS Impact (BAIS) III of 2008 to illustrate the proposed methods. The first AIDS impact survey in Botswana was conducted in 2001 and it was called the BAIS I. This was Botswana's first national population based household sexual behavioral survey. However, the study was limited to collecting the baseline information on some topics related to HIV/AIDS including behavioural changes. There was no HIV testing undertaken during BAIS I. In 2004, the second AIDS impact survey called BAIS II was conducted. It focused on identifying and measuring factors that are associated with HIV in Botswana. These factors include information on behaviour, knowledge, attitudes and cultural influences. More importantly, BAIS II also focused on the estimation of HIV prevalence amongst the population aged 18 months and above. BAIS III was conducted in 2008. The aim of BAIS III was to update existing information on the behavioral patterns of the populations aged 10-64 years and to estimate HIV prevalence and incidence rates among individuals aged 18 months and above. Estimation of HIV incidence is important for assessing the impact of prevention efforts and to track and monitor the growth of new infections. BAIS III was a cross sectional study. The unique feature of BAIS III is that it has biomarkers data on HIV incidence. It is this unique feature of BAIS III data that was exploited in the current work to show the application of the proposed methods.

1.2.2 Laboratory Testing

In stage I, a parallel testing algorithm using commercial ELISA test kits - Vironostika-HIV Uni-Form II plus O (Organon Teknika, Boxtel, The Netherlands) and Murex (Abbott, Wiesbaden, Germany) was used to screen blood samples. A specimen was considered HIV antibody positive if it was reactive on parallel ELISA testing otherwise it was considered HIV antibody negative. In stage II, in order to detect recent HIV-1 seroconversion, all the HIV positive specimens were tested again using a less sensitive test, Aware BED enzyme immunoassay (EIA) HIV-1 incidence Test (Calypte Biomedical Corporation, Portland, Oregon, USA). Specimens were considered recent HIV-1 seroconversion if they were not reactive to Aware BED EIA test (tested negative) for $ODn < 0.8$ otherwise the specimens were classified as long standing infections. More details of these Laboratory Testing procedures are provided in CSO (2008) report.

1.2.3 The Data

All persons aged 18 months and above were eligible for HIV testing using a sample of size n=21,414 during the survey period. But only 67% (sample size=14,351) provided blood specimen for HIV Testing. Of the 14,351 subjects who tested, 2521 tested HIV positive, $n_0 = 11,823$ tested HIV negative while 7 results were indeterminate. Out of the 2521 subjects who tested positive for HIV, there were $n_r = 149$ recent infections and $n_1 = 2319$ long standing infections. For the remaining subjects, $n_{others} = 53$ the specimens were reported missing (for 27 specimens the box was not found and for other 26 the blood samples were finished). Initially we assumed that the missing specimens are missing completely at random (MCAR) because the reasons for samples to be missing may be assumed to unrelated to outcome of interest whether observed or unobserved (Little and Rubin, 2002). However, to be on the safer side and the fact the MCAR assumption is too strong an assumption we later relaxed this assumption by considering and applying the missing at random assumption (MAR). This is a better assumption than the MCAR because it allows the missingness to depend on at least the observed data and can easily be dealt with using likelihood approaches.

1.2.4 Preliminary analysis

Data was collected on a number of variables including: age, sex, marital status, educational level, status on alcohol intake (whether or not one have ever taken an alcoholic drink), drug intake (whether or not one have ever taken drugs), and the number of partners during the survey period. Table 1.1 shows some preliminary data analysis excluding 53 subjects with specimen reported. So this analysis is based on $n = n_0 + n_r + n_1 = 11,823 + 149 + 2319 = 14,291$. We performed a chi-square test for association between outcome status and each of the above variables. The p-values are presented in the Table 1.1. Also Table 1.1 shows the percentage of missing data. We note that where we have missing data, the p-values may not be valid.

Table 1.1: Distribution of the sampled individuals into the 3-states disease model stratified according to some key demographic characteristics

Demographics		Sample n (n_0, n_r, n_1)	p-value	% of data missing
Sex	Male	6516 (5603, 55, 858)	< 0.01	0%
	Female	7775 (6220, 94, 1461)		
Age (y)	≤ 4	872 (854, 5, 13)	< 0.01	0%
	5-14	3382 (3242, 18, 122)		
	15-19	1449 (1395, 5, 49)		
	20-29	2967 (2401, 43, 523)		
	30-39	2142 (1287, 33, 822)		
	40-49	1429 (928, 27, 474)		
	50+	2050 (1716, 18, 316)		
Education	Non-formal	203 (146, 3, 54)	< 0.01	35%
	Primary	3247 (2586, 34, 627)		
	Secondary	4625 (3620, 56, 949)		
	Higher	1200 (1006, 12, 182)		
Marital status	Never married	6195 (5206, 74, 915)	< 0.01	27%
	Married	1465 (1162, 20, 283)		
	Living together	2307 (1523, 29, 755)		
	Separated	70 (45, 0, 25)		
	Divorced	100 (74, 0, 26)		
	Widowed	252 (152, 4, 96)		
Ever taken alcohol	Yes	3814 (2814, 42, 958)	< 0.01	27%
	No	6582 (5354, 85, 1143)		
Ever taken drugs	Yes	305 (231, 4, 70)	0.47	27%
	No	10086 (7934, 123, 2029)		
No. of sexual partners	0	1108 (776, 24, 308)	0.04	47%
	1	5573 (3985, 80, 1508)		
	2	642 (486, 10, 146)		
	> 2	182 (141, 1, 40)		

1.2.5 Limitations of the Data

One of the major limitations of this data set was that, except for age and sex, all other variables had a lot of missing information. All other variables have more than 25% of the data missing. For this reason, we only used age and sex to illustrate the proposed methods and for addressing the question of incidence dependence on covariates.

1.3 Thesis Objective

The objective of this thesis is to develop likelihood based methods that can be used to estimate HIV incidence when data is derived from cross sectional surveys and the disease classification is achieved by combining two biomarker or assay tests. The thesis builds on the work of Janssen et al. (1998) and extends the statistical framework developed by Balasubramanian and Lagakos (2010) to derive the maximum likelihood estimators of HIV incidence under different settings.

1.4 Thesis Outline

Chapter 2 reviews the four state model of Balasubramanian and Lagakos (2010) and introduces the reader to the three state disease progression model. We also propose a method of estimating incidence when we have information on the immediate past prevalence is present. The method is compared to that of Wang and Lagakos (2009). The method performs slightly better than the one of Wang and Lagakos (2009) in terms of efficiency gain. The chapter concludes by a discussion that also includes some limitations of this method.

In Chapter 3, we describe methods for simultaneously incorporating the sensitivity and specificity of the diagnostic tests for new infections. In particular, we derive the maximum likelihood estimator for incidence when the less sensitive test has sensitivity and specificity less than 100% assuming the sensitive test has 100% sensitivity and specificity.

In Chapter 4, we relax the assumption of constant incidence density and introduce the linear incidence density. This approach is applicable to situations where the incidence density is changing over time. However, the idea of linear incidence density is limited by the issues of identifiability hence we make an additional assumption that the past prevalence is known.

In Chapter 5, we present the method for adjusting the estimates of incidence when a proportion of subjects tested using an antibody sensitive test were not tested using a less sensitive test resulting in missing data. In particular, we considered how adjustments can be made in the log-likelihood when the missing data is assumed to be missing at random.

Chapter 6 describes the method for incorporating the important risk factors and extend the method of Magder and Hughes (1997) to incorporate the uncertainty of determining the outcome of interest which is HIV incidence.

In Chapter 7, we present the overall conclusion of our research which includes the limitations as well suggestions for future research.

Chapter 2

A likelihood estimation of HIV incidence incorporating information on past prevalence

SUMMARY. The prevalence and incidence of an epidemic are basic characteristics that are essential for study planning, assessing the effect of interventions and for determining public health priorities. A direct approach for estimating incidence is to undertake a longitudinal cohort study where a representative sample of disease free individuals are followed for a specified period of time and new cases of infection are observed and recorded. This approach is expensive, time consuming and prone to bias due to loss-to-follow-up. An alternative approach is to estimate incidence from cross sectional surveys using biomarkers to identify persons recently infected as in (Brookmeyer and Quinn, 1995; Janssen et al., 1998). This paper builds on the work of Janssen et al. (1998) and extends the theoretical framework proposed by Balasubramanian and Lagakos (2010) by incorporating information on past prevalence and deriving maximum likelihood estimators of incidence. The performance of the proposed method is evaluated through a simulation study, and its use is illustrated using data from the Botswana AIDS Impact (BAIS) III survey of 2008.

KEY WORDS: HIV incidence; Cohort studies; Cross sectional surveys; Maximum likelihood; previous prevalence

2.1 Introduction

In this chapter, we consider estimating HIV incidence when we have information on previous prevalence. In many settings, at the time of the cross sectional survey, the information on previous prevalences (existing/non-recent infections) from previous studies is often available. If the period between the immediate past survey and the current cross sectional survey is not too long, then one can use the prevalence from the past survey to improve incidence estimation in the current survey. Incorporating additional information of known past prevalence can lead to an efficiency gain in estimating incidence. This paper builds on the work of Janssen et al. (1998) and extends the theoretical framework proposed by Balasubramanian and Lagakos (2010) by incorporating information on past prevalence in order to improve the efficiency of incidence estimation.

It is known (McDougal et al., 2006; Barin et al., 2005; Chawla et al., 2007) that a varying proportion of individuals tested with BED assay produce the "false-recent" results long after seroconversion (assay non-progressors). That is, some long standing infection are misclassified as recent long after they have seroconverted (Karita et al., 2007; McDougal et al., 2006; Hargrove et al., 2008). Estimators that take into account false recent rate have been proposed (McWalter and Welte, 2010; Wang and Lagakos, 2009). Since our method is likelihood based, we extend the approach in Wang and Lagakos (2009) to make adjustments for false recent rate in the proposed model.

The rest of the chapter is structured as follows. In Section 2.2, we review models for HIV progression and the likelihood estimators for incidence. The proposed model that incorporates the information on previous prevalence is introduced in Section 2.3. A simulation study to investigate the performance of this model is presented in Section 2.4. We use the data from Botswana AIDS Impact (BAIS) III survey of 2008 in Section 2.5 to illustrate the use of the proposed method. Lastly, we end the chapter with a discussion in Section 2.6.

2.2 Models and the likelihood function

2.2.1 Review of the Four-State Balasubramanian-Lagakos (BL) Model

Balasubramanian and Lagakos (2010) generalized the HIV history of an individual through a four state model with the four states denoted by S_0, S_1, S_2 and S_3 . Where S_0 represents the uninfected state which extend from the time an individual is born to the time he/she develops the HIV antigens which are often detected by an antigen test (denoted by A). S_1 represents the "acute infection state" which extend from the initial infection to seroconversion. By seroconversion we refer to the development of HIV antibodies that are often detectable by a sensitive test such as standard ELISA denoted by E. S_1 is the state that Brookmeyer and Quinn (1995) used to determine the recent infections. In this setting, which is similar to Janssen et al. (1998) approach,

Table 2.1: Summary of the 4-States

State	Test Results	Implication
S_0	(A^-, E^-)	uninfected
S_1	(A^+, E^-)	acute infection
S_2	(E^+, D^-)	recent infection
S_3	(E^+, D^+)	nonrecent infection

S_2 represents the "recent infection" state where HIV antibodies are detectable by a sensitive diagnostic test but not yet through a less sensitive test such as BED assay or detuned ELISA denoted by D. Finally S_3 represents the "non-recent infection" state in which HIV antibodies are detectable by a less sensitive test (and sensitive test). Table 2.1 summarizes the four states in relation to the test result of a sample and what the implication is.

The above BL 4-state model can be considered as a hierarchical or longitudinal progression process where subjects are initially in state S_0 then they move sequentially to S_1 then S_2 and finally to S_3. Note that cross sectional sampling of subjects is done using diagnostic tests and this provides useful information about the state prevalence functions denoted by:

$$\pi_j(t) = P(\text{subject is in state } S_j \text{ at calender time t})$$

for j $= 1, 2, 3, 4$. The idea is to link the cross-sectional samples to longitudinal quantities so as to learn more about the underlying disease process.

2.2.2 The 3-state model, state probabilities and incidence

In some settings, S_0 and S_1 have been combined to form a single state such that the HIV history of an individual is conceptualized in three states. This is basically how the Janssen et al. (1998) estimator was formulated. In this paper, we follow the 3-state model approach and throughout this paper S_1 shall refer to combined S_0 and S_1 as they are in Section 2.2.1. That is, we consider a 3-state disease progression model for the natural history of HIV/AIDS noted by S_1, S_2 and S_3. Here S_1 represents the "pre-seroconversion" state which corresponds to the period in which an individual is either infected but have not yet seroconverted or is uninfected. S_2 represents the "recent infection" state where HIV antibodies are detectable by a sensitive diagnostic test but not yet through a less sensitive test. Finally S_3 represents the "non-recent infection" state in which HIV antibodies are detectable by a less sensitive test and sensitive test. We later consider the 4-state model of Wang and Lagakos (2009) to make adjustments for assay non-progressors.

We use similar notation as in (Wang and Lagakos, 2009; Claggett et al., 2012). Assume the random variable T denotes the calendar time of HIV seroconversion for someone born at time 0. Let t denote the calendar time of the cross sectional sample. Furthermore, let $f(t)$, $F(t)$ and $\lambda(t)$ denote the incidence density function (incidence density) for becoming infected, cumulative distribution function and the incidence rate/hazard rate of T at time $u \geq 0$ respectively. Let L_2 denote the sojourn or residence time in S_2 with the corresponding cumulative distribution denoted by $G_2(\cdot)$. We assume that L_2 has support in $[0, L_2^*]$, where $L_2^* < t$ and is independent of T. It follows that $G_2(0) = 0$ and $G_2(L_2^*) = 1$.

The prevalence probabilities for the 3-states at time t are given by:

$$\begin{aligned}
\pi_1(t) &\overset{def}{=} P(\text{in } S_1 \text{ at calender time t}) \\
&= 1 - F(t) \qquad (2.2.1) \\
\pi_2(t) &\overset{def}{=} P(\text{in } S_2 \text{ at calender time t}) \\
&= \int_{t-L_2^*}^{t} f(u)\,[1 - G_2(t-u)]\,\mathrm{d}u \qquad (2.2.2) \\
\pi_3(t) &= 1 - \pi_1(t) - \pi_2(t) \qquad (2.2.3)
\end{aligned}$$

Our interest is on the estimation of the HIV incidence rate at time t given by

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \qquad (2.2.4)$$

Since t is the fixed calender time at which the cross sectional sample is obtained, we shall hereafter denote $\lambda(t)$ by λ.

The Likelihood Function and HIV incidence estimation without previous prevalence

Suppose we select a random sample of size n from the population at some calendar time t. Let n_0 denote the number of subjects who test negative on a more sensitive test, n_r denote the number of subjects who test positive on a sensitive test and negative on a less sensitive test and n_1 denote the number of subjects who test positive on a less sensitive test such that $n = n_0 + n_r + n_1$. The n_r individuals are also referred to as "recent infections" while the n_1 individuals are referred to as "long standing infections" or "non-recent infections". Note that if one tests negative on a sensitive test then a less sensitive test is generally not done but assumed to be negative. Note that if one tests negative on a sensitive test then a less sensitive test is generally not given but assumed to be negative. In practice, a less sensitive test is administered among individuals who test positive on the sensitive test. If the results of the less sensitive test is negative then the individual is classified as a recent infection. Otherwise if the individual test positive on both tests then he/she is classified as a long standing infection. The commonly used less sensitive test for different HIV subtypes (B, E, D, C) is BED capture enzyme immunoassay, often referred to as BED assay, developed by the Centers for Disease Control and Prevention (CDC). For this reason we shall denote the less sensitive test by B instead of D as in the 4-state BL model.

The general likelihood function, L, corresponding to the 3-state model is given by

$$\mathrm{L} = [\pi_1(t)]^{n_0}[\pi_2(t)]^{n_r}[\pi_3(t)]^{n_1} \tag{2.2.5}$$

subject to $\pi_1(t)+\pi_2(t)+\pi_3(t) = 1$. Under the constant incidence function assumption, the likelihood function for the 3-states model can be written as:

$$\mathrm{L}(f,\theta) = [1-\theta_t]^{n_0}[f\mu]^{n_r}[\theta_t - f\mu]^{n_1} \tag{2.2.6}$$

assuming $\mu = E(L_2)$ is known.

The corresponding log likelihood function, $\ell(f,\theta)$ will be:

$$\ell(f,\theta) = n_0\log(1-\theta) + n_r\log(f\mu) + n_1\log(\theta - f\mu) \tag{2.2.7}$$

The MLEs for θ and f can be obtained by joint maximization of Eq (2.2.7) with respect to f and θ and the closed form solutions can be found. The partial derivatives with respect to θ and f, respectively, are:

$$\frac{\partial\ell(f,\theta)}{\partial\theta} = \frac{-n_0}{1-\theta} + \frac{n_1}{\theta - f\mu} \tag{2.2.8}$$

$$\frac{\partial \ell(f,\theta)}{\partial f} = \frac{n_r}{f} - \frac{n_1\mu}{\theta - f\mu} \tag{2.2.9}$$

Equating Eq (2.2.8) and Eq (2.2.9) to zero and solving for $\hat{\theta}$ and $\hat{f}$ we obtain:

$$\begin{aligned} \hat{\theta} &= \frac{n_r + n_1}{n} \\ \hat{f} &= \frac{n_r}{n\mu} \end{aligned}$$

where $n = n_0 + n_r + n_1$.

Since

$$\hat{\lambda} = \frac{\hat{f}}{1-\hat{\theta}}$$

then

$$\hat{\lambda} = \frac{n_r}{n_0\mu} \tag{2.2.10}$$

Eq (2.2.10) is the estimator of incidence rate proposed by Kaplan and Brookmeyer (1999) and it is arises as a special case of the 4-state model proposed by Balasubramanian and Lagakos (2010) when the standard ELISA and detuned ELISA were used as the sensitive and less sensitive tests respectively assuming the number of subjects who tested positive on HIV antigen test is negligible. Basically, the Balasubramanian and Lagakos (2010) model under this assumption is what we refer to as the 3-state model approach.

The other estimator of HIV incidence proposed by Janssen et al. (1998) uses the standard epidemiological relationship between prevalence and incidence rate and is given by:

$$\hat{\lambda} = \frac{n_r}{(n_0 + n_r)\mu} \tag{2.2.11}$$

As noted by Balasubramanian and Lagakos (2010), the two estimators are closer to each other since the number of new infections defined as the number of subjects testing positive on the sensitive test and negative on the less sensitive test is usually much smaller than the number testing negative on both tests and hence if we neglect n_r in the denominator of Eq (2.2.11) then the two estimators are the same.

Variance estimation and confidence intervals

Let $\xi = (f, \theta)$ and $\hat{\xi} = (\hat{f}, \hat{\theta})$ denote the MLEs for ξ respectively. To construct the confidence intervals for ξ or functions of its components, the standard errors and the covariances can be obtained from the Hessian matrix of the log-likelihood function (Cox and Hinkley, 1974). The variance of $\hat{\xi}$ is given by,

$$\text{Var}(\hat{\xi}) = -\text{E}[H]^{-1} = -\text{E}\begin{bmatrix} h_{\hat{f},\hat{f}} & h_{\hat{f},\hat{\theta}} \\ h_{\hat{\theta},\hat{f}} & h_{\hat{\theta},\hat{\theta}} \end{bmatrix}^{-1}$$

where

$$\begin{aligned} h_{\hat{f},\hat{f}} &= \frac{\partial^2 \ell(f,\theta)}{\partial f^2} = -\frac{n_r}{f^2} - \frac{n_1 \mu^2}{(\theta - f\mu)^2} \\ h_{\hat{f},\hat{\theta}} &= \frac{\partial^2 \ell(f,\theta)}{\partial \theta \partial f} = \frac{n_1 \mu}{(\theta - f\mu)^2} \\ h_{\hat{\theta},\hat{\theta}} &= \frac{\partial^2 \ell(f,\theta)}{\partial \theta^2} = -\frac{n_0}{(1-\theta)^2} - \frac{n_1}{(\theta - f\mu)^2} \end{aligned}$$

When f and θ are replaced by their MLEs, the estimated variance of $\hat{\lambda}$,

$$\text{var}(\hat{\lambda}) = \frac{n_r(n_0 + n_r)}{n_0^3 \mu^2}$$

Wang and Lagakos (2009) approximate estimate of the variance of $\hat{\lambda}$,

$$\mathrm{var}(\hat{\lambda}) \approx \frac{n_r}{n_0^2\mu^2} \tag{2.2.12}$$

This is because $n_r \ll n_0$ hence $n_0 + n_r \approx n_0$.

2.3 The extended model allowing for past prevalence

Suppose HIV prevalence at time $t_1 < t$, $F(t_1)$, is known. Thereafter we denote $F(t_1)$ by θ_0. Let $\tau = t - t_1$ denote the time between the current study at time t and the immediate past study at time t_1. Let the prevalence at time t be $F(t) = \theta$. Assuming constant density for HIV seroconversion in a short period of time right before the cross sectional survey, that is,

$$f(u) = f \text{ for } u \in [t - L_2^*, t]$$

then θ is given by

$$\begin{aligned} \theta &= \theta_0 + \int_{t_1}^{t} f(u)du \\ &= \theta_0 + f(t - t_1) \text{ (for } t_1 < L_2^*) \\ &= \theta_0 + f\tau \end{aligned} \tag{2.3.1}$$

Since

$$\lambda = \frac{f}{1-\theta} = \frac{f}{1-\theta_0 - f\tau}$$

then

$$f = \frac{\lambda(1-\theta_0)}{1+\lambda\tau} \tag{2.3.2}$$

The likelihood function, in terms of λ, is then given by:

$$\mathrm{L}(\lambda) = \left[1-\theta_0-\frac{\lambda(1-\theta_0)\tau}{1+\lambda\tau}\right]^{n_0}\left[\frac{\lambda(1-\theta_0)\mu}{1+\lambda\tau}\right]^{n_r}\left[\theta_0+\frac{\lambda(1-\theta_0)(\tau-\mu)}{1+\lambda\tau}\right]^{n_1}$$

Hence the maximum likelihood estimator of λ, $\hat{\lambda}$, when previous prevalence is known is given by

$$\hat{\lambda} = \frac{-b+\sqrt{b^2-4ac}}{2a} \tag{2.3.3}$$

where

$$\begin{aligned} a &= n_0\tau\left[\tau-(1-\theta_0)\mu\right] \\ b &= (n_0+n_1)\tau\theta_0-(n_r+n_1)\left[\tau-(1-\theta_0)\mu\right] \\ c &= -n_r\theta_0 \end{aligned}$$

for $b^2-4ac>0$. Note that to ensure that the estimated incidence rate is a positive quantity we use the positive root of the quadratic equation arising from the maximization of the log-likelihood equation with respect to with λ. See the Appendix for a derivation of the quadratic equation.

We note that the proposed estimator, $\hat{\lambda}$, reduces to the simple estimator when we have no information on previous prevalence when τ equal μ, the mean window period. That is, for $\tau=\mu$

$$\hat{\lambda} = \frac{n_r}{n_0\mu}.$$

This is the estimator of HIV incidence, given in Eq (2.2.10), when we have no previous prevalence.

To construct confidence intervals, we need the variance of $\hat{\lambda}$. Given $\frac{\partial^2 \ell}{\partial \lambda^2}$, the approximate variance of $\hat{\lambda}$,

$$\mathrm{var}(\hat{\lambda}) = -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \lambda^2}\right]^{-1}$$

where

$$\frac{\partial^2 \ell}{\partial \lambda^2} = \frac{n\tau^2}{(1+\lambda\tau)^2} - \frac{n_r}{\lambda^2} - \frac{n_1\left[\tau - (1-\theta_0)\mu\right]^2}{\left[\theta_0 + \lambda\left\{\tau - (1-\theta_0)\mu\right\}\right]^2} \tag{2.3.4}$$

We obtain the estimate of the variance of $\hat{\lambda}$ by replacing λ in Eq (2.3.4) with $\hat{\lambda}$, its MLE given in Eq (2.3.3).

Also we note that for $\tau = \mu$

$$\mathrm{var}(\hat{\lambda}) = \frac{n_r(n_0 + n_r)}{n_0^3 \mu^2} \approx \frac{n_r}{n_0^2 \mu^2}$$

as in Wang and Lagakos (2009).

2.3.1 Incorporating false recent rate into the extended model

The model we have introduced in Section 2.3 does not take into account individuals who remain negative on the less-sensitive assay indefinitely. It is recognized that a fraction of HIV infected subjects, referred to as assay non-reactors/non-progressors or elite suppressors, repeatedly test negative on a less sensitive test, such as BED assay, long after they have seroconverted leading in what has been termed “false recent rate” (Wang and Lagakos, 2009; McWalter and Welte, 2010; Brookmeyer, 2010b; Hargrove et al., 2008; Novitsky et al., 2009; Karita et al., 2007). Failure to account for this false recent rate results in estimates of incidence that are biased because they overestimate the number of subjects who are recently infected (Wang and Lagakos, 2009; Hargrove et al., 2008; Karita et al., 2007). Estimators to address this problem have been proposed (McWalter and Welte, 2010; Wang and Lagakos, 2009). Since the estimator we have introduced in Section 2.3 is also a maximum likelihood estimator, to incorporate the false recent rate we extend the approach of Wang and Lagakos (2009).

We consider the four-state progressive-disease model as in Wang and Lagakos (2009). Under this model, State 1 represents the "pre-seroconversion" state which corresponds to the period in which an individual is either infected but have not yet seroconverted or is uninfected. State 2 represents the "recent infection" state where HIV antibodies are detectable by a sensitive diagnostic test but not yet through a less sensitive test (though the test will eventually become reactive to the less-sensitive assay). State 3 represents the "non-recent infection" state in which HIV antibodies are detectable by a less sensitive test and sensitive test. Finally state 4 represents subjects who are detectable by the sensitive test, but who will permanently remain nonreactive to the less-sensitive test.

2.3.2 The likelihood function and parameter estimation

Let p denote the proportion of subjects who will become reactive at some point after seroconversion. Hence the false recent rate, $1 - p$, denotes the proportion of subjects who remain negative on the less sensitive assay indefinitely.

To incorporate the false recent rate, the general likelihood function, L, that we introduced in Eq (2.2.5) can be extended as follows:

$$\mathrm{L} = [\pi_1(t)]^{n_0}[\pi_2(t)p + (1-\pi_1(t))(1-p)]^{n_r}[\pi_3(t)p]^{n_1} \tag{2.3.5}$$

subject to $\pi_1(t) + [\pi_2(t)p + (1-\pi_1(t))(1-p)] + \pi_3(t)p = 1$.

Specifically, likelihood we presented in Section 2.3 can be extended to incorporate the false recent rate as follows:

$$\begin{aligned}
\ell(\lambda) &= \left[1-\theta_0-\frac{\lambda(1-\theta_0)\tau}{1+\lambda\tau}\right]^{n_0} \\
&\times\left[\frac{\lambda(1-\theta_0)\mu p}{1+\lambda\tau}+\left\{\theta_0+\frac{\lambda(1-\theta_0)\tau}{(1+\lambda\tau)}\right\}(1-p)\right]^{n_r} \\
&\times\left[p\left\{\theta_0+\frac{\lambda(1-\theta_0)(\tau-\mu)}{1+\lambda\tau}\right\}\right]^{n_1}
\end{aligned}$$

The corresponding log likelihood is

$$\begin{aligned}
\ell(\lambda) &= n_0\ln\left[1-\theta_0-\frac{\lambda(1-\theta_0)\tau}{1+\lambda\tau}\right] \\
&+n_r\ln\left[\frac{\lambda(1-\theta_0)\mu p}{1+\lambda\tau}+\left\{\theta_0+\frac{\lambda(1-\theta_0)\tau}{(1+\lambda\tau)}\right\}(1-p)\right] \\
&+n_1\ln\left[p\left\{\theta_0+\frac{\lambda(1-\theta_0)(\tau-\mu)}{1+\lambda\tau}\right\}\right] \\
&\propto -n\ln(1+\lambda\tau)+n_r\ln\left[\lambda\mu p-\lambda\mu p\theta_0+\theta_0-\theta_0 p+\lambda\tau-\lambda\tau p\right] \\
&+n_1\ln\left[-\lambda\mu+\lambda\mu\theta_0+\theta_0+\lambda\tau\right]
\end{aligned}$$

The partial derivative with respect to λ is

$$\frac{\partial\ell}{\partial\lambda} = -\frac{n\tau}{1+\lambda\tau}+\frac{n_r m_r}{u_r+\lambda m_r}+\frac{n_1 m_1}{u_1+\lambda m_1} \tag{2.3.6}$$

where

$$\begin{aligned} n &= n_0 + n_r + n_1 \\ m_r &= \mu p(1-\theta_0) + \tau(1-p) \\ u_r &= \theta_0(1-p) \\ m_1 &= -\mu(1-\theta_0) + \tau \\ u_1 &= \theta_0 \end{aligned}$$

The MLE of $\hat{\lambda}$ is obtained by setting Eq (2.3.6) to 0 and is given by:

$$\hat{\lambda} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{2.3.7}$$

where

$$\begin{aligned} a &= m_1 m_r \tau n_0 \\ b &= m_r \tau u_1(n_0 + n_1) + m_1 \tau u_r(n_0 + n_r) - m_1 m_r(n_r + n_1) \\ c &= n\tau u_r u_1 - n_r m_r u_1 - n_1 m_1 u_r \end{aligned}$$

We consider the positive root given by Eq (2.3.7) to ensure that the values of incidence rate are within acceptable range. Note that for this setting, values of incidence rate can be negative when $1-p$ is unusually large say greater than 0.05.

The approximate variance of $\hat{\lambda}$, obtained by replacing λ by its maximum likelihood estimates given in Eq (2.3.7), is

$$\mathrm{var}[\hat{\lambda}] = -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \lambda^2}\right]^{-1}$$

where

$$\frac{\partial^2 \ell}{\partial \lambda^2} = \frac{n\tau^2}{(1+\lambda\tau)^2} - \frac{n_r m_r^2}{(u_r + \lambda m_r)^2} - \frac{n_1 m_1^2}{(u_1 + \lambda m_1)^2} \tag{2.3.8}$$

and m_r, u_r, m_1 and u_1 are as defined before.

2.4 Simulation Study

We performed a simulation study to evaluate the performance of the proposed estimator of HIV incidence using 1000 simulations. We generated a sample of size $n = 3000$. The data was simulated from a multinomial distribution with state probabilities for the unadjusted model given by:

$$\begin{aligned} \pi_1(t) &= 1 - \theta_0 - \frac{\lambda(1-\theta_0)\tau}{1+\lambda\tau} \\ \pi_2(t) &= \frac{\lambda(1-\theta_0)\mu}{1+\lambda\tau} \\ \pi_3(t) &= 1 - \pi_1(t) - \pi_2(t) \end{aligned}$$

For the adjusted model that takes into account the false recent rate the probabilities are:

$$\begin{aligned} \pi_1(t) &= 1 - \theta_0 - \frac{\lambda(1-\theta_0)\tau}{1+\lambda\tau} \\ \pi_2(t) &= \frac{\lambda(1-\theta_0)\mu p}{1+\lambda\tau} + \left\{\theta_0 + \frac{\lambda(1-\theta_0)\tau}{(1+\lambda\tau)}\right\}(1-p) \\ \pi_3(t) &= 1 - \pi_1(t) - \pi_2(t) \end{aligned}$$

Wang and Lagakos (2009) conducted an extensive simulation to evaluate the method that adjusts for false recent rate described in Section 2.3.1 in the absence of previous prevalence. We consider the proposed (unadjusted) model when we have previous prevalence (Model 2 denoted as M2) and compare it with the standard (unadjusted) model with no data on previous prevalence (Model 1 denoted as M1). We also consider the proposed (adjusted) model when we have previous prevalence (Model 4 denoted by M4) and compare it with the standard (adjusted) model (Model 3 denoted by M3) in order to determine in which settings does previous prevalence improve the precision of incidence estimation.

For each simulation, we calculate the average point estimate of incidence based on the proposed model. The true value of the incidence rate, $\lambda(t)$ denoted by (I_{true}) is taken to be 3%. The average variance (standard errors) of these estimates are obtained from the observed Fisher information. We also obtain the empirical estimates of the variance from the 1000 simulated point estimate of incidence. The results of the simulation study are presented in Table 2.2 for $\theta_0 = 0.10, 0.171, 0.2$. For each fixed values of θ_0 we assess how efficiency gain is affected for $\tau = 0.1, 0.3, 0.5, 0.6, 1.0$ years. Also presented are the values of the coverage probabilities. That is, the proportion of times out of 1000 simulations the 95% confidence interval covers the true value of HIV incidence. The results are presented for both unadjusted and adjusted models (with and without previous prevalence).

We note that in all cases the simulated values of λ are closer to the true value of the incidence rate, $I_{true} = 0.03$. In general, $\hat{\text{SE}}(\hat{\lambda})$ and $\text{E}\hat{\text{SE}}(\hat{\lambda})$, the estimated standard errors and the corresponding empirical estimates, respectively, are very close to each other for all the different time points for both models. In most cases, the coverage probabilities are close to the nominal 95% confidence interval. As depicted in Figure 2.1, suitable values of τ are those between the mean window period (where the two curves overlap) and one year. In particular, for the unadjusted model, there is

Table 2.2: *Results of a simulation study comparing M2 and M1 with* $N = 3000$ *and* $I_{true} = 0.03$. $\hat{\lambda}$, $\hat{SE}(\hat{\lambda})$ *and* $E\hat{S}E(\hat{\lambda})$ *represents the average estimates, estimated standard errors and the corresponding empirical estimates respectively from 1000 simulations.* $95\%CV$ *denotes the proportion of experiments in which* I_{true} *is contained in the nominal* 95% *confidence interval.*

Simulation Results with $I_{true} = 0.03$							
		Previous HIV Prevalence					
		$\theta_0 = 0.10$		$\theta_0 = 0.171$		$\theta_0 = 0.20$	
τ(in years)	Estimates	M2	M1	M2	M1	M2	M1
0.1	$\hat{\lambda}$	0.0298	0.0299	0.0298	0.0298	0.0300	0.0300
	$\hat{\text{SE}}(\hat{\lambda})$	0.0049	0.0051	0.0052	0.0053	0.0054	0.0054
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0051	0.0054	0.0053	0.0054	0.0053	0.0055
	95%CV	93.1	93.3	94.6	93.8	95.4	95.2
0.3	$\hat{\lambda}$	0.0299	0.0299	0.0298	0.0298	0.0302	0.0302
	$\hat{\text{SE}}(\hat{\lambda})$	0.0051	0.0051	0.0054	0.0053	0.0055	0.0055
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0053	0.0054	0.0055	0.0055	0.0057	0.0057
	95%CV	92.8	92.7	93.7	93.8	94.0	94.1
0.5	$\hat{\lambda}$	0.0300	0.0300	0.0297	0.0297	0.0300	0.0300
	$\hat{\text{SE}}(\hat{\lambda})$	0.0052	0.0052	0.0054	0.0053	0.0055	0.0055
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0053	0.0053	0.0056	0.0056	0.0055	0.0055
	95%CV	93.5	92.8	93.5	93.3	94.7	94.3
0.6	$\hat{\lambda}$	0.0296	0.0299	0.0297	0.0297	0.0300	0.0300
	$\hat{\text{SE}}(\hat{\lambda})$	0.0051	0.0051	0.0053	0.0053	0.0055	0.0055
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0054	0.0054	0.0057	0.0058	0.0056	0.0057
	95%CV	92.7	92.4	92.7	92.9	93.9	93.8
1	$\hat{\lambda}$	0.0300	0.0300	0.0299	0.0300	0.0302	0.0302
	$\hat{\text{SE}}(\hat{\lambda})$	0.0048	0.0052	0.0051	0.0054	0.0053	0.0055
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0050	0.0056	0.0052	0.0056	0.0054	0.0057
	95%CV	93.7	94.0	94.9	94.2	94.5	94.5

an efficiency gain for values of τ between the mean window period and one year. In these settings we examined incorporating past prevalence leads to modest efficiency gain when τ is close to one year.

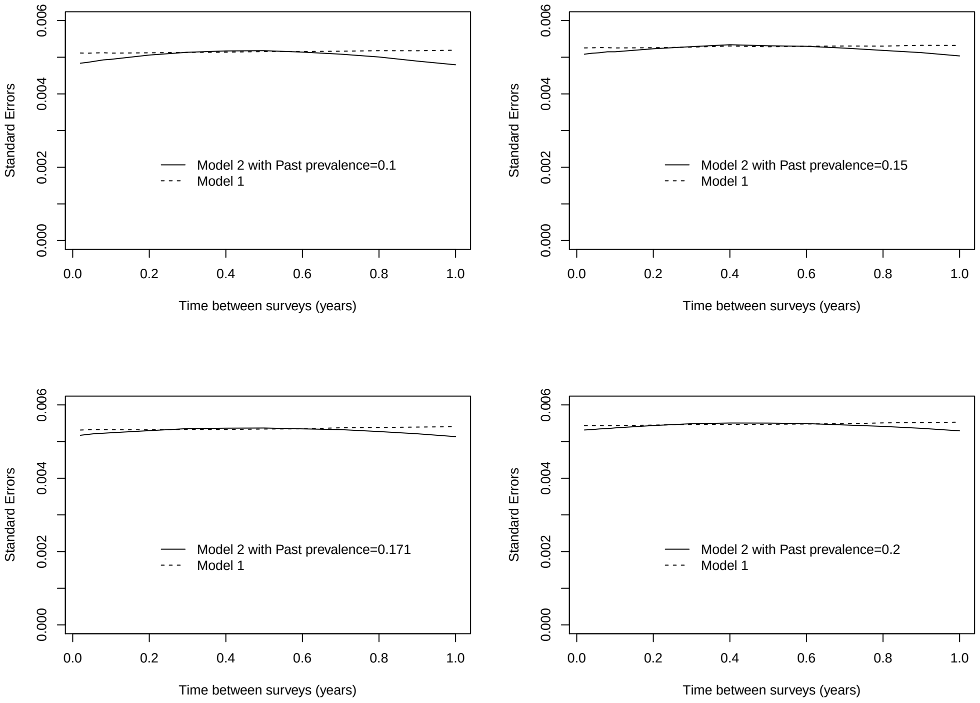

Figure 2.1: Graph of standard errors against time (years) between surveys using the data in Table 2.2

We also compared the adjusted models (with previous prevalence, $M4$ and without previous prevalence, $M3$). When we incorporate the false recent rate, we see that the estimates of λ are close to the true values in all cases and that the standard error increases for a higher proportion of false recent rate increases (see Table 2.3). Wang and Lagakos (2009) observed similar results. The estimated standard errors are closer to the corresponding empirical estimates in both cases of false recent proportions namely 5% and 2%. The relative efficiency in the proposed adjusted model compared to the standard adjusted model of Wang and Lagakos (2009) is around one.

Table 2.3: *Results of a simulation study with the incorporation of the False Recent Rate (FRR) denoted by p for* $N = 3000$ *and* $I_{true} = 0.03$. $\hat{\lambda}$, $\hat{SE}(\hat{\lambda})$ *and* $E\hat{SE}(\hat{\lambda})$ *represents the average estimates, estimated standard errors and the corresponding empirical estimates respectively from 1000 simulations.* $95\%CV$ *denotes the proportion of experiments in which* I_{true} *is contained in the nominal* 95% *confidence interval.*

Simulation Results with $I_{true} = 0.03$ incorporating FRR									
		$\theta_0 = 0.10$				$\theta_0 = 0.20$			
		$p = 0.95$		$p = 0.98$		$p = 0.95$		$p = 0.98$	
τ(y)	Est	$M4$	$M3$	$M4$	$M3$	$M4$	$M3$	$M4$	$M3$
0.3	$\hat{\lambda}$	0.0306	0.0306	0.0304	0.0304	0.0304	0.0303	0.0304	0.0304
	S$\hat{\mathrm{E}}(\hat{\lambda})$	0.0063	0.0063	0.0056	0.0057	0.0078	0.0079	0.0065	0.0065
	E$\hat{\mathrm{S}}$E$(\hat{\lambda})$	0.0064	0.0064	0.0058	0.0058	0.0080	0.0080	0.0068	0.0068
	95%CV	94.5	94.8	94.5	94.2	94.8	94.2	93.1	93.2
0.5	$\hat{\lambda}$	0.0304	0.0304	0.0304	0.0304	0.0304	0.0303	0.0303	0.0302
	S$\hat{\mathrm{E}}(\hat{\lambda})$	0.0061	0.0064	0.0056	0.0057	0.0076	0.0079	0.0065	0.0066
	E$\hat{\mathrm{S}}$E$(\hat{\lambda})$	0.0064	0.0066	0.0059	0.0060	0.0080	0.0082	0.0068	0.0069
	95%CV	94.3	93.9	93.6	93.9	93.8	94.1	93.7	93.2
0.6	$\hat{\lambda}$	0.0305	0.0305	0.0303	0.0303	0.0305	0.0305	0.303	0.0304
	S$\hat{\mathrm{E}}(\hat{\lambda})$	0.0061	0.0064	0.0056	0.0057	0.0076	0.0080	0.0064	0.0066
	E$\hat{\mathrm{S}}$E$(\hat{\lambda})$	0.0063	0.0067	0.0059	0.0060	0.0079	0.0083	0.0068	0.0069
	95%CV	94.7	94.6	93.9	93.7	94.2	93.6	93.6	93.5
1	$\hat{\lambda}$	0.0306	0.0307	0.0303	0.0303	0.0305	0.0307	0.0305	0.0305
	S$\hat{\mathrm{E}}(\hat{\lambda})$	0.0054	0.0066	0.0051	0.0058	0.0068	0.0082	0.0060	0.0067
	E$\hat{\mathrm{S}}$E$(\hat{\lambda})$	0.0054	0.0057	0.0053	0.0060	0.0071	0.0084	0.0063	0.0069
	95%CV	94.5	95.6	94.0	94.0	93.7	93.3	93.7	94.0

2.5 Illustration of the proposed model using BAIS III data set

We applied the proposed model to the Botswana AIDS impact survey (BAIS III) data of 2008. BAIS III was the third sexual behaviour national population level survey. More importantly, BAIS III has gone beyond the traditional aims of assessing knowledge, attitude and behavior regarding HIV and AIDS to estimating the levels of HIV prevalence and incidence in the general population aged 18 months and above (CSO, 2008). This was an important exercise particularly with regard to assessment of the impact of national programs and it also provided the basis of future research including this current paper. BAIS III was a cross sectional study of which includes the use of biomarkers to estimate incidence.

Estimated HIV incidence rate per 100 person-years (I) was calculated using the four estimators:

a) Model 1 (M1): The standard unadjusted estimator (Kaplan and Brookmeyer, 1999; Balasubramanian and Lagakos, 2010)

b) Model 2 (M2): The proposed unadjusted estimator incorporating past prevalence

c) Model 3 (M3): The standard adjusted estimator (Wang and Lagakos, 2009)

d) Model 4 (M4): The proposed adjusted estimator incorporating past prevalence

As cautioned by (Hargrove et al., 2008; McWalter and Welte, 2010; Wang and Lagakos, 2009), the use of locally available estimated parameters cannot be overemphasized. Because the locally estimated parameters are not available the values used in this paper are for illustration purposes only. However, the comparison and conclusions still remain valid with these assumed parameter values. In both scenarios of Model 1 versus Model 2 and Model 3 versus Model 4 we assume a mean window period of 155 days (CSO, 2008). We use a value of $\tau = 1$ year. For models 3 and 4, we assume a false recent rate, $1-p = 1.5\%$. In Botswana, between 2004 and 2008, overall national HIV prevalence increased from 17.1% to 17.6% (CSO, 2008) for individuals aged 1.5 years and above. For purposes of illustration we used simple extrapolation method to estimate the previous prevalence for 2007. The values of θ_0 are provided in Table 2.4.

Table 2.4: Values of θ_0

Demographics		Sample n (n_0, n_r, n_1)	$\theta_0 \approx \frac{n_1+n_r}{n} - 0.00125$
Sex	Male	6516 (5603, 55, 858)	0.139
	Female	7775 (6220, 94, 1461)	0.199
Age (y)	≤ 4	872 (854, 5, 13)	0.019
	5-14	3382 (3242, 18, 122)	0.040
	15-19	1449 (1395, 5, 49)	0.036
	20-29	2967 (2401, 43, 523)	0.190
	30-39	2142 (1287, 33, 822)	0.398
	40-49	1429 (928, 27, 474)	0.349
	50+	2050 (1716, 18, 316)	0.162
Overall	All groups	14291 (11823, 149, 2319)	0.171

The 95% confidence intervals (95% CI) are provided for all the 4 models. We also use all the four models to estimate group specific HIV incidence by sex and age. The results for all the four models are presented in Table 2.5.

In both cases, Model 2 (compared to Model 1) and Model 4 (compared to Model 3) produce better estimates of precision as shown by the narrow confidence intervals. These are settings where we have assumed the previous prevalence is known. We can see that, in all the four models, incidence is higher for females compared to males as it was reported in the BAIS III report. Also, incidence is higher for age groups 20-49 years compared to other age groups. There is a substantial decrease in the estimated incidence when the assumed false recent rate is 1.5%. It is therefore important to make necessary adjustments where there is need because failure to do that will result in overestimation of incidence while making unnecessary adjustments will lead to underestimation of incidence.

Table 2.5: Estimated HIV incidence by gender and age using all the four Models

Estimated HIV incidence in Botswana					
Demographics		I_{M1}(95% CI)	I_{M2}(95% CI)	I_{M3}(95% CI)	I_{M4}(95% CI)
Sex	Male	2.31 (1.70, 2.92)	1.97 (1.46, 2.48)	1.76 (1.14, 2.38)	1.45 (0.94, 1.96)
	Female	3.56 (2.84, 4.28)	3.00 (2.41, 3.59)	2.72 (2.00, 3.44)	2.19 (1.59, 2.79)
Age (y)	1-4	1.38 (0.17, 2.59)	0.89 (0.14, 1.64)	1.32 (0.10, 2.54)	0.83 (0.08, 1.58)
	5-14	1.31 (0.71, 1.91)	1.02 (0.57, 1.47)	1.17 (0.56, 1.78)	0.89 (0.44, 1.34)
	15-19	0.84 (0.10, 1.58)	0.70 (0.11, 1.29)	0.72 (0.00, 1.47)	0.58 (0.00, 1.18)
	20-29	4.22 (2.96, 5.48)	3.43 (2.43, 4.43)	3.44 (2.17, 4.71)	2.68 (1.67, 3.69)
	30-39	6.04 (3.98, 8.10)	5.20 (3.46, 6.94)	3.75 (1.67, 5.83)	3.06 (1.20, 4.82)
	40-49	6.85 (4.27, 9.43)	5.69 (3.59, 7.79)	5.02 (2.42, 7.62)	3.94 (1.82, 6.06)
	50+	2.47 (1.33, 3.61)	2.12 (1.17, 3.07)	1.81 (0.66, 2.96)	1.50 (0.53, 2.47)
Overall	All groups	2.97 (2.49, 3.45)	2.51 (2.12, 2.90)	2.26 (1.78, 2.74)	1.84 (1.44, 2.24)

2.6 Discussion

We propose a method for estimating incidence of a disease, such as HIV, from a cross-sectional study when previous prevalence is known. We find that in some settings such a modification can modestly improve the precision of the estimator of incidence. In order to compare our model, which assumes known previous prevalence, with the standard model which assumes no previous prevalence we assume a known value of the mean window period of 155 days, the same as the one that was used in BAIS III data analysis (CSO, 2008). In practice, the assumed false recent rate and the mean window need to be estimated externally and therefore there is a need to account for their uncertainty. However, the proposed method do not take into account the uncertainty in the false recent rate and the mean window period. Although uncertainty in the false recent rate and the mean window period affects estimators of incidence, we note that for comparison between the two methods, the uncertainties of the two estimators will not affect the conclusions. This is because the two models (with and without previous prevalence) were compared under the same conditions of not accounting for uncertainties. Methods of handling these uncertainties have been proposed by Wang and Lagakos (2010) through the idea of augmented designs where all or some individuals who are tested as recently infected are followed until they exit the recent infection state. The design allows for internal estimation of both the false recent rate and the mean window period and subsequently take into account the uncertainties associated with these estimators. To ensure that our model did not overestimate incidence as a result of assay non-progression, we extend the idea of (McWalter and Welte, 2010; Wang and Lagakos, 2009) through a likelihood approach to account for false recent rate.

It is important to note that in the era of improved standard of care where most people living with HIV receive antiretroviral therapy (ARV) which may suppress the viral load below detection limit, it is possible that some individuals on ARV may potentially ‘return’ or ‘revert’ back to the recent infection state (Wang and Lagakos, 2009; Brookmeyer, 2010b; McWalter and Welte, 2010). However it has been noted that where ARV usage is documented, all subjects on treatment can be assumed to be in non-recent state because treatment initiation is often long after sero-conversion;

usually after 1 to 2 years depending on the genetic structure of the individual (Wang and Lagakos, 2009) which can slow or enhance disease progression. Another point worth noting is that previous prevalence estimates may not always be informative about current prevalence estimates, especially when there are innovations in diagnostics and/or treatments. To address this, we shall consider alternative methods to improve our estimates of incidence. Examples of such methods is where we relax the assumption of constant incidence density function by considering the linear incidence density function.

Alternatively, to overcome this problem, a novel though expensive approach is to design an assay, with high sensitivity and specificity, for identifying new infections based on the characteristics of HIV gene diversification within an infected individual (Park et al., 2011).

One of the findings of the current research is that incidence estimates differ between males and females and between age groups hence future research should aim to incorporate covariates dependence on incidence rate estimation. An additional future enhancement to the proposed method will also be to incorporate uncertainty in past prevalence.

APPENDIX: Estimating Quadratic Equation

The log likelihood function, $\ell(\lambda)$, where now only λ is unknown, for the unadjusted model where the previous prevalence is known, can be simplified to:

$$\begin{aligned}\ell(\lambda) &= -n_0\log(1+\lambda\tau)+k+n_r\log(\lambda)-n_r\log(1+\lambda\tau)\\ &\quad +n_1\log\left[\theta_0+\lambda(\tau-\mu+\theta_0\mu)\right]-n_1\log(1+\lambda\tau)\end{aligned}$$

Hence we can write,

$$\ell(\lambda) = -n\log(1+\lambda\tau)+k+n_r\log(\lambda)+n_1\log\left[\theta_0+\lambda\left\{\tau-(1-\theta_0)\mu\right\}\right]$$

Where

$$k = n_0\log(1-\theta_0)+n_r\log\left[(1-\theta_0)\mu\right]$$

and

$$n = n_0+n_r+n_1.$$

Thus

$$\frac{\partial\ell}{\partial\lambda} = -\frac{n\tau}{1+\lambda\tau}+\frac{n_r}{\lambda}+\frac{n_1\left[\tau-(1-\theta_0)\mu\right]}{\theta_0+\lambda\left[\tau-(1-\theta_0)\mu\right]}$$

and by setting $\frac{\partial\ell}{\partial\lambda}=0$ leads to

$$\frac{n\tau}{1+\lambda\tau} = \frac{n_r}{\lambda}+\frac{n_1\left[\tau-(1-\theta_0)\mu\right]}{\theta_0+\lambda\left[\tau-(1-\theta_0)\mu\right]}$$

Simplifying the above equation, we obtain a quadratic equation in λ:

$$\lambda^2 n_0\tau\left[\tau-(1-\theta_0)\mu\right]+\lambda\left[n\tau\theta_0-n_r\theta_0\tau-(n_r+n_1)\left\{\tau-(1-\theta_0)\mu\right\}\right]-n_r\theta_0=0$$

Chapter 3

Estimating HIV Incidence with adjustment for Sensitivity and Specificity

SUMMARY. Diagnostic tests used to determine recent infections are prone to errors. The common error being the one where long standing or prevalent infections are misclassified as recent. As a result, estimates of HIV incidence derived from cross sectional surveys using biomarkers such as the detuned ELISA or the BED capture enzyme immunoassay have been reported to be significantly higher than from prospective cohort studies (Karita et al., 2007; McDougal et al., 2006; Hargrove et al., 2008). This then led to proposals for adjustments of these assay based estimates to correct for this misclassification. Adjustment procedures for handling these misclassifications were first proposed by Parekh et al. (2002) then later by (McDougal et al., 2006; Hargrove et al., 2008). Another adjusted estimator of HIV incidence was developed by McWalter and Welte (2010). The same estimator was developed by Wang and Lagakos (2009) using the method of maximum likelihood estimation. However, this estimator does not take into account individuals who are misclassified as non-recent while in fact they are recent. In this paper, we propose a new estimator of incidence rate that account for both of these misclassifications.. The method is applied to data from Botswana HIV/AIDS impact study of 2008.

KEYWORDS: HIV incidence, Sensitivity, Specificity, BED assay, Maximum likelihood

3.1 Introduction

Cross-sectional approach for estimating HIV incidence, through a novel two-stage serologic sensitive/less sensitive testing algorithm for detecting recent HIV seroconversion (STARHS), has offered advantages to traditional longitudinal cohort studies in terms of cost, follow up bias and time (Brookmeyer et al., 1995; Janssen et al., 1998; Wang and Lagakos, 2009). In stage one, a sensitive test, such as standard the ELISA, is used to diagnose HIV infection then a less sensitive test such as the standard BED capture enzyme immunoassay (BED assay) is used in the second stage to distinguish between long standing infections and recent infections among those who tested positive for HIV in stage one. However, there has been concerns that BED assays do not properly take into account individuals with long window periods (non-progressors/non-reactors), individuals on anti-retroviral treatments (ARTs) and those with AIDS defining conditions and often classify such cases as recent infections while they are not (Karita et al., 2007; McDougal et al., 2006; Hargrove et al., 2008). Misclassification could arise because the proportion of IgG that is HIV antibody could fall below the threshold in response to either ARTs or onset of some opportunistic infections (Hallett et al., 2009; Hayashida et al., 2008). As a consequence of this, estimates of HIV incidence derived from cross-sectional surveys of biomarkers has been reported to be higher than the cohort estimates (Karita et al., 2007; McDougal et al., 2006; Hargrove et al., 2008). This then led to proposals for adjustments of these assay based estimates to correct for this misclassification. Adjustment procedures for handling these misclassifications were first proposed by Parekh et al. (2002) then later by (McDougal et al., 2006; Hargrove et al., 2008; McWalter and Welte, 2010; Wang and Lagakos, 2009; Welte et al., 2009).

The McDougal et al. (2006) method of adjustment uses the sensitivity (the proportion of recent specimens that test positive, that is, below the normalized optical density threshold) within the mean window period (μ) and both the short term and long term specificities (the proportion of longstanding specimens that test negative, that is, above normalized optical density threshold). In this case, the short term specificity refers to the specificity of BED in the period between μ and 2μ while long term specificity refers to the specificity of BED for the period more than 2μ.

Upon realizing that the McDougal et al. (2006) adjustments were over-parameterized, Hargrove et al. (2008) developed an alternative adjustment procedure with few parameters. Generally the Hargrove et al. (2008) adjustment is a simplified version of the McDougal et al. (2006) with sensitivity and short term specificity assumed to be equal. In particular, the Hargrove et al. (2008) estimator depends mainly on the the false recent rate and the mean window period.

A different strategy for adjustment for the 4-state disease progression model that takes into account specificity and sensitivity of diagnostic test was proposed by Balasubramanian and Lagakos (2010) using the method of maximum likelihood estimation.

The 4-state model of Balasubramanian and Lagakos (2010) requires that all subjects who test negative on a sensitive test be tested again using another test to determine the number of newly infected but have not seroconverted. This could be an expensive exercise as reported by Janssen et al. (1998) who subsequently proposed a simple 3-state model.

Another adjusted estimator of HIV incidence was developed by McWalter and Welte (2010) using mathematical models that take into account subjects who repeatedly test negative on the less sensitive test and positive on the sensitive test long after they have seroconverted. The same estimator was developed by Wang and Lagakos (2009) using the method of maximum likelihood estimation. However, this estimator does not take into account individuals who are misclassified as non-recent while in fact they are recent. We argue that as long as BED is used as a diagnostic test then it is subject to this misclassification error which will have an effect on the estimated incidence rate.

We propose an estimator of incidence that takes into account subjects who are misclassified as non-recent while in fact they are recent. The paper builds on the work of Balasubramanian and Lagakos (2010) by considering different strategies for adjusting estimated HIV incidence for the 3-state disease progression model when the BED assay is used as the diagnostic test assuming the sensitivity and specificity of the antibody test is 100%. We also show that if specificity is equal to one then the proposed estimator reduces to the one proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009). We note that what we define as sensitivity is what (McWalter and Welte, 2010; Wang and Lagakos, 2009) define as the proportion of assay progressors (p). This is because assay progressors are subjects who will be correctly classified to be in state 3 as we shall see in Section 3.2 when we look at the preliminary concepts. In the remaining sections we shall proceed as follows: the proposed method is introduced in Section 3.3. In Section 3.4 we conduct a simulation study to evaluate the performance of the proposed estimator. We illustrate its use in Section 3.5 and finally the discussion is presented in Section 3.6.

3.2 Preliminary Concepts

We consider the 3-state longitudinal disease progression model for the natural history of HIV/AIDS and classification of subjects is by a diagnostic test such that state 1 represents the pre-seroconversion period. We denote state 1 by S_1 because it is corresponding to the period in which an individual is either infected but have have not yet seroconverted or is uninfected. State 2, denoted by S_2, represents the "recent infection" state where HIV antibodies are detectable by a sensitive diagnostic test but not yet through a less sensitive test such as BED assay. Finally state 3, denoted by S_3, represents the "non-recent infection" state in which HIV antibodies are detectable by a less sensitive test (and sensitive test). We assume the incidence density is constant.

Definition 3.2.1. We define sensitivity (s_B) for the BED assay as the probability that the BED assay is positive given that the subject is in state S_3. That is,

$$s_B = P(B + |S_3).$$

Definition 3.2.2. We define specificity (p_B) for the BED assay as the probability that the BED assay is negative given that the subject is in state S_1 or S_2. So

$$p_B = P(B - | S_1, S_2)$$

Since s_B and p_B are probabilities, then they must be greater than zero.

3.3 Incorporating Imperfect Sensitivity and Specificity

Consider a random sample of size n from the population at some calendar time t. Let n_0 be the number of subjects in S_1, n_r be the number of subjects in S_2 and finally n_1 be the number of subjects in S_3 such that $n = n_0 + n_r + n_1$.

Using similar arguments as in Balasubramanian and Lagakos (2010), it can be shown that the prevalence probabilities for the 3-states at time t are

$$\begin{aligned}
\pi_1(t) &= 1 - \theta \\
\pi_2(t) &= f\mu \\
\pi_3(t) &= \theta - f\mu
\end{aligned}$$

The objective is to estimate HIV incidence rate,

$$\lambda = \frac{f}{1 - \theta}$$

where f is the incidence density function and for the purpose of the current analysis, is assumed to be constant as in Balasubramanian and Lagakos (2010), and θ is the HIV prevalence at time t.

It follows that the trinomial log-likelihood function for this setting is

$$\ell(f,\theta) = n_0\text{log}[1-\theta] + n_r\text{log}[f\mu] + n_1\text{log}[\theta - f\mu] \tag{3.3.1}$$

assuming μ is known.

Since

$$f = \lambda(1-\theta)$$

then Eq (3.3.1) can be rewritten as

$$\ell(\lambda,\theta) = n_0\text{log}[1-\theta] + n_r\text{log}[\lambda(1-\theta)\mu] + n_1\text{log}[\theta - \lambda(1-\theta)\mu] \tag{3.3.2}$$

Finally Eq (3.3.2) can be modified to incorporate sensitivity and specificity of the BED assay and the corresponding log-likelihood function, with $s_B + p_B \neq 1$, is

$$\begin{aligned}\ell(\lambda,\theta) \;&=\; n_0\text{log}[(1-\theta)p_B] \\ &\quad +n_r\text{log}[\lambda(1-\theta)\mu p_B + \{\theta - \lambda(1-\theta)\mu\}(1-s_B)] \\ &\quad +n_1\text{log}[\lambda(1-\theta)\mu(1-p_B) + \{\theta - \lambda(1-\theta)\mu\}s_B] \end{aligned} \tag{3.3.3}$$

The maximum likelihood estimators for λ and θ are respectively:

$$\hat{\lambda} = \frac{n_r s_B - (1-s_B)n_1}{n_0\mu(s_B + p_B - 1)} \tag{3.3.4}$$

$$\hat{\theta} = \frac{n_1 + n_r}{n} \tag{3.3.5}$$

where $n = n_0 + n_r + n_1$

The corresponding estimates of variances of $\hat{\lambda}$ and $\hat{\theta}$ can be obtained as the diagonals of the inverse of the matrix of negative second derivatives of $\ell(\lambda,\theta)$ with λ and θ replaced by their maximum likelihood estimates. These can be shown to be

$$\begin{aligned}\text{var}(\hat{\lambda}) &= \frac{(n_r+n_1)(ns_B-2n_1)s_B+n_0n_1(1-2s_B)+n_1^2}{n_0^3\mu^2(p_B+s_B-1)^2} && (3.3.6)\\ \text{var}(\hat{\theta}) &= \frac{n_0(n_r+n_1)}{n^3}\\ &= \frac{n_0(n-n_0)}{n^3} && (3.3.7)\end{aligned}$$

The approximate covariance between $\hat{\lambda}$ and $\hat{\theta}$ is given by:

$$\text{cov}(\hat{\lambda},\hat{\theta}) = \frac{(n_r+n_1)s_B-n_1}{n_0n\mu(p_B+s_B-1)} \qquad (3.3.8)$$

where $n = n_0 + n_r + n_1$.

When the BED assay is assumed to have perfect sensitivity and specificity then the MLEs for λ and θ respectively simplifies to:

$$\hat{\lambda} = \frac{n_r}{n_0\mu} \qquad (3.3.9)$$

$$\hat{\theta} = \frac{n_1+n_r}{n} \qquad (3.3.10)$$

Eq (3.3.9) is the estimator of incidence under the standard 3-state model (Kaplan and Brookmeyer, 1999; Balasubramanian and Lagakos, 2010). Likewise, the corresponding estimates of the variances, of $\hat{\lambda}$ and $\hat{\theta}$, covariance between $\hat{\lambda}$ and $\hat{\theta}$ assuming perfect sensitivity and specificity of the BED respectively simplifies to:

$$\begin{aligned}\text{var}(\hat{\lambda}) &= \frac{n_r(n_0+n_r)}{n_0^3\mu^2} && (3.3.11)\\ \text{var}(\hat{\theta}) &= \frac{n_0(n-n_0)}{n^3} && (3.3.12)\end{aligned}$$

and

$$\text{cov}(\hat{\lambda}, \hat{\theta}) \quad = \quad \frac{n_r}{n_0 n \mu} \tag{3.3.13}$$

3.3.1 The case of $p_B = 1$ and $S_B < 1$

Note that our main interest is in the estimation of the rate of new infection or incidence rate, λ. Consider the estimator in Eq (3.3.4). When $p_B = 1$ it reduces to

$$\hat{\lambda} = \frac{n_r s_B - (1 - s_B) n_1}{n_0 \mu s_B} \tag{3.3.14}$$

where now $s_B = p$, the proportion of assay progressors as defined by Wang and Lagakos (2009).

The variance of $\hat{\lambda}$ reduces to

$$\text{var}(\hat{\lambda}) = \frac{(n_r + n_1)(n s_B - 2n_1) s_B + n_0 n_1 (1 - 2s_B) + n_1^2}{n_0^3 \mu^2 s_B^2} \tag{3.3.15}$$

where $n = n_0 + n_r + n_1$.

3.3.2 The case when $s_B = 1$ and $p_B < 1$

When $s_B = 1$ the estimator in Eq (3.3.4) reduces to

$$\hat{\lambda} = \frac{n_r}{n_0 \mu p_B} \tag{3.3.16}$$

and the variance reduces to

$$\text{var}(\hat{\lambda}) = \frac{n_r (n_0 + n_r)}{n_0^3 \mu^2 p_B^2} \tag{3.3.17}$$

3.4 Simulation

In this section, we conduct a simulation study to evaluate the performance of the proposed estimator of HIV incidence when sensitivity and specificity are incorporated. In this study 1000 simulations and a sample of size $n = 3000$ have been used. The data was simulated from a trinomial distribution with the following parameters

$$\begin{aligned}\pi_1(t) &= (1-\theta)p_B \\ \pi_2(t) &= \lambda(1-\theta)\mu p_B + \{\theta - \lambda(1-\theta)\mu\}(1-s_B) \\ \pi_3(t) &= 1 - \pi_1(t) - \pi_2(t)\end{aligned}$$

See Eq (3.3.3) for more details.

The results of the simulation study are given in Table 3.1.

From Table 3.1, we can see that in almost all the cases, except when $(s_B = 0.97, p_B = 0.97)$, the coverage probabilities are close to the nominal 95% level. When (s_B, p_B) gets smaller then the coverage probabilities gets smaller than the nominal level and the standard errors also gets large. The coverage probability is more sensitive to the case when $(s_B < 1, p_B = 1)$ than when $(s_B = 1, p_B < 1)$. The $(s_B < 1, p_B = 1)$ setting corresponds to the McWalter and Welte (2010); Wang and Lagakos (2009) estimator. The standard errors are smaller for $(s_B = 1, p_B < 1)$ than for $(s_B < 1, p_B = 1)$. When $s_B = 1$ and p_B gets much smaller, say $p_B < 0.97$ then the proposed estimator overestimates the true incidence rate even though the coverage probabilities are closer to the nominal level. Consistent with the theory and reality, precision improves for larger values of (s_B, p_B), say $(s_B \geq 0.99, p_B \geq 0.99)$. It worth noting that if $(s_B < 0.95, p_B < 0.95)$ then the method will underestimate incidence unless either

Table 3.1: *Results of a simulation study are provided for the proposed model for different values of sensitivity* (s_B) *and specificity* (p_B)*. The true value of incidence rate,* $I_{true} = 0.03$ *and the true value of theta is* $\theta = 0.20$*.* $\hat{\lambda}$*,* $\hat{SE}(\hat{\lambda})$ *and* $E\hat{SE}(\hat{\lambda})$ *represents the average estimates, estimated standard errors and the corresponding empirical estimates respectively from 1000 simulations.* 95%*CV denotes the proportion of experiments in which* I_{true} *is contained in the nominal* 95% *confidence interval.*

Simulation Results with $I_{true} = 0.03$ and $\theta = 0.20$					
(s_B, p_B)	Estimates		(s_B, p_B)	Estimates	
(1, 1)	$\hat{\lambda}$	0.0301	(0.98, 0.98)	$\hat{\lambda}$	0.0300
	$\hat{SE}(\hat{\lambda})$	0.0055		$\hat{SE}(\hat{\lambda})$	0.0067
	$E\hat{SE}(\hat{\lambda})$	0.0056		$E\hat{SE}(\hat{\lambda})$	0.0069
	95%CV	94.1		95%CV	94.1
(0.99, 0.99)	$\hat{\lambda}$	0.0302	(0.97, 0.97)	$\hat{\lambda}$	0.0290
	$\hat{SE}(\hat{\lambda})$	0.0061		$\hat{SE}(\hat{\lambda})$	0.0072
	$E\hat{SE}(\hat{\lambda})$	0.0065		$E\hat{SE}(\hat{\lambda})$	0.0076
	95%CV	93.3		95%CV	92.3
(0.99,1)	$\hat{\lambda}$	0.0302	(0.97, 1)	$\hat{\lambda}$	0.0305
	$\hat{SE}(\hat{\lambda})$	0.0060		$\hat{SE}(\hat{\lambda})$	0.0069
	$E\hat{SE}(\hat{\lambda})$	0.0062		$E\hat{SE}(\hat{\lambda})$	0.0072
	95%CV	93.4		95%CV	93.3
(1, 0.99)	$\hat{\lambda}$	0.0305	(1, 0.97)	$\hat{\lambda}$	0.0312
	$\hat{SE}(\hat{\lambda})$	0.0056		$\hat{SE}(\hat{\lambda})$	0.0057
	$E\hat{SE}(\hat{\lambda})$	0.0057		$E\hat{SE}(\hat{\lambda})$	0.0059
	95%CV	94.1		95%CV	94.4
(1,0.90)	$\hat{\lambda}$	0.0338	(0.90, 1)	$\hat{\lambda}$	0.0301
	$\hat{SE}(\hat{\lambda})$	0.0064		$\hat{SE}(\hat{\lambda})$	0.0095
	$E\hat{SE}(\hat{\lambda})$	0.0064		$E\hat{SE}(\hat{\lambda})$	0.0095
	95%CV	93.4		95%CV	94.3
(0.90, 0.90)	$\hat{\lambda}$	0.0008	(0.95, 0.95)	$\hat{\lambda}$	0.0252
	$\hat{SE}(\hat{\lambda})$	0.0119		$\hat{SE}(\hat{\lambda})$	0.0085
	$E\hat{SE}(\hat{\lambda})$	0.0120		$E\hat{SE}(\hat{\lambda})$	0.0087
	95%CV	32.0		95%CV	87.5

sensitivity and/or specificity is 100%.

3.5 Application to BAIS III data set

3.5.1 Introduction

BAIS III of 2008 from Botswana is used to illustrate the use of the proposed estimator. For purposes of illustration, we assume a mean window period of $\mu = E(L_2) = 155$ days which is the same mean window period that was used in BAIS III report (CSO, 2008). Table 3.2 presents a summary of the data necessary for this analysis. The data shows the distribution of the sampled individuals into the 3-states disease model but stratified according to sex and age respectively

Table 3.2: Summary of BAIS III data stratified by age and sex

Demographics		Sample n (n_0, n_r, n_1)
Sex	Male	6516 (5603, 55, 858)
	Female	7775 (6220, 94, 1461)
Age (y)	≤ 4	872 (854, 5, 13)
	5-14	3382 (3242, 18, 122)
	15-19	1449 (1395, 5, 49)
	20-29	2967 (2401, 43, 523)
	30-39	2142 (1287, 33, 822)
	40-49	1429 (928, 27, 474)
	50+	2050 (1716, 18, 316)
Overall	All groups	14291 (11823, 149, 2319)

3.5.2 Results

We compared the adjusted and unadjusted estimates of HIV incidence rate (I) expressed as per 100 person year using two models;

- The proposed adjusted estimator, M1
- The unadjusted estimator, when $(s_B = 1, p_B = 1)$, M2

Table 3.3 presents the estimated HIV incidence rate (in 100 person years) for both models together with the corresponding 95% confidence intervals. We estimated HIV incidence rate stratified by sex and age.

Table 3.3: Adjusted (adj) and unadjusted (unadj) estimates of HIV incidence (I) in Botswana by sex and age using M1 and M2

Estimates of HIV incidence in Botswana adjusted for $(p_B,\, s_B)$			
		(0.99, 0.99)	(1, 1)
Demographics		I_{M1}(95% CI)	I_{M2}(95% CI)
Sex	Male	1.97 (1.35, 2.59)	2.31 (1.70, 2.93)
	Female	3.03 (2.30, 3.76)	3.56 (2.83, 4.28)
Age (y)	1-4	1.36 (0.13, 2.58)	1.38 (0.17, 2.59)
	5-14	1.23 (0.62, 1.84)	1.31 (0.70, 1.91)
	15-19	0.77 (0.02, 1.51)	0.84 (0.10, 1.59)
	20-29	3.74 (2.45, 5.02)	4.22(2.95, 5.49)
	30-39	4.56 (2.47, 6.66)	6.04 (3.95, 8.12)
	40-49	5.69 (3.06, 8.33)	6.85 (4.22, 9.47)
	50+	2.05 (0.89, 3.21)	2.47 (1.32, 3.62)
Overall		2.53 (2.04, 3.01)	2.97 (2.49, 3.45)

We can see from Table 3.3 that the estimated HIV incidence rate is smaller (as expected) for the adjusted estimator as compared to the unadjusted estimator. However, the confidence intervals are wider for the adjusted estimates. For example, for females, the unadjusted (Model 2) estimate (and 95% CI) of HIV incidence rate is

3.56 (2.83, 4.28) and the width of the confidence interval is 1.45 while for the adjusted (Model 1) the estimates are 3.03 (2.30, 3.76) and the width of the confidence interval is 1.46. The HIV incidence rate is higher for the middle age groups (20-49 years) in agreement with what is reported in the literature, for example, in CSO (2008). The results also show that incidence rate is higher in females than males.

3.6 Discussion

Accurate estimation of HIV incidence require diagnostic tests with very high sensitivity and specificity. We have proposed an estimator of incidence rate that incorporates the sensitivity and specificity of the BED as a diagnostic test. We have seen that if the BED is assumed to have 100% sensitivity, then failure to adjust for specificity (if indeed there are subjects who have been misclassified to be in state 3 while they are supposed to be state 2) will underestimate incidence rate. If BED is assumed to have 100% specificity, it follows that failure to account for sensitivity (if indeed there are subjects who have been misclassified to be in state 2 while they are supposed to be state 3) will lead to incidence rate estimates that are too high. The estimator that we have proposed makes the two adjustments simultaneously.

We advice that adjustments of estimates of HIV incidence in this setting should be incorporated with appropriate values of sensitivity and specificity. However, as noted by Wang and Lagakos (2009), these values are often not available and therefore need to be estimated from follow up studies where the time of seroconversion will be known (although not exactly) with some accuracy. Other estimates can be obtained from the literature provided the characteristics of the study population are similar. In the current analysis we considered the fact that BAIS III covered the whole country and hence given the diversity of the groups, then it is possible to have individuals who might be wrongly diagnosed as recent infections and those who might have been wrongly diagnosed as non-recent infections and therefore adjustments for both sensitivity and specificity are very important for accurate estimation of HIV incidence rate. However, we did not have estimates of sensitivity and specificity therefore we assumed some values for illustration purposes.

All subjects on ARTs were included in the analysis and classified as long standing infections because treatment initiation usually starts several years after seroconversion as proposed by Wang and Lagakos (2009). Luckily, in the BAIS III data set, all subjects on ARTs were classified as non-recent by the BED test. We also noted that about 18 subjects refused to disclose whether or not they are on ARTs. All these subjects were later classified by the BED test as non-recent. It is highly likely that refusal to disclose ARTs usage was associated with ARTs usage. More research is needed to address the effect of ART usage on the estimation of HIV incidence rate.

We assumed a known mean widow period of 155 days. But if a biased estimate of the mean window period is used then all estimates of HIV incidence will also be biased. Methods for improving the accuracy of the estimates of HIV incidence derived from cross sectional surveys of biomarkers include augmented studies such as the one proposed by (Wang and Lagakos, 2010; Claggett et al., 2012) are needed so as to get better estimates of the sensitivity and mean window period though it may be of little use to the estimation of specificity. The current analysis did not take into account the uncertainties in the mean window period, sensitivity and specificity. Other techniques for handling the uncertainty in the mean window period were discussed by Cole et al.

(2006) who proposed the Monte Carlo-based confidence intervals to account for the random error of the mean widow period. Chu and Cole (2006) proposed the Bayesian method for incorporating the uncertainty in the mean widow period while Janssen et al. (1998) used the Bonferroni-box procedure with exact Poisson assumption on n_r, the number of subjects testing positive on the sensitive test and negative on the less sensitive test. Thus an additional future enhancement to the proposed method will also be to incorporate uncertainties in the mean window period, sensitivity and specificity.

APPENDIX: The log-likelihood equations for estimating HIV incidence rate and prevalence

The likelihood function is

$$\begin{aligned} \mathrm{L}(\lambda,\theta) &= [(1-\theta)p_B]^{n_0} \\ &\quad \times[\lambda(1-\theta)\mu p_B + \{\theta-\lambda(1-\theta)\mu\}(1-s_B)]^{n_r} \\ &\quad \times[\lambda(1-\theta)\mu(1-p_B) + \{\theta-\lambda(1-\theta)\mu\}s_B]^{n_1} \end{aligned}$$

And the corresponding log-likelihood function is

$$\begin{aligned} \ell(\lambda,\theta) &= n_0\log[(1-\theta)p_B] \\ &\quad +n_r\log[\lambda(1-\theta)\mu p_B + \{\theta-\lambda(1-\theta)\mu\}(1-s_B)] \\ &\quad +n_1\log[\lambda(1-\theta)\mu(1-p_B) + \{\theta-\lambda(1-\theta)\mu\}s_B] \end{aligned}$$

The partial derivatives with respect to λ and θ respectively are:

$$\begin{aligned}
\frac{\partial \ell(\lambda,\theta)}{\partial \lambda} &= \frac{n_r[(1-\theta)\mu p_B - (1-\theta)\mu(1-s_B)]}{\lambda(1-\theta)\mu p_B + (\theta - \lambda(1-\theta)\mu)(1-s_B)} \\
&\quad + \frac{n_1[(1-\theta)\mu(1-p_B) - (1-\theta)\mu s_B]}{\lambda(1-\theta)\mu(1-s_B) + \{\theta - \lambda(1-\theta)\mu\}\, s_B} \\
\frac{\partial \ell(\lambda,\theta)}{\partial \theta} &= -\frac{n_0}{(1-\theta)} \\
&\quad + \frac{n_r[-\lambda\mu p_B + (1+\lambda\mu)(1-s_B)]}{\lambda(1-\theta)\mu p_B + \{\theta - \lambda(1-\theta)\mu\}\,(1-s_B)} \\
&\quad + \frac{n_1[-\lambda\mu(1-p_B) + (1+\lambda\mu)s_B]}{\lambda(1-\theta)\mu(1-p_B) + \{\theta - \lambda(1-\theta)\mu\}\, s_B}
\end{aligned}$$

Chapter 4

Estimation of HIV incidence under linear incidence density function

SUMMARY. The method for estimating HIV incidence proposed by Balasubramanian and Lagakos (2010) and also used in Gabaitiri et al. (2013) assumes that the incidence density function is constant over time. This could be a reasonable assumption in settings where changes in HIV incidence overtime are very small and the time between the two surveys as proposed by Gabaitiri et al. (2013) is small. However, if HIV incidence is dropping due to the effect of preventive measures such as frequent condoms use, the impact of other risk reduction programmes, and also if the period between two successive surveys is large then the constant incidence density assumption may not be reasonable. In this paper, we relax this assumption and derive the maximum likelihood estimator (MLE) of HIV incidence. In particular, we derive the MLE of HIV incidence when the incidence density is assumed to be linear. We assume that the main mode of transmission in the target population is via heterosexual. The proposed method is illustrated using data from the Botswana AIDS Impact (BAIS) III survey of 2008.

KEYWORDS: HIV incidence, linear incidence density, Maximum likelihood, previous prevalence

4.1 Introduction

In this chapter, we relax the assumption of constant incidence density. We note that although the method proposed by Balasubramanian and Lagakos (2010) provides a general mathematical framework for the estimation of HIV incidence, it also assumes that the incidence density is constant over time. This could be a reasonable assumption in settings where changes in HIV incidence are very small overtime. However, if HIV incidence is dropping due to the effect of preventive measures such as frequent use of condoms during sex, the impact of other risk reduction programmes, and also if the period between two successive surveys is large then the constant incidence density assumption may not be reasonable. In this paper, we relax this assumption and derive the maximum likelihood estimator (MLE) of HIV incidence. In particular, we derive the MLE of HIV incidence when the incidence density is assumed to be linear.

We shall proceed as follows, in Section 4.2 we shall review the constant incidence density model. Section 4.3 we introduce the linear incidence density function and develop the likelihood function and then derive the maximum likelihood estimator (MLE) of HIV incidence under this assumption. Inference measures including estimates of standard errors of the MLE of incidence will be considered in Section 4.3.1. In Section 4.4 we shall perform a simulation study to investigate some properties of the developed estimator and in Section 4.5 we shall apply our method to the Botswana HIV/AIDS Impact Survey (BAIS III) data of 2008. A final discussion of the chapter

Table 4.1: Summary of the 3-States

State	Elisa	BED	Implication
S_1	E^-	B^-	pre-seroconversion
S_2	E^+	B^-	recent infection
S_3	E^+	B^+	non-recent infection

will be provided in Section 4.6.

4.2 A review of the constant incidence density model

Consider the 3-state longitudinal disease progression model for the natural history of HIV/AIDS and classification of subjects is by a diagnostic test such that state 1 represents the pre-seroconversion period denoted by S_1 because it corresponds to the period in which an individual is either infected but have have not yet seroconverted (acute infection period) or is uninfected. State 2, denoted by S_2, represents the "recent infection" state where HIV antibodies are detectable by a sensitive diagnostic test but not yet through a less sensitive test such as the BED assay. Finally state 3, denoted by S_3, represents the "non-recent infection" state in which HIV antibodies are detectable by a less sensitive test (and sensitive test). We assume two diagnostic tests for HIV are administered on an individual (typically standard ELISA denoted by E and BED assay denoted by B). Table 4.1 summarizes the three states assuming perfect sensitivity and specificity of the diagnostic tests.

Note that the three states model is a special case of the four state model described by Balasubramanian and Lagakos (2010), where individuals in S_1 are further sub-divided into uninfected and infected but not yet sero-converted.

Suppose $T \geq 0$ denotes the calendar time of HIV infection for someone born at time 0. Furthermore, let $f(t)$, $F(t)$ and $\lambda(t)$ denote the density function (incidence density), cumulative distribution function and hazard function for becoming infected at time t respectively. Let L_2 denote the sojourn or residence time in S_2 with the corresponding cumulative distribution denoted by $G_2(\cdot)$ as in Balasubramanian and Lagakos (2010) and the corresponding probability density function given by $g_2(\cdot)$. We assume that L_2 has support in $[0, L_2^*]$, where $L_2^* < t$ and is independent of T.

The prevalence probabilities for the 3-states model at time t are given by:

$$\begin{aligned}
\pi_1(t) &\overset{def}{=} P(\text{in } S_1 \text{ at calender time t}) \\
&= 1 - F(t) \qquad (4.2.1) \\
\pi_2(t) &\overset{def}{=} P(\text{in } S_2 \text{ at calender time t}) \\
&= \int_{t-L_2^*}^{t} f(u)\,[1 - G_2(t-u)]\,\mathrm{d}u \qquad (4.2.2) \\
\pi_3(t) &= 1 - \pi_1(t) - \pi_2(t) \qquad (4.2.3)
\end{aligned}$$

Eq (4.2.2) has been used as a basis for estimating HIV incidence from snapshots or cross sectional samples. See Kaplan and Brookmeyer (1999) for more details. The underlying idea is that there exist a maximum window period such that $1-G_2(L_2^*) = 0$ and under the constant the incidence density function $f(u) = f$ is constant over $[t - L_2^*, t]$. Hence prevalence of new infections at time t, $\pi_2(t) = P$ is given by

$$\begin{aligned}
P &= f \int_{t-L_2^*}^{t} [1 - G_2(t-u)]\,\mathrm{d}u \\
&= f \int_{0}^{L_2^*} [1 - G_2(y)]\,\mathrm{d}y \\
&= fE(L_2) \qquad (4.2.4)
\end{aligned}$$

where $E(L_2)$, commonly referred to as the mean window period by Brookmeyer (2009), is the mean residence time is S_2.

However, Balasubramanian and Lagakos (2010) proposed a more general framework that combines the information from all the three states through a likelihood approach using a multinomial distribution and subsequently derived the maximum likelihood estimator of HIV incidence together with the corresponding estimates of their standard errors. Generally the hazard or incidence rate is given by

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{1 - \theta} \tag{4.2.5}$$

where θ is the cross-sectional prevalence of the disease at time t. Once the form $f(t)$ have been assumed, then $f(\cdot)$ becomes an additional function to be estimated. The constant incidence density assumption simplifies Eq (4.2.2) although in this paper we claim that such an assumption may not hold in some settings. In this paper, the idea advanced is to relax this assumption and consider other forms of $f(t)$ over $[t - L_2^*, t]$. In particular, we consider a linear form of $f(t)$ over $[t - L_2^*, t]$.

4.3 Linear Incidence Density Form

Consider a random sample of size n from the population at some calendar time t. Let n_0 be the number of subjects in S_1, n_r be the number of subjects in S_2 and finally n_1 be the number of subjects in S_3 such that $n = n_0 + n_r + n_1$. Then given the derivations of $\pi_1(t)$, $\pi_2(t)$ and $\pi_3(t)$ in Eqs (4.2.1), (4.2.2) and (4.2.3) respectively,

$$P(n_0, n_r, n_1 | \pi_1(t), \pi_2(t), \pi_3(t)) \sim \text{trinomial}(n; \pi_1(t), \pi_2(t))$$

where $\pi_3(t) = 1 - \pi_1(t) - \pi_2(t)$.

We shall assume that the density function, $f(u) = a + bu$ for $u \in [t - L_2^*, t]$, where the parameters a and b are unknown. The idea is to estimate a, b and θ and subsequently estimate HIV incidence from the equation

$$\lambda(t) = \frac{f(t)}{1-\theta} = \frac{a+bt}{1-\theta} \tag{4.3.1}$$

Incorporating the linear density function in the expression for $\pi_2(t)$ in Eq (4.2.2) we get

$$\begin{aligned}
\pi_2(t) &= \int_{t-L_2^*}^{t} f(u)\left[1 - G_2(t-u)\right] \mathrm{d}u \\
&= \int_{t-L_2^*}^{t} (a+bu)\left[1 - G_2(t-u)\right] \mathrm{d}u \\
&= \int_{t-L_2^*}^{t} a\left[1 - G_2(t-u)\right] \mathrm{d}u + b\int_{t-L_2^*}^{t} u\left[1 - G_2(t-u)\right] \mathrm{d}u \\
&= a\int_{0}^{L_2^*} \left[1 - G_2(y)\right] \mathrm{d}y + b\int_{0}^{L_2^*} (t-y)\left[1 - G_2(y)\right] \mathrm{d}y \\
&= aE(L_2) + b\int_{0}^{L_2^*} t\left[1 - G_2(y)\right] \mathrm{d}y - b\int_{0}^{L_2^*} y\left[1 - G_2(y)\right] \mathrm{d}y \\
&= aE(L_2) + btE(L_2) - bA
\end{aligned}$$

where

$$\begin{aligned}
A &= \int_{0}^{L_2^*} y\left[1 - G_2(y)\right] \mathrm{d}y \\
&= \left[\frac{y^2}{2}\{1 - G_2(y)\}\right]_0^{L_2^*} + \int_{0}^{L_2^*} \frac{y^2}{2} g_2(y) \mathrm{d}y \\
&= 0 + \frac{E(L_2^2)}{2} \text{ , since } 1 - G_2(L_2^*) = 0 \\
&= \frac{E(L_2^2)}{2}
\end{aligned}$$

Finally then,

$$\pi_2(t) = aE(L_2) - \frac{1}{2}b\left[E(L_2^2) - 2btE(L_2)\right]$$

Let $E(L_2) = \mu$ and $E(L_2^2) = \gamma$. Then the prevalence probabilities for the 3-states model at time t under the linear density function assumption are:

$$\begin{aligned}
\pi_1(t) &= 1 - \theta \\
\pi_2(t) &= a\mu - \frac{1}{2}b(\gamma - 2t\mu) \\
\pi_3(t) &= 1 - \pi_1(t) - \pi_2(t) \\
&= \theta - a\mu + \frac{1}{2}b(\gamma - 2t\mu)
\end{aligned}$$

Assuming the mean and variance of L_2, $E(L_2)$ and $Var(L_2)$ respectively, are known, the corresponding trinomial log-likelihood function, $\ell(a, b, \theta)$, will be:

$$\ell(a,b,\theta) = n_0\log(1-\theta) + n_r\log[a\mu - \frac{1}{2}b(\gamma - 2t\mu)] + n_1\log[\theta - a\mu + \frac{1}{2}b(\gamma - 2t\mu)] \quad (4.3.2)$$

The above log-likelihood function is over-parameterized. To ensure identifiability we need to make additional assumptions about the past prevalence.

In most practical situations, past HIV prevalence is often known from previous studies. Hence the standard 3-state model can be extended to incorporate this information. The idea of incorporating past prevalence to improve the estimation of HIV incidence based laboratory assay data was first proposed by Gabaitiri et al. (2013). In particular, we focus on the immediate past HIV prevalence.

Suppose HIV prevalence at time t_1, for $t_1 < t$, is known, that is, $F(t_1) = \theta_0$ is known. In general let $\tau = t - t_1$ denote the time between the current study at time t and the immediate past study at time t_1. We refer to HIV prevalence at time t_1 as past prevalence. Generally the prevalence at time t, $F(t) = \theta$. Assuming linear incidence density, that is,

$$f(u) = a + bu \text{ for } u \in [t - L_2^*, t]$$

θ is given by

$$\begin{aligned} \theta &= \theta_0 + \int_{t_1}^{t} f(u)du \\ &= \theta_0 + \int_{t_1}^{t} (a + bu)du \\ &= \theta_0 + a(t - t_1) + \frac{b}{2}(t^2 - t_1^2) \end{aligned} \tag{4.3.3}$$

Note that if $t_1 = 0$ then $\tau = t$. Therefore we can write $\theta = \theta_0 + a\tau + \frac{b}{2}\tau^2$ which depends on the time between two successive studies, $\tau = t$. For brevity, we shall let $t_1 = 0$. To avoid confusion, we shall use τ to denote the period between successive studies and t to denote the time of the current cross sectional study. If we substitute $\theta_0 + a\tau + \frac{b}{2}\tau^2$ for θ in Eq (4.3.2) then the new log likelihood function will be:

$$\begin{aligned} \ell(a, b) &= n_0 \log(1 - \theta_0 - a\tau - \frac{b\tau^2}{2}) + n_r \log[a\mu - \frac{1}{2}b(\gamma - 2\tau\mu)] \\ &\quad + n_1 \log[\theta_0 + a(\tau - \mu) + \frac{1}{2}b(\tau^2 + \gamma - 2\tau\mu)] \end{aligned} \tag{4.3.4}$$

There are only two unknowns (‘a’ and ‘b’) in Eq (4.3.4) under the assumption that μ and γ are known. The resulting two partial derivatives with respect to ‘a’ and ‘b’ are, respectively:

$$\begin{aligned}\frac{\partial \ell(a,b)}{\partial a} &= -\frac{n_0\tau}{1-\theta_0-a\tau-\frac{b\tau^2}{2}}+\frac{n_r\mu}{a\mu-\frac{1}{2}b(\gamma-2\tau\mu)}\\ &\quad +\frac{n_1(\tau-\mu)}{\theta_0+a(\tau-\mu)+\frac{1}{2}b(\tau^2+\gamma-2\tau\mu)}\\ \frac{\partial \ell(a,b)}{\partial b} &= -\frac{\frac{1}{2}n_0\tau^2}{1-\theta_0-a\tau-\frac{b\tau^2}{2}}-\frac{\frac{1}{2}n_r(\gamma-2\tau\mu)}{a\mu-\frac{1}{2}b(\gamma-2\tau\mu)}\\ &\quad +\frac{\frac{1}{2}n_1(\tau^2+\gamma-2\tau\mu)}{\theta_0+a(\tau-\mu)+\frac{1}{2}b(\tau^2+\gamma-2\tau\mu)}\end{aligned}$$

Thus the MLEs of ‘a’ and ‘b’ denoted by $\hat{a}$ and $\hat{b}$ can be obtained by equating the above two partial derivatives to zero and solving for ‘a’ and ‘b’ simultaneously. Thus $\hat{a}$ and $\hat{b}$ are respectively given by

$$\hat{a} \;=- \;\frac{(2\mu\tau-\gamma)(n_1+n_r-n\theta_0)-n_r\tau^2}{n\tau(\gamma-\tau\mu)} \tag{4.3.5}$$

$$\hat{b} \;=- \;\frac{2[n_r\tau-(n_1+n_r)\mu+n\mu\theta_0]}{n\tau(\gamma-\tau\mu)} \tag{4.3.6}$$

It follows that the MLE of HIV incidence at the time t, $\lambda(t)$, will be

$$\hat{\lambda(t)} = \frac{\hat{a}+\hat{b}\tau}{1-\hat{\theta}} \tag{4.3.7}$$

where $\hat{\theta} = \theta_0 + \hat{a}\tau + \frac{1}{2}\hat{b}\tau^2$ by the invariance property of MLEs and $\hat{a}$ and $\hat{b}$ are as in Eq (4.3.5) and Eq (4.3.6) respectively. Note that the key aim here is the ability to estimate incidence density function, which is assumed linear in our case. A further advantage of using information on immediate past prevalence is that we can measure time from that study.

4.3.1 Inference

To construct the approximate confidence intervals for (a, b, θ or λ) standard errors for their MLEs or functions of their components can be obtained from the Hessian matrix, H, of the log-likelihood function. See Cox and Hinkley (1974) for more details. In the proposed model, we assume that all subjects are tested using two diagnostic tests consisting of the standard ELISA assay and the BED assay. Let $\beta = (a, b)$. Then $\hat{\beta} = (\hat{a}, \hat{b})$ denotes of the MLEs for the parameter vector β. Thus the variance of $\hat{\beta}$ is given by,

$$\operatorname{Var}(\hat{\beta}) = H = -\mathrm{E}\begin{bmatrix} h_{\hat{a},\hat{a}} & h_{\hat{a},\hat{b}} \\ h_{\hat{b},\hat{a}} & h_{\hat{b},\hat{b}} \end{bmatrix}^{-1} \tag{4.3.8}$$

where

$$\begin{aligned}
h_{\hat{a},\hat{a}} &= -\frac{n_0\tau^2}{\left[1-\theta_0-a\tau-\frac{b\tau^2}{2}\right]^2} - \frac{n_r\mu^2}{\left[a\mu-\frac{1}{2}b(\gamma-2\tau\mu)\right]^2} \\
&\quad -\frac{n_1(\tau-\mu)^2}{\left[\theta_0+a(\tau-\mu)+\frac{1}{2}b(\tau^2+\gamma-2\tau\mu)\right]^2} \\
h_{\hat{a},\hat{b}} &= -\frac{\frac{1}{2}n_0\tau^3}{\left[1-\theta_0-a\tau-\frac{b\tau^2}{2}\right]^2} + \frac{\frac{1}{2}n_r(\gamma-2\tau\mu)\mu}{\left[a\mu-\frac{1}{2}b(\gamma-2\tau\mu)\right]^2} \\
&\quad -\frac{\frac{1}{2}n_1(\tau-\mu)(\tau^2+\gamma-2\tau\mu)}{\left[\theta_0+a(\tau-\mu)+\frac{1}{2}b(\tau^2+\gamma-2\tau\mu)\right]^2} \\
h_{\hat{b},\hat{b}} &= -\frac{\frac{1}{4}n_0\tau^4}{\left[1-\theta_0-a\tau-\frac{b\tau^2}{2}\right]^2} - \frac{\frac{1}{4}n_r(\gamma-2\tau\mu)^2}{\left[a\mu-\frac{1}{2}b(\gamma-2\tau\mu)\right]^2} \\
&\quad -\frac{\frac{1}{4}n_1(\tau^2+\gamma-2\tau\mu)^2}{\left[\theta_0+a(\tau-\mu)+\frac{1}{2}b(\tau^2+\gamma-2\tau\mu)\right]^2}
\end{aligned}$$

where of course $h_{\hat{a},\hat{b}} = h_{\hat{b},\hat{a}}$.

If we replace 'a' and 'b' by their respective MLEs given in Eq (4.3.5) and Eq (4.3.6) respectively in Eq (4.3.8), it follows that the approximate variance of $\hat{a}$, variance of $\hat{b}$ and covariance between $\hat{a}$ and $\hat{b}$ are respectively:

$$\mathrm{var}(\hat{a}) = \frac{n_1 n_r \tau^4 + n_0 n_1(\gamma - 2\tau\mu)^2 + n_0 n_r(\tau^2 + \gamma - 2\tau\mu)^2}{\tau^2 n^3(\gamma - \tau\mu)^2} \tag{4.3.9}$$

$$\mathrm{var}(\hat{b}) = \frac{4[n_1(n_r\tau^2 + n_0\mu^2) + n_0 n_r(\tau - \mu)^2]}{\tau^2 n^3(\gamma - \tau\mu)^2} \tag{4.3.10}$$

$$\mathrm{cov}(\hat{a}, \hat{b}) = \frac{2[n_0 n_1 \mu(\gamma - 2\tau\mu) - n_r n_1 \tau^3 + n_0 n_r(\mu - \tau)(\tau^2 + \gamma - 2\tau\mu)]}{\tau^2 n^3(\gamma - \tau\mu)^2} \tag{4.3.11}$$

Finally, the estimated variance of HIV incidence,$\lambda(t)$, using the Delta Method, at time t will be

$$\mathrm{var}(\lambda\hat{(t)}) = \frac{1}{(1-\hat{\theta})^2}\mathrm{var}(\hat{f}) + \frac{\hat{f}^2}{(1-\hat{\theta})^4}\mathrm{var}(\hat{\theta}) + \frac{2\hat{f}}{(1-\hat{\theta})^3}\mathrm{cov}(\hat{f}, \hat{\theta}) \tag{4.3.12}$$

where

$$\begin{aligned}
\mathrm{var}(\hat{f}) &= \mathrm{var}(\hat{a}) + \tau^2\mathrm{var}(\hat{b}) + 2\tau\mathrm{cov}(\hat{a}, \hat{b}) \\
\mathrm{cov}(\hat{f}, \hat{\theta}) &= \tau\mathrm{var}(\hat{a}) + \frac{\tau^3}{2}\mathrm{var}(\hat{b}) + 1.5\tau^2\mathrm{cov}(\hat{a}, \hat{b}) \\
\mathrm{var}(\hat{\theta}) &= \tau^2\mathrm{var}(\hat{a}) + \frac{\tau^4\mathrm{var}(\hat{b})}{4} + \tau^3\mathrm{cov}(\hat{a}, \hat{b})
\end{aligned}$$

4.3.2 Incorporation of the false recent rate

The model we have proposed assumes that all persons who were non-reactive to the less sensitive assay eventually react to the sensitive assay. However, a proportion of

HIV infected subjects may remain non-reactive to the less sensitive assay indefinitely long after the sero-conversion period leading to overestimation of the estimated HIV incidence rate as discussed by (Wang and Lagakos, 2009; McWalter and Welte, 2010; Brookmeyer, 2010b; Hargrove et al., 2008; Novitsky et al., 2009; Karita et al., 2007). This proportion is called the "false recent rate". To incorporate the false recent rate, we extend the approach of Wang and Lagakos (2009).

The revised likelihood function and parameter estimation

Let p denote the proportion of subjects who will become reactive at some point after seroconversion as in Wang and Lagakos (2009). Then the false recent rate will be $q = 1 - p$ which denotes the proportion of subjects who remain negative on the less sensitive assay indefinitely. Then the log-likelihood incorporating the false recent rate will be:

$$\begin{aligned} \ell(a,b) &= n_0 \log(1 - \theta_0 - a\tau - \frac{b\tau^2}{2}) \\ &\quad + n_r \log[(a\mu - \frac{1}{2}b(\gamma - 2\tau\mu))p + (\theta_0 + a\tau + \frac{b\tau^2}{2})(1-p)] \\ &\quad + n_1 \log[\theta_0 + a(\tau - \mu) + \frac{1}{2}b(\tau^2 + \gamma - 2\tau\mu)] \qquad (4.3.13) \end{aligned}$$

The MLEs of 'a' and 'b', given by $\hat{a}$ and $\hat{b}$, when the false recent rate has been incorporated, can be obtained by joint maximization of Eq (4.3.13) with respect to 'a' and 'b'. Thus the corresponding MLEs, $\hat{a}$ and $\hat{b}$ are respectively:

$$\begin{aligned}\hat{a} &= -\frac{-2np\mu\tau\theta_0 + 2(n_1+n_r)p\mu\tau}{np\tau(\gamma-\tau\mu)} \\ &\quad -\frac{np\gamma\theta_0 - (n_1+n_r)p\gamma - (n_1+n_r)\tau^2 p + n_1\tau^2}{np\tau(\gamma-\tau\mu)} \qquad (4.3.14) \\ \hat{b} &= -\frac{2[np\mu\theta_0 + (n_1+n_r)p\tau - p\mu(n_1+n_r) - n_1\tau]}{np\tau(\gamma-\tau\mu)} \qquad (4.3.15)\end{aligned}$$

Hence the MLE of HIV incidence at the time t, $\hat{\lambda}(t)$, which is now a function of p or equivalently $q = 1 - p$ will be

$$\lambda(\hat{t}) = \frac{\hat{a} + \hat{b}\tau}{1 - \hat{\theta}} \qquad (4.3.16)$$

where $\hat{\theta} = \theta_0 + \hat{a}\tau + \frac{1}{2}\hat{b}\tau^2$ by the invariance property of MLEs and $\hat{a}$ and $\hat{b}$ are as provided in Eq (4.3.14) and Eq (4.3.15) respectively.

The corresponding variances and covariance are:

$$\begin{aligned}\mathrm{var}(\hat{a}) &= \frac{n_0 n_1(-2p\mu\tau + p\gamma + p\tau^2 - \tau^2)^2}{n^3\tau^2p^2(\gamma-\tau\mu)^2} \\ &\quad +\frac{p^2 n_0 n_r(\tau^2+\gamma-2\tau\mu)^2 + n_1 n_r\tau^4}{n^3\tau^2p^2(\gamma-\tau\mu)^2} \qquad (4.3.17) \\ \mathrm{var}(\hat{b}) &= \frac{4[n_0n_1(p\mu+\tau-p\tau)^2 + n_0n_rp^2(\tau-\mu)^2 + n_1n_r\tau^2]}{n^3\tau^2p^2(\gamma-\tau\mu)^2} \qquad (4.3.18)\end{aligned}$$

$$\begin{aligned}\mathrm{cov}(\hat{a},\hat{b}) &= \frac{2[n_0n_1(p\mu+\tau-p\tau)(-2p\mu\tau+p\gamma+p\tau^2-\tau^2)]}{n^3\tau^2p^2(\gamma-\tau\mu)^2} \\ &\quad +\frac{[n_0n_rp^2(\mu-\tau)(\tau^2+\gamma-2\tau\mu) - n_1n_r\tau^3]}{n^3\tau^2p^2(\gamma-\tau\mu)^2} \qquad (4.3.19)\end{aligned}$$

Finally, the expression for the estimated variance of HIV incidence,$\hat{\lambda}(t)$, is the same as the one provided in in Eq 4.3.12.

4.3.3 Testing For b=0

When $b = 0$ the log likelihood function in Eq 4.3.4 reduces to

$$\ell(a) = n_0\log(1-\theta_0-a\tau)+n_r\log(a\mu)+n_1\log(\theta_0+a\tau-a\mu) \tag{4.3.20}$$

Then,

$$\hat{\lambda(t)} = \frac{-d+\sqrt{d^2-4ac}}{2a} \tag{4.3.21}$$

where

$$\begin{aligned} a &= n_0\tau\left[\tau-(1-\theta_0)\mu\right] \\ d &= (n_0+n_1)\tau\theta_0-(n_r+n_1)\left[\tau-(1-\theta_0)\mu\right] \\ c &= -n_r\theta_0 \end{aligned}$$

Eq (4.3.21) is the estimator of HIV incidence rate proposed by Gabaitiri et al. (2013). So the constant incidence estimator including information on previous prevalence proposed by Gabaitiri et al. (2013) is a special case of the proposed linear density estimator in the current paper.

In practice, one may want to carry out a statistical test of the goodness-of-fit between two models. Basically a relatively more complex model is compared to a simpler model to see if it fits a particular dataset significantly better. If so, then the additional parameters of the more complex model are often used in subsequent analyses otherwise a simple model is used. The standard procedure is to use the likelihood ratio test for nested models. In this setting, the objective is to test if ‘b’ is significantly different from zero. The Wald test can also be used. The two tests are asymptotically equivalent. Under the reduced model (or the null hypothesis that $b = 0$), the test statistic is given by

$$Z^2 = \left[\frac{\hat{b}}{\sqrt{\text{var}(\hat{b})}}\right]^2 \overset{app}{\sim} \chi_1^2 \text{ for large samples.}$$

Where $\hat{b}$ and $\text{var}(\hat{b})$, under the reduced model, are respectively:

$$\begin{aligned} \hat{b} &= -\frac{2[n_r\tau - (n_1 + n_r)\mu + n\mu\theta_0]}{n\tau(\gamma - \tau\mu)} \\ \text{var}(\hat{b}) &= \frac{4[n_1(n_r\tau^2 + n_0\mu^2) + n_0n_r(\tau - \mu)^2]}{\tau^2 n^3(\gamma - \tau\mu)^2} \end{aligned}$$

That is, for large samples, the distribution of the test statistic is approximately chi-square with one degree of freedom.

As with all hypothesis tests (or significance tests), we can easily use the p-value (the probability of getting a result as extreme as that observed for the test statistic when the null hypothesis is true) to make decisions at common significance cut-points. For instance, if the p-value is below 0.05 we reject the null hypothesis that $b = 0$ at the 5% level of significance. The significance level represents the chance that the reduced model (when $b = 0$) is rejected when it is actually the correct model for the data.

4.4 Simulation Study

We performed a simulation study to evaluate the performance of the proposed estimator of HIV incidence using 1000 simulations. We generated a sample of size $n = 4000$. The data was simulated from a multinomial distribution with parameters (for the unadjusted model):

$$\begin{aligned}
\pi_1(t) &= 1 - \theta_0 - a\tau - \frac{b\tau^2}{2} \\
\pi_2(t) &= a\mu - \frac{1}{2}b(\gamma - 2\tau\mu) \\
\pi_3(t) &= \theta_0 + a(\tau - \mu) + \frac{1}{2}b(\tau^2 + \gamma - 2\tau\mu)
\end{aligned}$$

For the adjusted model, we have

$$\begin{aligned}
\pi_1(t) &= 1 - \theta_0 - a\tau - \frac{b\tau^2}{2} \\
\pi_2(t) &= a\mu - \frac{1}{2}b(\gamma - 2\tau\mu)p + (\theta_0 + a\tau + \frac{1}{2}b\tau^2)(1-p) \\
\pi_3(t) &= \theta_0 + a(\tau - \mu) + \frac{1}{2}b(\tau^2 + \gamma - 2\tau\mu)
\end{aligned}$$

For each simulation, we calculate the average point estimates of a_{true}, b_{true} and I_{true}, the true values for a, b and λ (HIV incidence rate), respectively based on the proposed model. The average variance (standard errors) of these estimates are obtained from the observed Fisher information. We also obtain the empirical estimates of the variance from the 1000 simulated point estimates of incidence.

The results of the simulation study are presented in Table 4.2 for $\theta_0 = 0.10, 0.171, 0.20$. For each fixed values of θ_0 we assess how the point estimates and their standard errors will change when $\tau = 0.5$ and 1.0 years. Also presented are the values of the coverage probabilities. That is, the proportion of times out of 1000 simulations the 95% confidence interval covers the true values of a, b and λ. We assumed $\mu = E(L_2) =$ 155 days and $Var(L_2) = 155^2$ $(days)^2$ as under an exponential model. Table 4.2 shows the results for the unadjusted model.

Table 4.2: *Results of a simulation study with* $N = 4000$ *and* $a_{true} = 0.01$, $b_{true} = 0.04$ *and* (I_{true})*. Note that* I_{true} *is given in parenthesis.* $\hat{a}$, $\hat{b}$, *and* $\hat{\lambda}$ *represents the average estimates of the parameters.* $\hat{SE}(\hat{a})$, $\hat{SE}(\hat{b})$ *and* $\hat{SE}(\hat{\lambda})$ *represents the estimated standard errors of the parameter estimates.* $\hat{ESE}(\hat{a})$, $\hat{ESE}(\hat{a})$ *and* $\hat{ESE}(\hat{\lambda})$ *represents the estimated empirical standard errors. All are based on 1000 simulations.* 95%*CV denotes the proportion of experiments in which* a_{true}, b_{true} *and* I_{true} *is contained in the nominal* 95% *confidence interval.*

Simulation Results with $a_{true} = 0.01$, $b_{true} = 0.04$ and (I_{true})				
		Previous HIV Prevalence		
τ(in years)	Estimates	$\theta_0 = 0.10$	$\theta_0 = 0.171$	$\theta_0 = 0.20$
0.5	$\hat{a}$	0.0099	0.0099	0.0100
	$\hat{\text{SE}}(\hat{a})$	0.0052	0.0060	0.0063
	$\text{E}\hat{\text{S}}\text{E}(\hat{a})$	0.0052	0.0061	0.0062
	95%CV	94.1	94.7	95.7
	$\hat{b}$	0.0402	0.0389	0.0393
	$\hat{\text{SE}}(\hat{b})$	0.0555	0.0690	0.0731
	$\text{E}\hat{\text{S}}\text{E}(\hat{b})$	0.0553	0.0694	0.0722
	95%CV	95.6	96.0	95.8
	$\hat{\lambda}$	0.0338 (0.0337)	0.0361 (0.0366)	0.0378 (0.0380)
	$\hat{\text{SE}}(\hat{\lambda})$	0.0266	0.0360	0.0395
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0266	0.0362	0.0392
	95%CV	95.7	95.7	95.3
1	$\hat{a}$	0.0117	0.0128	0.0118
	$\hat{\text{SE}}(\hat{a})$	0.0420	0.0502	0.0528
	$\text{E}\hat{\text{S}}\text{E}(\hat{a})$	0.0427	0.0506	0.0531
	95%CV	93.8	94.2	93.9
	$\hat{b}$	0.0371	0.0351	0.0369
	$\hat{\text{SE}}(\hat{b})$	0.0764	0.0902	0.0946
	$\text{E}\hat{\text{S}}\text{E}(\hat{b})$	0.0779	0.0912	0.0955
	95%CV	93.5	94.3	94.1
	$\hat{\lambda}$	0.0560 (0.0575)	0.0598 (0.0626)	0.0630 (0.0649)
	$\hat{\text{SE}}(\hat{\lambda})$	0.0398	0.0502	0.0528
	$\text{E}\hat{\text{S}}\text{E}(\hat{\lambda})$	0.0407	0.0510	0.0551
	95%CV	93.6	93.3	93.6

Table 4.3 and Table 4.4 shows the results for the adjusted model. The results presented are for $p = 0.95$ and $p = 0.98$ respectively.

Under these settings, for both the unadjusted and adjusted estimator, the average estimates of all the parameters namely a, b and $\lambda = I_{true}$ are close to the true values reflecting their validity. The standard errors of the unadjusted estimators are consistently lower than those of the adjusted estimators. It is also assuring to note that the estimated standard errors and the empirical estimates are closer to each other for all the estimates.

Also noticeable is that when the previous HIV prevalence increases, then the estimated standard errors increases but the parameter estimates are least affected. The coverage probabilities are closer to the nominal value. The estimated standard errors decreases as the proportion of non-progressors (false recent rate) decreases.

4.5 Application to BAIS III data set

We use the data from BAIS III of 2008 to illustrate the use of the proposed estimator. We assumed the mean window period, $\mu = E(L_2) = 155$ days and $Var(L_2) = 155^2$ $(days)^2$. We assume $\tau = 4$ years as it would be the case in most cross sectional surveys. The results are presented in Table 4.6 and Table 4.7.

In Botswana, between 2004 and 2008, overall the national HIV prevalence increased from 17.1% to 17.6% CSO (2008) for individuals aged 1.5 years and above. For purposes of illustration we used simple extrapolation method to estimate the previous prevalence for 2004. The values of θ_0 are provided in Table 4.5.

Table 4.3: *Results of a simulation study with the incorporation of the False Recent Rate (FRR) given by $q = 1 - p$ with $N = 3000$ and $a_{true} = 0.01$, $b_{true} = 0.04$ and (I_{true}). Note that I_{true} is given in parenthesis. $\hat{a}$, $\hat{b}$, and $\hat{\lambda}$ represents the average estimates of the parameters. $\hat{SE}(\hat{a})$, $\hat{SE}(\hat{b})$ and $\hat{SE}(\hat{\lambda})$ represents the estimated standard errors of the parameter estimates. $E\hat{SE}(\hat{a})$, $E\hat{SE}(\hat{a})$ and $E\hat{SE}(\hat{\lambda})$ represents the empirical parameter estimates. All are based on 1000 simulations. 95%CV denotes the proportion of experiments in which a_{true}, b_{true} and I_{true} is contained in the nominal* 95% *confidence interval. Here P=0.95.*

Simulation Results incorporating FRR with $a_{true} = 0.01$, $b_{true} = 0.04$ and (I_{true})				
			$p = 0.95$	
			Previous HIV Prevalence	
τ(in years)	Estimates	$\theta_0 = 0.10$	$\theta_0 = 0.171$	$\theta_0 = 0.20$
0.5	$\hat{a}$	0.0109	0.0105	0.0101
	$\hat{\text{SE}}(\hat{a})$	0.0065	0.0079	0.0084
	$\text{E}\hat{\text{SE}}(\hat{a})$	0.0067	0.0082	0.0085
	95%CV	93.7	93.6	94.1
	$\hat{b}$	0.0387	0.0398	0.0427
	$\hat{\text{SE}}(\hat{b})$	0.0577	0.0720	0.0765
	$\text{E}\hat{\text{SE}}(\hat{b})$	0.0592	0.0717	0.0764
	95%CV	94.4	94.9	94.3
	$\hat{\lambda}$	0.0341 (0.0337)	0.0374 (0.0366)	0.0401 (0.0380)
	$\hat{\text{SE}}(\hat{\lambda})$	0.0270	0.0365	0.0402
	$\text{E}\hat{\text{SE}}(\hat{\lambda})$	0.0276	0.0362	0.0400
	95%CV	94.3	95.2	94.9
1	$\hat{a}$	0.0096	0.0110	0.0084
	$\hat{\text{SE}}(\hat{a})$	0.0463	0.0559	0.0590
	$\text{E}\hat{\text{SE}}(\hat{a})$	0.0458	0.0542	0.0589
	95%CV	94.6	94.7	94.5
	$\hat{b}$	0.0416	0.0389	0.0436
	$\hat{\text{SE}}(\hat{b})$	0.0857	0.1028	0.1083
	$\text{E}\hat{\text{SE}}(\hat{b})$	0.0848	0.0999	0.1083
	95%CV	94.8	94.2	94.0
	$\hat{\lambda}$	0.0588 (0.0575)	0.0623 (0.0625)	0.0673 (0.0649)
	$\hat{\text{SE}}(\hat{\lambda})$	0.0456	0.0589	0.0643
	$\text{E}\hat{\text{SE}}(\hat{\lambda})$	0.0452	0.0575	0.0644
	95%CV	94.3	94.1	94.2

Table 4.4: *Results of a simulation study with the incorporation of the False Recent Rate (FRR) given by $q = 1 - p$ with $N = 3000$ and $a_{true} = 0.01$, $b_{true} = 0.04$ and (I_{true}). Note that I_{true} is given in parenthesis. $\hat{a}$, $\hat{b}$, and $\hat{\lambda}$ represents the average estimates of the parameters. $\hat{SE}(\hat{a})$, $\hat{SE}(\hat{b})$ and $\hat{SE}(\hat{\lambda})$ represents the estimated standard errors of the parameter estimates. $E\hat{SE}(\hat{a})$, $E\hat{SE}(\hat{a})$ and $E\hat{SE}(\hat{\lambda})$ represents the empirical parameter estimates. All are based on 1000 simulations. 95%CV denotes the proportion of experiments in which a_{true}, b_{true} and I_{true} is contained in the nominal 95% confidence interval. Here P=0.98.*

Simulation Results incorporating FRR with $a_{true} = 0.01$, $b_{true} = 0.04$ and (I_{true})				
			$p = 0.98$	
			Previous HIV Prevalence	
τ(in years)	Estimates	$\theta_0 = 0.10$	$\theta_0 = 0.171$	$\theta_0 = 0.20$
0.5	$\hat{a}$	0.0102	0.0102	0.0103
	$\hat{\text{SE}}(\hat{a})$	0.0057	0.0069	0.0072
	$\text{E}\hat{\text{SE}}(\hat{a})$	0.0058	0.0068	0.0074
	95%CV	94.6	94.7	94.3
	$\hat{b}$	0.0395	0.0434	0.0409
	$\hat{\text{SE}}(\hat{b})$	0.0563	0.0702	0.0744
	$\text{E}\hat{\text{SE}}(\hat{b})$	0.0583	0.0694	0.0759
	95%CV	94.9	95.0	94.5
	$\hat{\lambda}$	0.0338 (0.0337)	0.0392 (0.0366)	0.0393 (0.0380)
	$\hat{\text{SE}}(\hat{\lambda})$	0.0268	0.0363	0.0398
	$\text{E}\hat{\text{SE}}(\hat{\lambda})$	0.0278	0.0359	0.0406
	95%CV	94.8	94.8	94.3
1	$\hat{a}$	0.0126	0.0130	0.0119
	$\hat{\text{SE}}(\hat{a})$	0.0438	0.0525	0.0552
	$\text{E}\hat{\text{SE}}(\hat{a})$	0.0445	0.0520	0.0553
	95%CV	93.3	94.1	93.6
	$\hat{b}$	0.0357	0.0347	0.0367
	$\hat{\text{SE}}(\hat{b})$	0.0801	0.0952	0.1001
	$\text{E}\hat{\text{SE}}(\hat{b})$	0.0817	0.0944	0.1006
	95%CV	93.0	94.5	93.7
	$\hat{\lambda}$	0.0554 (0.0575)	0.0596 (0.0625)	0.0628 (0.0575)
	$\hat{\text{SE}}(\hat{\lambda})$	0.0421	0.0537	0.0584
	$\text{E}\hat{\text{SE}}(\hat{\lambda})$	0.0431	0.0532	0.0589
	95%CV	93.7	94.6	93.8

Table 4.5: Values of θ_0

Demographics		Sample n (n_0, n_r, n_1)	$\theta_0 \approx \frac{n_1+n_r}{n} - 0.005$
Sex	Male	6516 (5603, 55, 858)	0.135
	Female	7775 (6220, 94, 1461)	0.195
Age (y)	≤ 4	872 (854, 5, 13)	0.016
	5-14	3382 (3242, 18, 122)	0.036
	15-19	1449 (1395, 5, 49)	0.032
	20-29	2967 (2401, 43, 523)	0.186
	30-39	2142 (1287, 33, 822)	0.394
	40-49	1429 (928, 27, 474)	0.346
	50+	2050 (1716, 18, 316)	0.158
Overall	All groups	14291 (11823, 149, 2319)	0.168

Table 4.6: *Estimated HIV incidence rate by gender and age under the linear incidence density model with $p = 1$.*

			$p = 1$	
Demographics		$\hat{a}$ (95% CI)	$\hat{b}$ (95% CI)	$\hat{\lambda}$ (95% CI)
Sex	Male	-0.022 (-0.030, -0.015)	0.012 (0.009, 0.015)	2.89 (2.13, 3.66)
	Female	-0.033 (-0.041, -0.025)	0.017 (0.014,0.021)	4.48 (3.57, 5.38)
Age	≤ 4	-0.014 (-0.027, -0.000)	0.007 (0.000, 0.014)	1.63 (0.12, 3.13)
	5-14	-0.013 (-0.020, -0.006)	0.007 (0.004, 0.011)	1.62 (0.87, 2.37)
	15-19	-0.007 (-0.016, 0.002)	0.004 (0.000, 0.009)	1.03 (0.11, 1.96)
	20-29	-0.041 (-0.054, -0.027)	0.021 (0.015, 0.027)	5.31 (3.73, 6.90)
	30-39	-0.043 (-0.061, -0.025)	0.022 (0.014, 0.030)	7.61 (5.01, 10.21)
	40-49	-0.054 (-0.077, -0.031)	0.028 (0.017, 0.038)	8.65 (5.39, 11.91)
	50+	-0.023 (-0.037, -0.010)	0.012 (0.006, 0.018)	3.10 (1.67, 4.53)
Overall	All groups	-0.029 (-0.034, -0.023)	0.015 (0.012, 0.017)	3.73 (3.13, 4.33)

Table 4.7: *Estimated HIV incidence rate by gender and age under the linear incidence density model with* $p = 0.98$.

			$p = 0.98$	
Demographics		$\hat{a}$ (95% CI)	$\hat{b}$ (95% CI)	$\hat{\lambda}$ (95% CI)
Sex	Male	-0.014 (-0.022,-0.007)	0.008 (0.004, 0.011)	1.96 (1.19, 2.73)
	Female	-0.022 (-0.030, -0.014)	0.012 (0.008, 0.015)	3.04 (2.13, 3.95)
Age	≤ 4	-0.014 (-0.027, -0.001)	0.008 (0.001, 0.014)	1.63 (0.13, 3.14)
	5-14	-0.011 (-0.018, -0.003)	0.006 (0.002, 0.010)	1.39 (0.64, 2.15)
	15-19	-0.005 (-0.015, 0.004)	0.003 (-0.001, 0.008)	0.82 (0.00, 1.75)
	20-29	-0.030 (-0.044, -0.016)	0.016 (0.009, 0.022)	3.99 (2.40, 5.58)
	30-39	-0.020 (-0.039, -0.001)	0.011 (0.002, 0.019)	3.71 (1.09, 6.33)
	40-49	-0.034 (-0.058, -0.010)	0.017 (0.006, 0.028)	5.53 (2.26, 8.81)
	50+	-0.014 (-0.028, 0.000)	0.008 (0.001, 0.014)	1.97 (0.53, 3.41)
Overall	All groups	-0.019 (-0.024, -0.013)	0.010 (0.007, 0.012)	2.53 (1.93, 3.13)

Table 4.8: *Estimated HIV incidence rate by gender and age under the linear incidence density model with* $p = 0.95$.

			$p = 0.95$	
Demographics		$\hat{a}$ (95% CI)	$\hat{b}$ (95% CI)	$\hat{\lambda}$ (95% CI)
Sex	Male	-0.002 (-0.010,0.007)	0.001 (-0.002, 0.005)	0.49 (-0.31, 1.28)
	Female	-0.004 (-0.013, 0.005)	0.002 (-0.001, 0.006)	0.78 (-0.15, 1.71)
Age	≤ 4	-0.012 (-0.026, 0.001)	0.007 (0.000, 0.014)	1.48 (-0.04, 2.29)
	5-14	-0.007 (-0.015, 0.000)	0.004 (0.001, 0.008)	1.03 (0.26, 1.80)
	15-19	-0.002 (-0.012, 0.008)	0.002 (-0.003, 0.006)	0.48 (-0.46, 1.43)
	20-29	-0.013 (-0.028, 0.002)	0.007 (0.000, 0.014)	1.89 (0.27, 3.51)
	30-39	0.017 (-0.003, 0.038)	-0.008 (-0.017, 0.001)	-2.44 (-5.18, 0.29)
	40-49	-0.002 (-0.027, 0.024)	0.001 (-0.010, 0.013)	0.61 (-2.75, 3.98)
	50+	0.001 (-0.014, 0.016)	0.000 (-0.006, 0.007)	0.20 (-1.29, 1.69)
Overall	All groups	-0.003 (-0.009, 0.003)	0.002 (-0.001, 0.005)	0.64 (0.03, 1.26)

Table 4.6, Table 4.7 and Table 4.8 show the parameter estimates $\hat{a}$, $\hat{b}$ and $\hat{\lambda}$ (the incidence rate), together with their 95% confidence limits under three scenarios when $p = 1$, $p = 0.98$ and $p = 0.95$ respectively. The estimates are also structured according to sex (male and female) and age groups. An overall estimate for each parameter estimates is also given in the last row of the table. Noticeably, as expected, the incidence estimate for $p = 0.95$ and $p = 0.98$ are smaller than those for $p = 1$ for both sexes, all ages and overall while the estimates for $p = 0.95$ are smaller than those for $p = 0.98$.

We also tested the null hypothesis that $b = 0$ using the Wald test. That is, we tested whether or not the linear incidence density assumption was reasonable for this data under these settings as opposed to the constant incidence density assumption (the previous prevalence is assumed known). The results shows that, for $p = 1$ and $p = 0.98$, we reject the constant incidence density assumption and conclude that the linear incidence assumption is reasonable under these settings ($p - values < 0.01$). But when $p = 0.95$ the constant density assumption seems to be reasonable for this data ($p - value = 0.131$). Tables 4.6 and 4.7 also provide a very interesting information about the linear incidence density function. That is, the slopes in the two cases increase and reach a maximum at age class 40 - 49 and then drops. This is an indication that the force of infection or the intensity of the disease is more pronounced in the younger age groups then reaches a maximum at age group 40 -49 and less pronounced in the older age groups beyond age 50 and above. Further, in terms of gender, the slopes indicate that the force of infection is higher in females than males. But 4.8 does not show any of these patterns possibly because the linear incidence density assumption is not reasonable for the setting where $p = 0.95$.

4.6 Discussion

We proposed a method for estimating HIV incidence rate when the incidence density is assumed to be linear. The proposed model works well under the current setting and we were also able to compare this model with one under constant incidence assumption. The methods can be used to estimate incidence of any other disease provided similar data as the one used here is available.

The most important limitation of the cross sectional approach is that $\mu = E(L_2)$ and $Var(L_2)$ (in the case of the proposed estimator) are not known reliably and this will have an impact on the accuracy of the estimated incidence. Methods for accurate estimation of the two have been proposed Wang and Lagakos (2010); Claggett et al. (2012) though they have not been used in practical settings. We also note that the assumed three state model does not take mortality into account and thus the state prevalence probabilities that we developed may not be estimated correctly. This is because, as noted by Balasubramanian and Lagakos (2010), the risk of death is higher for subjects in the late stage of HIV. However, the introduction of anti-retroviral therapy (ART) has improved survival amongst individuals with late stage of HIV and thus lowering HIV related mortality. Further research is needed for on this area particulary on the effect of non-HIV (and HIV) related mortality on incidence estimation.

Although ART improves survival amongst individuals with late stage of HIV, it may also complicate the estimation of incidence since it tends to increase $1 - p$, the proportion of non-progressors. Two approaches have been proposed in the literature for handling individuals on ART. One is to identify and completely remove all individuals on ART from the cross sectional samples as suggested in McDougal et al. (2006). The danger of removing all subjects on ARTs is that the sample is now modified and this could lead to bias in the estimated incidence. As noted by Claggett et al. (2012), this can also have an impact on the resulting prevalence estimate if ART use is common in the study sites. Another approach is that all subjects on treatment can be assumed to be in the non-recent state because treatment initiation is often long after sero-conversion as suggested by Wang and Lagakos (2009). Further research is needed in this area.

Chapter 5

Estimation of HIV Incidence from a Cross-Sectional Sample with Missing Data

SUMMARY. The use of novel biomarkers to identify recent infections in cross sectional samples offers a lot of promise in the estimation of HIV incidence. Identification of new infections is done through a dual antibody testing system in which specimens are tested with a sensitive HIV antibody assay and those that are positive are then tested with a less sensitive assay. However, for some reasons, known or unknown, a proportion of specimens that tested positive on the standard antibody assay may not be tested by the less sensitive assay resulting in missing data. This means the standard method used to estimate incidence may lead to biased estimates and their standard errors. In this paper, maximum likelihood method for estimating incidence when some specimens that tested positive on the standard antibody test are missing is described. The proposed method is illustrated using data from the Botswana AIDS Impact (BAIS) III of 2008.

KEYWORDS: HIV incidence, BED assay, missing data, Maximum likelihood

5.1 Introduction

Estimation of human immunodeficiency virus (HIV) incidence is necessary for assessing the impact of programs and for monitoring the spread of HIV infection. A direct, though expensive approach way to estimate incidence, is through longitudinal cohort studies which may also need a long time to carry out. The use of novel biomarkers to identify recent infections in cross sectional samples offers a lot of promise in the estimation of HIV incidence and circumvent some of the problems associated with the cohort approach. Identification of new infections is done through a dual antibody testing system in which specimens are tested with a sensitive HIV antibody assay (typically ELISA) and then by a less sensitive assay Janssen et al. (1998). Specimens testing negative on the sensitive HIV antibody assay are considered negative or not yet seroconverted while specimens testing positive on the sensitive antibody assay are tested again using a less sensitive assay, typically BED-EIA (commonly referred to as BED assay). Specimens testing negative on the less sensitive assay are considered to be recent infections, while those testing positive are considered long standing or established infections (Janssen et al., 1998; Parekh et al., 2002; Parekh and McDougal, 2005). However, for some reasons, known or unknown, a proportion of specimens that tested positive on the sensitive antibody assay are not tested by the less sensitive assay resulting in missing data. This is particularly common in population based cross-sectional surveys. This means that the already developed standard dual testing methods used to estimate incidence will lead to biased incidence estimates and their standard errors. Ad hoc methods for incorporating missing data assumes that the missing data is missing completely at random (MCAR). That is, the events that lead to any particular data-item being missing are independent of both the observable variables and of unobservable parameters of interest (Little and Rubin, 2002). However MCAR is a too restrictive assumption and difficult to justify in reality. WHO (2011) describes how one can adjust for missing data under this assumption. Another way of handling missing data was introduced by Chu and Cole (2006) through the method of maximum likelihood under the missing at random (MAR) assumption as described by Little and Rubin (2002).

In this paper, we extend the idea of Chu and Cole (2006) to describe how one can incorporate missing data in the incidence estimation formula proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009) under MAR assumption. In Section 5.2 we describe the maximum likelihood method for incorporating the missing data under MAR. In Section 5.3, we illustrate the use of the method using data from the Botswana AIDS Impact (BAIS) III of 2008. We conclude with discussion in Section 5.4.

5.2 The extended model incorporating missing data

We consider the 3-state longitudinal disease progression model for the natural history of HIV/AIDS and classification of subjects is by a diagnostic test such that state 1 represents the pre-seroconversion period. We denote state 1 by S_1 because it is corresponding to the period in which an individual is either uninfected or is infected but has not yet seroconverted. Thus strictly speaking state 1 can be subdivided into two finer sub-states as in Balasubramanian and Lagakos (2010). State 2, denoted by S_2, represents the "recent infection" state where HIV antibodies are detectable by a sensitive diagnostic test but not yet through a less sensitive test such as BED assay. Finally state 3, denoted by S_3, represents the "non-recent infection" state in which HIV antibodies are detectable by a less sensitive test (and sensitive test). For simplicity and tractability, we assume the incidence density is constant as in Balasubramanian and Lagakos (2010). We note that Gabaitiri et al. (2012) did relax the assumption of a constant incidence density but the focus of the current paper is first to deal with the problem of missing outcomes.

Suppose n subjects are randomly selected from the population at some calendar time t. Let n_0 be the number of subjects in S_1, n_r be the number of subjects in S_2 and finally n_1 be the number of subjects in S_3 such that $n = n_0 + n_r + n_1$.

Using similar arguments as in Balasubramanian and Lagakos (2010), it can be shown that the prevalence probabilities for the 3-states at time t are

$$\begin{aligned}\pi_1(t) &= 1-\theta \\ \pi_2(t) &= f\mu \\ \pi_3(t) &= \theta - f\mu\end{aligned}$$

The objective is to estimate HIV incidence rate,

$$\lambda = \frac{f}{1-\theta}$$

where f is the incidence density function and for the purpose of the current analysis, is assumed to be constant as in Balasubramanian and Lagakos (2010), and θ is the HIV prevalence at time t.

It follows that the trinomial log-likelihood function for this setting (ignoring the constant term) is

$$\ell(f,\theta) = n_0\text{log}[1-\theta] + n_r\text{log}[f\mu] + n_1\text{log}[\theta - f\mu] \tag{5.2.1}$$

where μ, commonly known as the mean window period, is assumed known. To be precise, the parameter μ is the mean residence time in S_2. That is, if we let T_2 be a random variable denoting the time of stay in S_2 then $\mu = E(T_2)$.

Since

$$f = \lambda(1-\theta)$$

then Eq (5.2.1) can be rewritten as

$$\ell(\lambda,\theta) = n_0\text{log}[1-\theta] + n_r\text{log}[\lambda(1-\theta)\mu] + n_1\text{log}[\theta - \lambda(1-\theta)\mu] \tag{5.2.2}$$

If data is missing because a proportion of specimens (n_m out of $n - n_0$) that tested positive on the sensitive antibody assay are not tested by the less sensitive assay, then the log-likelihood under the MAR assumption is

$$\ell(\lambda,\theta) = n_0\text{log}[1-\theta] + n_m\text{log}\theta + n_r\text{log}[\lambda(1-\theta)\mu] + n_1\text{log}[\theta - \lambda(1-\theta)\mu] \tag{5.2.3}$$

where n_m is number of specimens with missing BED assay results. Note that n_m is the number of missing specimens that tested positive on the sensitive test and they are part of prevalent cases (new and long-standing cases) hence the term $n_m\text{log}\theta$, where θ is the HIV prevalence at time t. The MLEs of λ and θ denoted by $\hat{\lambda}$ and $\hat{\theta}$ are, respectively

$$\begin{aligned}
\hat{\lambda} &= \frac{(n_1+n_r+n_m)n_r}{(n_r+n_1)n_0\mu} \\
&= \frac{n_r}{\left[\frac{n_1+n_r}{n_1+n_r+n_m}\right]n_0\mu} \\
&= \frac{n_r}{n_0^*\mu} && (5.2.4) \\
\hat{\theta} &= \frac{n_1+n_r+n_m}{n} && (5.2.5)
\end{aligned}$$

where $n = n_0+n_r+n_1+n_m$, $n_0^* = n_0\left[\frac{n_1+n_r}{n_1+n_r+n_m}\right]$. The log-likelihood in Eq (5.2.3) can be modified to take into account proportion of individuals (p) who remain negative on the less-sensitive assay indefinitely as proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009) to give

$$\begin{aligned}
\ell(\lambda_p,\theta) &= n_0\text{log}[1-\theta]+n_m\text{log}\theta+n_r\text{log}[\lambda_p(1-\theta)p\mu \\
&\quad +(1-p)\theta]+n_1\text{log}[p(\theta-\lambda_p(1-\theta)\mu)] && (5.2.6)
\end{aligned}$$

The MLEs of λ_p and θ, $\hat{\lambda}_p$ and $\hat{\theta}$ are, respectively

$$\begin{aligned}
\hat{\lambda_p} &= \frac{(n_1+n_r+n_m)(n_rp-(1-p)n_1)}{(n_r+n_1)n_0p\mu} \\
&= \frac{n_rp-(1-p)n_1}{n_0^*p\mu} && (5.2.7) \\
\hat{\theta} &= \frac{n_1+n_r+n_m}{n} && (5.2.8)
\end{aligned}$$

where $n = n_0+n_r+n_1+n_m$, $n_0^* = n_0\left[\frac{n_1+n_r}{n_1+n_r+n_m}\right]$.

Hence standard formulae for estimating standard errors such as the ones proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009) can be used to calculate the confidence interval for $\hat{\lambda}_p$ with n_0 replaced by n_0^*. The variance estimation formula for the estimated incidence rate $(\hat{\lambda}_p)$ proposed by Wang and Lagakos (2009) when p and the mean window period, μ, are assumed to be known is

$$\text{var}(\hat{\lambda}_p) \approx \frac{n_r}{n_0^2 p \mu^2} \tag{5.2.9}$$

Hence for the proposed estimator with missing data, we replace n_0 by n_0^*. The estimated variance of $\hat{\theta}$, obtained from the inverse of the matrix of negative second derivatives, with θ replaced by its MLE, $\hat{\theta}$, is

$$\text{var}(\hat{\theta}) = \frac{n_0(n_r + n_1 + n_m)}{n^3} \tag{5.2.10}$$

where $n = n_0 + n_r + n_1 + n_m$.

5.3 Application to BAIS III data set

Briefly, all persons aged 18 months and above were eligible for HIV testing and the target sample size was $n = 21,414$. But only 67% (sample size=14,351) provided blood specimen for HIV Testing. Of the 14,351 subjects who tested, 2521 tested HIV positive, $n_0 = 11,823$ tested HIV negative while 7 results were indeterminate. Out of the 2521 subjects who tested positive for HIV, there were $n_r = 149$ recent infections and $n_1 = 2319$ long standing infections. For the remaining subjects, $n_{others} = n_m = 53$ the specimens were reported missing (for 27 specimens the box was not found and for other 26 the blood samples were finished). The question is, what is the contribution of the 53 missing sample in the estimation of the HIV incidence? In fact there are two extremes; either they are all recent infections making the number in S_2 equal to 202 or they are all long standing infections making the number in S_3 equal to 2372. Obviously the most convincing and a flexible scenario is that of intermediate numbers, say $y*$, is assumed to be in S_2 and $n_m - y*$ is assumed to be in S_3. A summary of the data is presented in Table 5.1. In addition, Table 5.1 also presents the estimated HIV prevalence together with the corresponding 95% confidence intervals (95% CI).

Table 5.1: Summary of BAIS III data stratified by age and sex with missing outcome data

Demographics		Sample n (n_0, n_r, n_1, n_m)	$\hat{\theta}$ (95% CI)
Sex	Male	6536 (5603, 55, 858, 20)	14.3 (13.4, 15.1)
	Female	7808 (6220, 94, 1461, 33)	20.3 (19.4, 21.2)
Age (y)	≤ 19	5707 (5491, 28, 184, 4)	3.8 (3.3, 4.3)
	20-29	2985 (2401, 43, 523, 18)	19.6 (18.1, 21.0)
	30-39	2154 (1287, 33, 822, 12)	40.3 (38.2, 42.3)
	40+	3498 (2644, 45, 790, 19)	24.4 (23, 25.8)
Overall	All groups	14344 (11823, 149, 2319, 53)	17.6 (17.0 18.2)

Table 5.2 shows that estimated incidence using Eq (5.2.7) which we refer to as Model 1 (under MAR assumption). We note that when $n_m = 0$, that is, the number of missing specimens is zero, then Eq (5.2.7) reduces to the incidence estimation formula proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009) which we refer to as Model 2 (under MCAR assumption). We also present the estimated incidence using the formula proposed by (Wang and Lagakos, 2009). We assume $1 - p = 1.5\%$ and a mean window period of 155 days as in the BAIS III report (CSO, 2008).

We note that Model 1 produces values of the estimated incidence which are larger than those of Model 2. This is consistent with the theory. The estimated prevalence of HIV as well as the incidence are high for females and for the age groups 20-39. The difference in the estimated incidences in the two models is not large since the number of missing specimens is not too large. However, if we consider the fact that only 67% of the targeted individuals provided the blood specimen for HIV testing then the impact of missing information on the estimated incidence may be severe.

Table 5.2: Estimated HIV incidence by gender and age when is proportion of specimens that tested positive on the sensitive test are missing

		Model 1	Model 2
Demographics		I_{Model1}(95% CI)	I_{Model2}(95% CI)
Sex	Male	1.80 (1.18, 2.43)	1.76 (1.14, 2.38)
	Female	2.77 (2.05, 3.51)	2.72 (2.00, 3.44)
Age (y)	≤ 19	1.10 (0.65, 1.56)	1.08 (0.63, 1.53)
	20-29	3.55 (2.28, 4.86)	3.44 (2.17, 4.71)
	30-39	3.80 (1.72, 5.90)	3.75 (1.67, 5.83)
	40+	3.00 (1.82, 4.21)	2.94 (1.76, 4.12)
Overall	All groups	2.31 (1.83, 2.80)	2.26 (1.78, 2.74)

This is because it is not unreasonable to assume that those who refused to provide the blood specimen might have refused on the basis of their past risk exposure.

5.4 Discussion

We described a method for estimating incidence when a proportion of blood specimens that tested positive on the sensitive tested were not tested again using the less sensitive test and thus assumed to be MAR. Consistent with the theory, the methods produces values of the estimated incidence which are larger than those under MCAR. The method was extended to the method for estimating incidence proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009). However, the idea can be extended to other likelihood based estimators of incidence like the ones proposed by (Gabaitiri et al., 2013, 2012; Gabaitiri and Mwambi, 2012) in a similar way. Imputation methods for missing binary data with auxiliary variables can also be explored to actually predict the state of the missing individual as either in S_2 or S_3 given the sample is already positive under the sensitive test. Basically this would mean building a model for the conditional probability of being in state S_2 or S_3 given the sample was positive under the sensitive test. Finally, the formula for estimating incidence that account for the uncertainty of the mean window period and p (proportion of assay progressors) can be used for this setting with n_0 replaced by n_0^*.

Chapter 6

Incorporating the effect of covariates in the estimation of HIV incidence

SUMMARY. Understanding the risk factors for HIV incidence is crucial for allocation of resources and proper implementation of risk reduction programs and other intervention strategies. When data is collected longitudinally and the outcome of interest is time to event, methods such as the Cox proportional hazard method and related methods can be used to determine risk factors for HIV incidence. However, for settings where incidence is estimated from cross sectional surveys and biomarker based methods, this may not be possible unless one makes additional assumptions on event times. In this paper, we follow the procedure similar to that developed by Balasubramanian and Lagakos (2010) to investigate the risk factors associated with HIV incidence through a multiple logistic regression model. We extend the idea of Magder and Hughes (1997) to incorporate the uncertainty of the outcome of interest which in this case is HIV incidence. We use these methods to analyse data from the Botswana HIV/AIDS impact study of 2008.

KEYWORDS: HIV incidence, Non-progressors, Sensitivity, Specificity, BED assay, logistic regression

6.1 Introduction

Understanding the risk factors for HIV incidence is crucial for allocation of resources and proper implementation of risk reduction programs and other intervention measures. Longitudinal cohort studies allow investigations aimed at identifying the risk factors of HIV incidence through survival analysis techniques such as the Cox proportional hazard model due to Cox (1972). In the absence of cohort data, prevalence data have been used to study risk factors associated with acquiring human immunodeficiency virus (HIV) (Mermin et al., 2008). However, as noted by Mermin et al. (2008), such data may not reflect risk factors for recent infections.

However, availability of assays such as the BED capture enzyme immunoassay (BED assay) together with other standard antibody tests such as the ELISA has made it possible to measure incidence from cross-sectional surveys. This approach has offered advantages to traditional longitudinal cohort studies in terms of cost, follow-up bias and time (Brookmeyer et al., 1995; Janssen et al., 1998; Wang and Lagakos, 2009). But there have been challenges on how to assess risk factors for HIV incidence since survival techniques cannot be used for these settings unless one makes additional assumptions on the event times.

A flexible statistical framework which allows incorporation of the covariates information was developed by Balasubramanian and Lagakos (2010). The procedure uses the standard multiple logistic regression model.

This paper follows the procedure developed by Balasubramanian and Lagakos (2010) to investigate the risk factors associated with HIV incidence through a multiple logistic regression model. Since the outcome of interest which in this case is HIV incidence is measured with imperfect sensitivity and specificity, we extend the idea of Magder and Hughes (1997) to incorporate the sensitivity and specificity of the outcome. Gabaitiri and Mwambi (2012) have shown that if specificity equals to one then sensitivity is similar to the proportion of assay progressors (p) as defined by McWalter and Welte (2010); Wang and Lagakos (2009). That is, p is the proportion of subjects who will become reactive at some point after seroconversion and hence will be correctly classified to be in state 3. Therefore this model can be further extended to investigate the risk factors associated with HIV incidence for the estimator of incidence proposed by McWalter and Welte (2010); Wang and Lagakos (2009). We use these methods to analyse data from the Botswana HIV/AIDS impact study of 2008.

This paper is organized as follows, in Section 6.2 we consider the incidence rate ratio as a comparative measure. In Section 6.3, we look at the method for incorporating the covariates and extend it to incorporate imperfect sensitivity and specificity. We use these methods to analyse data from the Botswana HIV/AIDS impact study of 2008 in Section 6.4 and the discussion is presented in Section 6.5.

6.2 The incidence rate ratio

For settings where we are interested in the comparing the incidence rate between two groups, say, for example, between between females and males, we can can use the incidence rate ratio as the parameter of interest.

Let λ_1 and λ_2 be incidence rates for group 1 and group 2 respectively. Furthermore, let $\hat{\lambda}_1$ and $\hat{\lambda}_2$ be the estimators for λ_1 and λ_2 respectively. Then the estimated incidence rate ratio, $\hat{RR}$, is

$$\hat{RR} = \frac{\hat{\lambda}_1}{\hat{\lambda}_2} \tag{6.2.1}$$

To illustrate the use of incidence rate ratio as a comparative measure, we use the estimator of incidence proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009) given by

$$\hat{\lambda} = \frac{pn_r - (1-p)n_1}{n_0 p \mu} \tag{6.2.2}$$

where n_0 denotes the number of subjects who test negative on a more sensitive test, n_r denote the number of subjects who test positive on a sensitive test and negative on a less sensitive test and n_1 denote the number of subjects who test positive on a less sensitive test such that $n = n_0 + n_r + n_1$. The n_r individuals are also referred to as "recent infections" while the n_1 individuals are referred to as "long standing infections" or "non-recent infections". The parameter μ denotes the mean window period and p is the proportion of assay progressors as defined by Wang and Lagakos (2009). The estimator in Eq 6.2.2 can be stratified according to risk factors such as sex which yields tow independent groups.

For group 1, the estimator of incidence is

$$\hat{\lambda}_1 = \frac{pn_{r,1} - (1-p)n_{1,1}}{n_{0,1}p\mu} \tag{6.2.3}$$

and for group 2, the estimator of incidence is

$$\hat{\lambda}_2 = \frac{pn_{r,2} - (1-p)n_{1,2}}{n_{0,2}p\mu} \tag{6.2.4}$$

where $n_{r,1}$ and $n_{r,2}$ are the number of subjects who are "recent infections" in group 1 and 2 respectively. Similar arguments applies to other variables.

Assuming p and μ are the same in the two groups, it follows that

$$\hat{RR} = \frac{\hat{\lambda}_1}{\hat{\lambda}_2} = \frac{[pn_{r,1} - (1-p)n_{1,1}]n_{0,2}}{[pn_{r,2} - (1-p)n_{1,2}]n_{0,1}} \tag{6.2.5}$$

Next we derive the standard errors of the estimated incidence rate ratio using the delta method. Consider

$$log\hat{R}R = \text{log}\hat{\lambda}_1 - \text{log}\hat{\lambda}_2$$

Then

$$\begin{aligned} \text{var}[\text{log}\hat{RR}] &\approx \text{var}[\text{log}\hat{\lambda}_1] + \text{var}[\text{log}\hat{\lambda}_2] \\ &\approx \left(\frac{1}{\hat{\lambda}_1}\right)^2 \text{var}(\hat{\lambda}_1) + \left(\frac{1}{\hat{\lambda}_2}\right)^2 \text{var}(\hat{\lambda}_2) \end{aligned}$$

Therefore the $100(1-\alpha)\%$ confidence interval for incidence risk ratio is $[\exp(L), \exp(U)]$, where $L = log\hat{R}R - Z_{\frac{\alpha}{2}}\sqrt{\text{var}[\text{log}\hat{RR}]}$ and $U = log\hat{R}R + Z_{\frac{\alpha}{2}}\sqrt{\text{var}[\text{log}\hat{RR}]}$

6.3 Incorporating Covariate Dependence

6.3.1 Logistic Regression Method

In this section we consider the model for incorporating the covariates that was proposed by Balasubramanian and Lagakos (2010). Some of the standard notations and expressions are similar to those used by Balasubramanian and Lagakos (2010).

We note that estimation of HIV incidence involves a two stage algorithm such that in stage one, one takes a cross-sectional sample of size n from a population, and test each person using an initial testing algorithm usually based on a sensitive test (typically ELISA denoted by E). Individuals testing negative (E^-) are assumed to be uninfected at that time. In the second stage, individuals testing positive on the sensitive test (E^+) are tested again using a less-sensitive test (usually BED capture enzyme immunoassay often referred to as BED assay, B). Subjects that are found to be positive on a sensitive test but negative on the less sensitive test (E^+B^-) for HIV infection are considered to be recent HIV infections otherwise (E^+B^+) they are categorized as non-recent or long standing infections as also described in Janssen et al. (1998).

Now suppose $T \geq 0$ denotes the calendar time of HIV infection for someone born at time t_0. Furthermore, let $f(t|t_0)$, $F(t|t_0)$ and $\lambda(t|t_0)$ denote the density function (incidence density), cumulative distribution function and hazard function for becoming infected at time t for someone born at time t_0 respectively. Let L_2 denote the sojourn or residence time in the recent infection state with the corresponding cumulative distribution denoted by $G(\cdot)$. We assume that L_2 has support in $[0, L_2^*]$, where $L_2^* < t$

and is independent of T. For simplicity and for purposes of conveying the idea, we assume the incidence density function, $f(u)$ is constant overtime. That is

$$f(u) = f \text{ for } u \in [t - L_2^*, t].$$

Generally the hazard or incidence rate is given by

$$\lambda(t|t_0) = \frac{f(t|t_0)}{1 - F(t|t_0)}.$$

The prevalence probabilities for the 3-state model at time t are

$$\begin{aligned}\pi_1(t) &= 1 - \theta \\ \pi_2(t) &= f\mu \\ \pi_3(t) &= \theta - f\mu\end{aligned}$$

where $\theta = F(t|t_0)$. We are interested in the model that will associate X, a vector of covariates, to HIV incidence rate. To investigate this association, one can use the proportional hazards model

$$\lambda(t|X) = \lambda(t_0)e^{\beta X} \tag{6.3.1}$$

Note that when $X = 0$ (implying there is no covariate structure on incidence)

$$\lambda(t|X) = \lambda(t_0). \tag{6.3.2}$$

Standard techniques can be used to estimate the unknown parameter vector β in Eq (6.3.1). However, due to the unavailability of the event times, Balasubramanian and Lagakos (2010) proposed an alternative approach where the regression coefficients, β, in Eq (6.3.1) are estimated by fitting the logistic regression model. The logistic regression model assumes that the logarithm of the odds of the outcome is a linear function of the predictors.

The odds of the outcome E^+B^-, given that the outcome is E^+B^- or E^-B^- and vector of risk factor X, is

$$\begin{aligned}\frac{P[E^+B^-|E^+B^- \text{ or } E^-B^-, X]}{1-P[E^+B^-|E^+B^- \text{ or } E^-B^-, X]} &= \frac{\pi_2(t|t_0,X)}{1-\pi_2(t|t_0,X)}\\ &= \frac{\pi_2(t|t_0,X)}{\pi_1(t|t_0,X)}\\ &= \frac{fe^{\beta X}\mu}{1-\theta}\\ &= \lambda(t_0)e^{\beta X}\mu\end{aligned}$$

where $\lambda(t_0) = \frac{f}{1-\theta}$

Taking logarithm of the odds, we get

$$\begin{aligned}\log\left[\frac{\pi_2(t|t_0,X)}{\pi_1(t|t_0,X)}\right] &= \log\left[\lambda(t_0)e^{\beta X}\mu\right]\\ &= \log\left[\lambda(t_0)\mu\right]+\beta X\\ &= \alpha^*+\beta X\end{aligned}$$

where $\alpha^* = \text{log}\lambda(t_0)\mu$. Hence one can estimate the regression coefficients in Eq (6.3.1) by fitting a standard logistic regression model as proposed by Balasubramanian and Lagakos (2010). Furthermore, to fit a logistic regression model we must have a binary outcome. As proposed by Balasubramanian and Lagakos (2010) we can discard n_1 individuals who are “non-recent” and hence classified to be in state 3 and regard n_r subjects in state 2 who are classified as “recent” as successes and n_0 subjects in state 1 as failures. This is motivated by the way in which we defined the odds function (that is, we used information from state 1 and state 2). In a way this is a conditional logistic regression model where the inclusion criteria in the model estimation set is either a sample is recent or not yet sero-converted or uninfected and non-recent samples are excluded.

6.3.2 Incorporating Sensitivity and Specificity in the logistic regression

Assuming ELISA has perfect sensitivity and specificity, BED assay misclassification can arise when either recent infections are categorized as non-recent or the non-recent being categorized as recent. Such misclassification will lead to biased estimates of the odds ratios and their standard errors when standard logistic regression is used to model the relationship between HIV incidence rate and its risk factors as noted by Magder and Hughes (1997).

In the current paper, we define sensitivity as the probability that the BED assay is positive (test is positive) given that the subject is in state 3 (in this case we regard being diseased as being a long standing infection) while specificity is defined as the probability that the BED assay is negative (test is negative) given that the subject is in state 1 or 2 (in this case we regard being disease free as being recent infection or uninfected). In addition, Gabaitiri and Mwambi (2012) have shown that if specificity is equal to one then sensitivity is similar to the proportion of assay progressors as in McWalter and Welte (2010) and Wang and Lagakos (2009). Thus the model proposed by Magder and Hughes (1997) for adjustment for sensitivity and specificity for logistic regression with uncertain outcomes can be used to estimate unbiased odds ratio and their standard errors for these settings.

Suppose Y_i denotes the outcome of interest for the i^{th} subject such that

$$Y_i = \begin{cases} 1 & \text{if the } i^{th} \text{ subject is truly diseased} \\ 0 & \text{if the } i^{th} \text{ subject is truly nondiseased} \end{cases}$$

Let ω_i be the classification indicator such that

$$\omega_i = \begin{cases} 1 & \text{if the } i^{th} \text{ subject is classified as having the outcome} \\ 0 & \text{otherwise} \end{cases}$$

Suppose that $\hat{Y_i}$ is the probability that the i^{th} subject truly has the disease condition given ω_i and a vector of covariates X_i. It follows that if the i^{th} subject is classified as having the outcome ($\omega_i = 1$),

$$\hat{Y_i} = \frac{\mathrm{P}(Y_i = 1|X_i, \beta) * \text{sensitivity}}{\mathrm{P}(Y_i = 1|X_i, \beta) * \text{sensitivity} + \mathrm{P}(Y_i = 0|X_i, \beta) * (1 - \text{specificity})} \tag{6.3.3}$$

and if $\omega_i = 0$,

$$\hat{Y_i} = \frac{\mathrm{P}(Y_i = 1|X_i, \beta) * (1 - \text{sensitivity})}{\mathrm{P}(Y_i = 1|X_i, \beta) * (1 - \text{sensitivity}) + \mathrm{P}(Y_i = 0|X_i, \beta) * \text{specificity}} \tag{6.3.4}$$

where β is a $k \times 1$ vector of regression coefficients to be estimated and once the linear predictor is specified then by inverting logit link function,

$$\mathrm{P}(Y_i = 1|X_i, \beta) = \frac{\exp(\sum_{i=1}^{n} \beta_j X_{ij})}{1 + \exp(\sum_{i=1}^{n} \beta_j X_{ij})} \tag{6.3.5}$$

The regression coefficients, β, are estimated using the EM algorithm when sensitivity and specificity of the outcome measurement is incorporated. In the E step, the process starts by first settings β to an arbitrary value and computing the probability of Y_i for each subject.

For the maximization step, the data are then duplicated and each observation included twice, one with the outcome variable set to one (for the diseased) and another with the outcome set to zero (for the non-diseased). A weighted logistic regression model is fitted with weights equal to Y_i if the subject is diseased and $(1 - Y_i)$ if the subject is non-diseased. The new parameter estimates obtained from the fitted weighted logistic model are used to re-calculate new Y_i and the process is repeated until parameter estimates stop changing, that is, until convergence. The model was fitted using Statistical Analysis System (SAS) and the SAS macros were provided by Prof Laurence S. Magder, University of Maryland, United States of America. We modified them where necessary.

6.4 Data Analysis

6.4.1 Introduction

Botswana AIDS impact survey (BAIS) III of 2008 is the third sexual behavioral national population level survey. It included estimation of prevalence and incidence beyond traditional aims of assessing knowledge, attitude and behavior regarding HIV and AIDS. BAIS III was a cross sectional study. The objective is to analyse BAIS III data set in order to identify risk factors for HIV incidence in Botswana. However, due to a large proportion of missing data we only investigated two risk factors, namely sex and age because there was enough information on them.

In the analysis, we first consider the incidence rate ratio as a comparative measure. We also use the logistic regression model as proposed by Balasubramanian and Lagakos (2010) and extend the idea of Magder and Hughes (1997) to incorporate the uncertainty of the outcome of interest.

See (Gabaitiri et al., 2013) for a more detailed description of the data.

6.4.2 The Results

Table 6.1 shows the distribution of the sampled individuals into the 3-state disease model and the estimated incidence risk ratio (RR) and their corresponding 95% confidence intervals.

Table 6.1: Summary of BAIS III data stratified by age and sex and the estimated incidence risk ratio, RR and the corresponding 95% confidence intervals

		Sample	Estimates
Demographics		n (n_0, n_r, n_1)	RR (95% CI)
Sex	Male	6516 (5603, 55, 858)	1
	Female	7775 (6220, 94, 1461)	1.54 (0.99, 2.39)
Age (y)	≤ 19	5703 (5491, 28, 184)	1
	20-29	2967 (2401, 43, 523)	3.18 (1.82, 5.54)
	30-39	2142 (1287, 33, 822)	3.47 (1.74, 6.93)
	40+	3479 (2644, 45, 790)	2.72 (1.53, 4.84)

Table 6.2 shows the results for unadjusted logistic regression estimates of association between HIV incidence and demographic variables (age and sex) and the 95% confidence intervals for estimated odds ratios.

Table 6.2: Unadjusted logistic regression estimates of association between HIV incidence and demographic variables (age and sex) and the estimated odds ratio (OR) and their corresponding 95% confidence intervals at different combinations of (sensitivity denoted by sens and specificity denoted by spec)

		OR (95% CI)		
Demographics		(sen=0.99, spec=0.999)	(sen=0.985, spec=1)	(sen=1, spec=1)
Sex	Male	1	1	1
	Female	1.60 (1.11, 2.31)	1.54 (1.10, 2.15)	1.54 (1.10, 2.15)
Age (y)	≤ 19	1	1	1
	20-29	4.13 (2.35, 7.24)	3.51 (2.18, 5.67)	3.51 (2.18, 5.67)
	30-39	6.01 (3.35, 10.80)	5.03 (3.03, 8.35)	5.03 (3.03, 8.35)
	40+	3.91 (2.24, 6.83)	3.34 (2.08, 5.36)	3.34 (2.08, 5.36)

Table 6.3 presents the results for adjusted logistic regression estimates of association between HIV incidence and demographic variables (age and sex) and the 95% confidence intervals for estimated odds ratios.

Table 6.3: Adjusted logistic regression estimates of association between HIV incidence and demographic variables (age and sex) and the estimated odds ratio (OR) and their corresponding 95% confidence intervals

		OR (95% CI)		
Demographics		(sen=0.99, spec=0.999)	(sen=0.99, spec=1)	(sen=1, spec=1)
Sex	Male	1	1	1
	Female	1.57 (1.08, 2.28)	1.49 (1.06, 2.08)	1.49 (1.06, 2.08)
Age (y)	≤ 19	1	1	1
	20-29	4.16 (2.36, 7.34)	3.49 (2.16, 5.63)	3.49 (2.16, 5.63)
	30-39	6.09 (3.37, 11.00)	5.02 (3.02, 8.34)	5.02 (3.02, 8.34)
	40+	3.81 (2.17, 6.69)	3.22 (2.00, 5.18)	3.22 (2.00, 5.18)

Overall the results show that females compared to males have higher risk of HIV infection (RR=1.54 95% CI;0.99, 2.39) as we can see from Table 6.1. Also the risk of infection is higher for those aged 20 years and above compared to those aged ≤ 19 years. For example, when we compare those aged 30-39 years to those aged ≤ 19 years we have RR=3.18 (95% CI;1.82, 5.54). These finding are supported by the results from logistic regression model as we can see from Table 6.2. We also observe similar findings in the adjusted logistic regression model (Table 6.3). Furthermore, when we take into account the uncertainty of the outcome of interest, as determined by the diagnostic test, we see that there is an increase in the estimated odds ratios and their corresponding standard errors (as revealed by wider confidence intervals) for settings where we made adjustments for imperfect sensitivity and specificity simultaneously. But when we make adjustments for sensitivity only assuming perfect specificity, consistent with the model proposed by (McWalter and Welte, 2010; Wang and Lagakos, 2009) there is almost no change in the parameter estimates estimates and their standard errors compared to perfect sensitivity and specificity.

6.5 Discussion

We looked at methods for incorporating the effect of covariates. In particular, we focussed on the logistic regression method proposed by Balasubramanian and Lagakos (2010). We extended the method of Magder and Hughes (1997) to incorporate the uncertainty of determining the outcome of interest which in this case is HIV incidence. We also considered the comparative measure, namely, the incidence rate ratio for settings where we are interested in comparing the incidence rate between two groups. However this methods requires stratification by covariates which can be inefficient due to reduction in cell frequencies when there many covariates with many levels of stratification. In general, our analysis reveals that risk reduction programs and other intervention measures should be structured according to predictive risk factors, for example, according to sex and age.

We note that although data on other variables was available for the BAIS III data set, we only used sex and age as the risk factors in our analysis simple because other variables alone had more than 25% of the information missing thus making it impossible to include them in the analysis. The situation was worse when we try to investigate them in combination with others. Methods such as multiple imputation are usually recommended for settings where we have less than 20% of the data missing Little and Rubin (2002). Furthermore, missing data was not the focus of this research. More research is needed on this area particularly when we incorporate information on imperfect sensitivity and specificity of the diagnostic test used to measure the outcome of interest in the presence of missing data.

Chapter 7

General Conclusion

Accurate estimates of HIV incidence are crucial for planning and assessing the impact of interventions. The use of biomarkers offers a lot of promise and it circumvents some of the problems associated with estimating HIV incidence from longitudinal cohort studies. However, there are often methodological challenges on how reliable is the estimated incidence. Most of the existing and newly developed statistical methods have their advantages and disadvantages.

The major objective of this thesis is to develop statistical methods that can be used to estimate HIV incidence rate. However, most of the methods developed in this thesis can be used to estimate incidence of other diseases provided similar data as the one used here is available. The methods developed take advantage of the statistical framework developed by Balasubramanian and Lagakos (2010) to derive likelihood estimators for the incidence rate. We proposed a method to improve incidence estimation as well as precision of the estimated incidence rate by incorporating information on the immediate past prevalence. The advantage of this method is that it uses the available data on prevalence, in particular the immediate past prevalence, to improve on incidence estimate and its precision. However, the main disadvantage of the method is that there is a need to account for the uncertainty in the past prevalence in addition to the uncertainties in mean window period and the proportion of assay progressors. We further proposed a method for estimating incidence by relaxing the assumption of constant incidence density. In particular we assumed a linear incidence density. The advantage of this method is that it can be used in settings where changes in HIV incidence are over a long period of time and the period between two successive surveys, as proposed by Gabaitiri et al. (2013), is large. But, the disadvantage of the method is that the assumed linear incidence density function may not be suitable for the data hence further research that includes investigating other forms of incidence density is needed. The problem though will be identifiability of parameters. Further research can explore this. We also proposed a method for adjusting the estimates of incidence when a proportion of subjects tested using an antibody sensitive test were not tested using a less sensitive test resulting in missing data. In particular, we considered how adjustments can be made in the log-likelihood when the missing data is assumed to be missing at random. This is a flexible approach to missing address the problem of missing data, however, the problem often arises when this assumption do not hold but rather a more complicated assumption of missing not random is the one that is appropriate. A method that simultaneously makes adjustment for sensitivity and specificity was also considered in this thesis. The advantage of this method, compared to the one described by (Wang and Lagakos, 2009; McWalter and Welte, 2010), is that we are able to simultaneously make adjustments for false recent rate (1-sensitivity) and false non-recent rate (1-specificity). But unlike the method of

(Wang and Lagakos, 2009; McWalter and Welte, 2010) where one have to account for uncertainties in the mean window and proportion of assay progressors, this method adds in an additional parameter thus bringing in additional uncertainty which will results in large estimates of the standard errors and hence wide confidence intervals. We also showed how sensitivity is similar to the proportion of subjects who eventually transit the "recent infection" state to the "non-recent" infection state as described by Wang and Lagakos (2009). We also described a method for incorporating risk factors for HIV incidence and extended the method of Magder and Hughes (1997) to incorporate the uncertainty of determining the outcome of interest which in this case is HIV incidence. Logistic regression, which assumes that the logarithm of the odds of the outcome is a linear function of the predictors, was used to estimate the effect of various risk factors on HIV incidence. This is a standard method that is widely used in many areas including medical research because it is simple fit using most statistical softwares and also the estimated parameters are easy to interpret as odds ratios. The logistic regression was extended using the method of Magder and Hughes (1997).

Another point worth noting is that Wang and Lagakos (2009) have shown that the adjusted and unadjusted estimators of incidence they proposed can be biased when the wrong underlying model is assumed. The estimators of incidence proposed in this thesis are not resistant to this bias especially that the accuracy of our estimators depends largely on the accuracy of the mean window period, sensitivity, specificity and the past prevalence which are always unknown. However, more research is needed to assess the robustness of the assumptions to model mis-specification.

We however, as noted earlier in most of the chapters, that there are still challenges on the estimation of HIV incidence rate. One is how to get a reliable estimator of mean window period, sensitivity and specificity. In addition, for the proposed estimators, research is necessary to investigate how we can account for the uncertainties of the estimates of the unknown parameters. We have highlighted in most chapters that the proposed methods do not take mortality into account and thus the state prevalence probabilities that we developed may be biased. This is because, as noted by Balasubramanian and Lagakos (2010), the risk of death is higher for subjects in the late stage of HIV. However, the introduction of anti-retroviral therapy (ART) has improved survival amongst individuals with late stage of HIV and thus lowering HIV related mortality. Further research is needed on this area particulary on the effect of non-HIV (and HIV) related mortality on incidence estimation.

We used data from the Botswana AIDS Impact (BAIS) III of 2008 to illustrate the proposed methodology and only age and sex were used to illustrate the methods. Other variables alone had more than 25% of the information missing thus making it impossible to include them in the analysis. As we highlighted, the situation was worse when we try to investigate them in combination with others. Therefore methods for handling missing data such as multiple imputation could not be used since usually it is recommended for settings where we have less than 20% of the data missing (Little and Rubin, 2002). The situation is further exacerbated by the fact that only 67% of the sampled individuals refused to provided blood specimen for HIV Testing. Although the focus of this thesis was not on missing data, more research is still needed on this area.

Appendix A

A simulation study to compare the unadjusted incidence estimators with known and unknown previous prevalence under constant incidence density function

```
### This R function called "simulationkno1" is a
## simulation study that evaluates the performance of the
## proposed estimators of incidence rate
## for the model when the previous prevalence is known.
## In this simulation, we compare undjusted

## estimators for of incidence rate

## for the standard model and the proposed model

## when past prevalence is known

## The input variables are "timet"=time between

## two prevalence studies,

## "muL"=mean window period,

## "ftau.hat"=the previous prevalence,

## We define "f.true"=true value of the

## incidence density function

## "nsim"=number of simulations

## "phi.one", "phi.two", "phi.three" are

##  the prevalence probabilities

## "n"=overall sample size

## Then we generate "n00, n10, n11" about 1000, times.

## Estimate "theta.hat"="ftau.hat"+f2*timet

## Finally it estimates the incidence rate and its

## corresponding standard errors.

## and coverage probabilities


simulationkno1<-function(timet, muL, lam.true, fhat.tau,

 nsim){f.true<-lam.true*(1-fhat.tau)/(1+lam.true*timet);

theta.true<-fhat.tau+f.true*timet;

fhat.tau<-fhat.tau;
```

```
lam.true<-lam.true;
phi.one<-(1-fhat.tau-lam.true*(1-fhat.tau)*timet
/(1+lam.true*timet));
phi.two<-lam.true*(1-fhat.tau)*muL/(1+lam.true*timet);
phi.three<-1-phi.one-phi.two;
nsim<-nsim;
N<-3000;
set.seed(3000)
rcount<-rmultinom(nsim, size=N,
prob=c(phi.one, phi.two, phi.three));
n00<-matrix(rcount[1,]);
n10<-matrix(rcount[2,]);
n11<-matrix(rcount[3,]);
mat<-as.matrix(cbind(n00, n10, n11));
colnames(mat)<-c("n00","n10","n11");
lam.hat<-rep(0,nsim);
var.lam<-rep(0, nsim);
B<-rep(0,nsim);
var.lam.hat<-rep(0,nsim);
var.theta.hat<-rep(0, nsim);
lam.m3<-rep(0, nsim);
var.lam.m3<-rep(0, nsim);
i<-1
while(i<=nsim){
```

```
B[i]<-(mat[i,"n00"]+mat[i,"n11"])*timet*fhat.tau
-(mat[i,"n10"] +mat[i,"n11"])*(timet-(1-fhat.tau)*muL);
lam.hat[i]<-(-B[i]+sqrt(B[i]^2+4*mat[i,"n00"]*timet
*(timet-(1-fhat.tau)*muL)*(mat[i,"n10"]*fhat.tau)))
/(2*mat[i,"n00"]*timet*(timet-(1-fhat.tau)*muL));
####Variance+MLE matrix under constant model -begin
var.lam.hat[i]<--((((N*timet^2)/((1+lam.hat[i]*timet)^2))
-(mat[i,"n10"]/(lam.hat[i]^2))
-(mat[i,"n11"]*(timet-(1-fhat.tau)*muL)^2)/
((fhat.tau+lam.hat[i]*(timet-(1-fhat.tau)*muL))^2)))^(-1);
####Variance+MLE matrix under constant model -end
lam.m3[i]<-mat[i,"n10"]/(mat[i,"n00"]*muL);
var.lam.m3[i]<-mat[i,"n10"]/(mat[i,"n00"]*muL)^2;
i<-i+1
}
#Coverage probability - new (uadjusted) model starts here
lower<-rep(0, nsim);
upper<-rep(0, nsim);
cv<-rep(0, nsim);
for(i in 1:nsim){
lower[i]<-lam.hat[i]-1.96*sqrt(var.lam.hat[i]);
upper[i]<-lam.hat[i]+1.96*sqrt(var.lam.hat[i]);
if(lam.true>lower[i] && lam.true<upper[i]) cv[i]<-1
else cv[i]<-0;
```

```
}
#Coverage probability-new (uadjusted) model ends here

#Coverage probability-standard (unadjusted) model starts- M3
lower3<-rep(0, nsim);
upper3<-rep(0, nsim);
cv3<-rep(0, nsim);
for(i in 1:nsim){
lower3[i]<-lam.m3[i]-1.96*sqrt(var.lam.m3[i]);
upper3[i]<-lam.m3[i]+1.96*sqrt(var.lam.m3[i]);
if(lam.true>lower[i] && lam.true<upper3[i])
cv3[i]<-1 else cv3[i]<-0;
}
#Coverage probability -standard (unadjusted)
# model ends - M3

cvtable<-cbind(lower,upper,lam.true, cv)
cv<-table(cv)
cvtable3<-cbind(lower3,upper3,lam.true, cv3)
cv3<-table(cv3)

#Incidence estimation and its SE-begins
# -New (unadjusted) model
mean.lam.hat<-mean(lam.hat);
```

```
v.lam.hat.emp<-var(lam.hat);
se.lam.hat.emp<-sqrt(v.lam.hat.emp);
var.lam.hat.ave<-mean(var.lam.hat);
se.lam.hat.ave<-sqrt(var.lam.hat.ave);
#Incidence estimation and its SE-end
# -New(unadjusted) model

#Incidence estimation and its SE-begins
# -Std (unadjusted) model-M3
mean.lam.m3<-mean(lam.m3);
v.lam.m3.emp<-var(lam.m3);
se.lam.m3.emp<-sqrt(v.lam.m3.emp);
var.lam.m3.ave<-mean(var.lam.m3);
se.lam.m3.ave<-sqrt(var.lam.m3.ave);
#Incidence estimation and its SE-ends
# -std (unadjusted) model-M3

incidence.par<-cbind(lam.true,mean.lam.hat
,se.lam.hat.ave,se.lam.hat.emp,cv,cv3,
mean.lam.m3,se.lam.m3.ave,se.lam.m3.emp);
return(incidence.par);
}
```

Appendix B

A simulation study to evaluate the performance of adjusted new incidence estimator when previous prevalence is known under constant incidence density function

```
### This R function called "simulation-kno2" is a
# simulation study that evaluates the estimator for
# incidence rate for the model when past
# prevalence is known.
## The input variables are "timet"=time
```

```
# between two prevalence studies,
## "muL"=mean windon period,
# "ftau.hat"=the previous prevalence,
## and "p"=proportion of assay progressors
## We define "f.true"=true value
# of the incidence density
## function and "nsim"=number of simulations
## "phi.one", "phi.two", "phi.three" are the prevalence
## probabilities and "n"=overall sample size
## Then we generate "n00, n10, n11" about 1000, times.
## Estimate "theta.hat"="ftau.hat"+f2*timet
## Finally it estimates the incidence rate and its
## corresponding standard error.

simulationkno2<-function(timet, muL,
lam.true, fhat.tau,p,nsim){
f.true<-lam.true*(1-fhat.tau)/(1+lam.true*timet);
theta.true<-fhat.tau+f.true*timet;
fhat.tau<-fhat.tau;
p<-p;
lam.true<-lam.true;
phi.one<-(1-fhat.tau-lam.true*(1-fhat.tau)*timet/
(1+lam.true*timet));
phi.three<-p*(fhat.tau+lam.true*(1-fhat.tau)
*(timet-muL)/(1+lam.true*timet));

phi.two<-1-phi.one-phi.three;

nsim<-nsim;

N<-3000;

set.seed(3000)

rcount<-rmultinom(nsim, size=N,

prob=c(phi.one, phi.two, phi.three));

n00<-matrix(rcount[1,]);

n10<-matrix(rcount[2,]);

n11<-matrix(rcount[3,]);

mat<-as.matrix(cbind(n00, n10, n11));

colnames(mat)<-c("n00","n10","n11");


lam.hat<-rep(0,nsim);

A<-rep(0,nsim);

B<-rep(0,nsim);

C<-rep(0,nsim);

var.lam.hat<-rep(0,nsim);


#####Adjusted estimator


mr<-muL*p*(1-fhat.tau)+timet*(1-p);

ur<-fhat.tau*(1-p);

m1<--muL*(1-fhat.tau)+timet;
```

```
u1<-fhat.tau;

i<-1
while(i<=nsim){
A[i]<-m1*mr*timet*mat[i,"n00"];
B[i]<-mr*timet*u1*(mat[i,"n00"]+mat[i,"n11"])+m1*timet*ur*
(mat[i,"n00"]+mat[i,"n10"])-m1*mr*(mat[i,"n10"]+mat[i,"n11"]);
C[i]<-N*timet*ur*u1-mr*u1*mat[i,"n10"]-mat[i,"n11"]*m1*ur;
lam.hat[i]<-(-B[i]+sqrt(B[i]^2-4*A[i]*C[i]))/(2*A[i]);
#Variance+MLE matrix under constant model
var.lam.hat[i]<--((((N*timet^2)/((1+lam.hat[i]*timet)^2))
-(mat[i,"n10"]*mr^2/(ur+mr*lam.hat[i])^2)
-(mat[i,"n11"]*m1^2)/((u1+lam.hat[i]*m1)^2)))^(-1);
i<-i+1
}

#Coverage probability for true lambda-
# for the proposed model begins here
lower<-rep(0, nsim);
upper<-rep(0, nsim);
cv<-rep(0, nsim);
for(i in 1:nsim){
lower[i]<-lam.hat[i]-1.96*sqrt(var.lam.hat[i]);
upper[i]<-lam.hat[i]+1.96*sqrt(var.lam.hat[i]);
if(lam.true>lower[i] && lam.true<upper[i])
cv[i]<-1 else cv[i]<-0;
}
#Coverage probability for true lambda for
#the proposed model ends here

cvtable<-cbind(lower,upper,lam.true, cv)
cv<-table(cv)

#Incidence est-begin - Proposed model
mean.lam.hat<-mean(lam.hat);
v.lam.hat.emp<-var(lam.hat);
se.lam.hat.emp<-sqrt(v.lam.hat.emp);
var.lam.hat.ave<-mean(var.lam.hat);
se.lam.hat.ave<-sqrt(var.lam.hat.ave);
#cov1<-se.lam.hat.ave/muL;
#Incidence est-end - Proposed model

incidence.par<-cbind(lam.true,mean.lam.hat,
se.lam.hat.ave,se.lam.hat.emp,cv);

return(incidence.par);

}
```

Appendix C

A simulation study to compare the adjusted incidence estimators with known and unknown previous prevalence under constant incidence density function

```
### This R function called "simulation-kno3" is a
## simulation study that evaluates the estimator for
## incidence rate for the model
## of Wang-Lagakos (adjusted) and compare the
## estimator to the one proposed (adjusted))
## both adjusted for p where "p"=proportion of
## assay progressors. We use
## the prevalence probabilities same as under
## the proposed (adjusted) model.
## Therefore the input variables are "timet"=time
## between two prevalence studies,
## "muL"=mean windon period, "ftau.hat"=the previous
## prevalence,
## We define "f.true"=true value of the
## incidence density function
## "nsim"=number of simulations, and "p"
## "phi.one", "phi.two", "phi.three" are
## the prevalence probabilities
## "n"=overall sample size
## Then we generate "n00, n10, n11".
## Finally it estimates the incidence rate
## and its corresponding standard error.

simulationkno3<-function(timet, muL, lam.true,
fhat.tau, p, nsim) {
f.true<-lam.true*(1-fhat.tau)/(1+lam.true*timet);
theta.true<-fhat.tau+f.true*timet;
fhat.tau<-fhat.tau;
p<-p;
```

```
lam.true<-lam.true;

#phi.one3<-1-fhat.tau;

#phi.two3<-lam.true*(1-fhat.tau);

#phi.three3<-1-phi.one3-phi.two3;

phi.one<-(1-fhat.tau-lam.true*(1-fhat.tau)

*timet/(1+lam.true*timet));

phi.three<-p*(fhat.tau+lam.true*(1-fhat.tau)

*(timet-muL)/(1+lam.true*timet));

phi.two<-1-phi.one-phi.three;

nsim<-nsim;

#alpha<-0.1;

N<-3000;

set.seed(3000)

rcount<-rmultinom(nsim, size=N,

prob=c(phi.one, phi.two, phi.three));

n00<-matrix(rcount[1,]);

n10<-matrix(rcount[2,]);

n11<-matrix(rcount[3,]);

mat<-as.matrix(cbind(n00, n10, n11));

colnames(mat)<-c("n00","n10","n11");

lam.hat<-rep(0,nsim);

var.lam.hat<-rep(0,nsim);


i<-1
```

```
while(i<=nsim){
#Adjusted estimator
lam.hat[i]<-(p*mat[i,"n10"]-(1-p)
*mat[i,"n11"])/(p*mat[i,"n00"]*muL);
#Variance+MLE matrix under constant model
var.lam.hat[i]<-mat[i,"n10"]/(p*mat[i,"n00"]^2*muL^2);
i<-i+1
}
#Coverage probability - Wang-Lagakos model-begin
lower<-rep(0, nsim);
upper<-rep(0, nsim);
cv<-rep(0, nsim);
for(i in 1:nsim){
lower[i]<-lam.hat[i]-1.96*sqrt(var.lam.hat[i]);
upper[i]<-lam.hat[i]+1.96*sqrt(var.lam.hat[i]);
if(lam.true>lower[i] && lam.true<upper[i])
cv[i]<-1 else cv[i]<-0;
}
#Coverage probability - Wang-Lagakos model-end
cvtable<-cbind(lower,upper,lam.true, cv)
cv<-table(cv)
#Incidence est-begin - New model
mean.lam.hat<-mean(lam.hat);
v.lam.hat.emp<-var(lam.hat);
```

```
se.lam.hat.emp<-sqrt(v.lam.hat.emp);
var.lam.hat.ave<-mean(var.lam.hat);
se.lam.hat.ave<-sqrt(var.lam.hat.ave);
incidence.par<-cbind(lam.true,mean.lam.hat,
se.lam.hat.ave,se.lam.hat.emp,cv);
return(incidence.par);
}
```

Appendix D

A simulation study to evaluate the performance of the new incidence estimator that makes adjustments for sensitivity and specificity simultaneously under constant incidence density function

```
### This R function called "simulationsp" is a
## simple simulation study that estimates the
## incidence rate for the new model that
## makes adjustments for sensitivity (s)
## and specificity (p) simultaneously.
## "number of simulation=nsim"
## "phi.one", "phi.two" & "phi.three" are states
## prevalence probabilities. Let overall n=3000.
## Then we generate "n00, n10, n11" 1000 times.
## Finally it estimates the incidence rate,
## its corresponding standard error,
## empirical standard errors.
## and coverage probabilities

simulationsp<-function(muL, lam.true, theta.true,
s, p, nsim) {
theta.true<-theta.true;
s<-s;
p<-p;
lam.true<-lam.true;
phi.one<-(1-theta.true)*p;
phi.two<-lam.true*(1-theta.true)*muL*p
+(theta.true-lam.true
*(1-theta.true)*muL)*(1-s) ;
phi.three<-1- phi.one-phi.two;
nsim<-nsim;
N<-3000;
```

```
set.seed(3000)
rcount<-rmultinom(nsim, size=N,
prob=c(phi.one, phi.two, phi.three));
n00<-matrix(rcount[1,]);
n10<-matrix(rcount[2,]);
n11<-matrix(rcount[3,]);
mat<-as.matrix(cbind(n00, n10, n11));
colnames(mat)<-c("n00","n10","n11");

lam.hat<-rep(0,nsim);
var.lam.hat<-rep(0,nsim);

i<-1
while(i<=nsim){
#####Adjusted estimator for Sensitivity and specificity
lam.hat[i]<-(s*mat[i,"n10"]-(1-s)*mat[i,"n11"])/
((s+p-1)*mat[i,"n00"]*muL);
####Variance for the estimator adjusted for
Sensitivity and specificity
var.lam.hat[i]<-((mat[i,"n10"]+mat[i,"n11"])
*(N*s-2*mat[i,"n11"])*s+mat[i,"n00"]*mat[i,"n11"]
*(1-2*s)+mat[i,"n11"]^2)/((s+p-1)^2*mat[i,"n00"]^3*muL^2);
i<-i+1
}
```

```
#Coverage probability - The proposed model with
sensitivity and specificity
lower<-rep(0, nsim);
upper<-rep(0, nsim);
cv<-rep(0, nsim);
for(i in 1:nsim){
lower[i]<-lam.hat[i]-1.96*sqrt(var.lam.hat[i]);
upper[i]<-lam.hat[i]+1.96*sqrt(var.lam.hat[i]);
if(lam.true>lower[i] && lam.true<upper[i])
cv[i]<-1 else cv[i]<-0;
}

###Coverage interval for true lambda ends here

cvtable<-cbind(lower,upper,lam.true, cv)
cv<-table(cv)

#Incidence est-begin - New model adjusted
for Sensitivity and specificity
mean.lam.hat<-mean(lam.hat);
v.lam.hat.emp<-var(lam.hat);
se.lam.hat.emp<-sqrt(v.lam.hat.emp);
var.lam.hat.ave<-mean(var.lam.hat);
```

```
se.lam.hat.ave<-sqrt(var.lam.hat.ave);
#Incidence est-end - new model


incidence.par<-cbind(lam.true,mean.lam.hat
,se.lam.hat.ave,se.lam.hat.emp,cv);


return(incidence.par);


}
```

Appendix E

A simulation study to evaluate the performance of the new incidence estimator under linear incidence density function

```
### This R function called "simulationlm" is a
# simulation study that evaluates the
## incidence rate for the proposed model.
# when the incidence density is assumed to be linear
## It calculates the "a.hat", "b.hat", and "theta.hat".
## The input variables are "a.true", "b.true",
## "muL", "varL", "ftau.hat",
## "number of simulation=nsim"
## It first gives "theta.true"
## The "phi.one", "phi.two", "phi.three"
## are the states prevalence probabilities
## "n"=sample size, "nsim"= the number of simulations
## Then we generate "n00, n10, n11".

simulationlm<-function(a.true, b.true, tau, mu,
varL, theta0, p, nsim) {
gamma<-varL+mu^2;
p<-p;
theta.true<-theta0+a.true*tau+0.5*b.true*tau^2;
f.true<-a.true+b.true*tau;
lam.true<-f.true/(1-theta.true);
phi.one<-1-theta.true;
phi.two<-a.true*mu-0.5*b.true*(gamma-2*tau*mu)
*p+(1-p)*(theta0+a.true*tau+0.5*b.true*tau^2);
phi.three<-1-phi.one-phi.two;
nsim<-nsim;
n<-4000;
set.seed(30000)
rcount<-rmultinom(nsim, size=n,
prob=c(phi.one, phi.two, phi.three));
n0<-matrix(rcount[1,]);
nr<-matrix(rcount[2,]);
```

```
n1<-matrix(rcount[3,]);
mat<-as.matrix(cbind(n0, nr, n1));
colnames(mat)<-c("n0","nr","n1");
a<-rep(0,nsim);
b<-rep(0,nsim);
var.a.hat<-rep(0,nsim);
var.b.hat<-rep(0, nsim);
var.theta.hat<-rep(0,nsim);
var.lam.hat<-rep(0, nsim);
cov.ab.hat<-rep(0, nsim);
theta.hat<-rep(0, nsim);
lam.hat<-rep(0, nsim);
var.f<-rep(0, nsim);
cov.ftheta<-rep(0, nsim);
var.theta<-rep(0, nsim);
var.lam.hat<-rep(0, nsim);
f<-rep(0, nsim);
i<-1
while(i<=nsim){
a[i]<--(-2*n*mu*tau*p*theta0+2*mu*p*tau*(mat[i,"n1"]
+mat[i,"nr"])+n*gamma*p*theta0-p*gamma*(mat[i,"n1"]
+mat[i,"nr"])-(tau^2)*p*(mat[i,"n1"]+mat[i,"nr"])
+mat[i,"n1"]*tau^2)/(n*tau*p*(gamma-tau*mu));
```

```
b[i]<--2*(n*mu*p*theta0+tau*p*(mat[i,"n1"]
+mat[i,"nr"])-mu*p*(mat[i,"n1"]+mat[i,"nr"])
-mat[i,"n1"]*tau)/(n*tau*p*(gamma-tau*mu));

var.a.hat[i]<-(mat[i,"n0"]*mat[i,"n1"]*(-2*mu*p*tau
+p*gamma+tau^2*p-tau^2)^2+p^2*mat[i,"n0"]*mat[i,"nr"]
*(tau^2+gamma-2*tau*mu)^2+mat[i,"n1"]*mat[i,"nr"]
*tau^4)/(n^3*tau^2*p^2*(gamma-tau*mu)^2);

var.b.hat[i]<-4*(mat[i,"n0"]*mat[i,"n1"]*(mu*p
+tau-p*tau)^2+mat[i,"n0"]*mat[i,"nr"]*p^2*(mu-tau)^2
+mat[i,"n1"]*mat[i,"nr"]*tau^2)/(n^3*tau^2*p^2
*(gamma-tau*mu)^2);

cov.ab.hat[i]<-2*(mat[i,"n0"]*mat[i,"n1"]
*(mu*p+tau-p*tau)*(-2*mu*p*tau+p*gamma+tau^2*p-tau^2)
+mat[i,"n0"]*mat[i,"nr"]*p^2*(mu-tau)*(tau^2
+gamma-2*tau*mu)-mat[i,"n1"]*mat[i,"nr"]*tau^3)
/(n^3*tau^2*p^2*(gamma-tau*mu)^2);

theta.hat[i]<-theta0+a[i]*tau+0.5*b[i]*tau^2;
f[i]<-a[i]+b[i]*tau;
lam.hat[i]<-f[i]/(1-theta.hat[i]);
```

```
var.f[i]<-var.a.hat[i]+tau^2*var.b.hat[i]
+2*tau*cov.ab.hat[i];
cov.ftheta[i]<-tau*var.a.hat[i]+0.5*tau^3*var.b.hat[i]
+1.5*tau^2*cov.ab.hat[i];

var.theta[i]<-tau^2*var.a.hat[i]
+tau^4/4*var.b.hat[i]+tau^3*cov.ab.hat[i];

var.lam.hat[i]<-(1/(1-theta.hat[i])^2)*var.f[i]
+(f[i]^2/(1-theta.hat[i])^4)*var.theta[i]+(2*f[i]
/(1-theta.hat[i])^3)*cov.ftheta[i];
i<-i+1
}

#Coverage probability for a;
lowera<-rep(0, nsim);
uppera<-rep(0, nsim);
cva<-rep(0, nsim);
for(i in 1:nsim){
lowera[i]<-a[i]-1.96*sqrt(var.a.hat[i]);
uppera[i]<-a[i]+1.96*sqrt(var.a.hat[i]);
if(a.true>lowera[i] && a.true<uppera[i])
cva[i]<-1 else cva[i]<-0;
}
```

```
cvtablea<-cbind(lowera,uppera,cva);
cva<-table(cva);

#Coverage probability for b;
lowerb<-rep(0, nsim);
upperb<-rep(0, nsim);
cvb<-rep(0, nsim);
for(i in 1:nsim){
lowerb[i]<-b[i]-1.96*sqrt(var.b.hat[i]);
upperb[i]<-b[i]+1.96*sqrt(var.b.hat[i]);
if(b.true>lowerb[i] && b.true<upperb[i])
cvb[i]<-1 else cvb[i]<-0;
}
cvtableb<-cbind(lowerb,upperb,cvb);
cvb<-table(cvb);

#Coverage probability for lam;
lowerl<-rep(0, nsim);
upperl<-rep(0, nsim);
cvl<-rep(0, nsim);
for(i in 1:nsim){
lowerl[i]<-lam.hat[i]-1.96*sqrt(var.lam.hat[i]);
upperl[i]<-lam.hat[i]+1.96*sqrt(var.lam.hat[i]);
if(lam.true>lowerl[i] && lam.true<upperl[i])
```

```
cvl[i]<-1 else cvl[i]<-0;
}
cvtablel<-cbind(lowerl,upperl,cvl);
cvl<-table(cvl);
#Incidence estimation under linear density-begin
mean.a<-mean(a);
mean.b<-mean(b);
mean.lam<-mean(lam.hat);
v.a.emp<-var(a);
se.a.emp<-sqrt(v.a.emp);
v.b.emp<-var(b);
se.b.emp<-sqrt(v.b.emp);
v.l.emp<-var(lam.hat);
se.l.emp<-sqrt(v.l.emp);
var.a.ave<-mean(var.a.hat);
se.a.ave<-sqrt(var.a.ave);
var.b.ave<-mean(var.b.hat);
se.b.ave<-sqrt(var.b.ave);
var.l.ave<-mean(var.lam.hat);
se.l.ave<-sqrt(var.l.ave);
t<-phi.one+phi.two+phi.three;
#Incidence estimation under linear density-end
incidence.par<-cbind(a.true,mean.a,se.a.ave,
se.a.emp,b.true,mean.b,se.b.ave,
se.b.emp,lam.true,mean.lam,se.l.ave,
se.l.emp,cva,cvb,cvl,t);
return(incidence.par);
}
```

Bibliography

Bacchetti, P., Segal, M., and Jewell, N. P. (1993). Backcalculation of HIV infection rates. *Statistical Science* **8**, 82–119.

Balasubramanian, R. and Lagakos, S. W. (2010). Estimating HIV incidence based on combined prevalence testing. *Biometrics* **66**, 1–10. DOI: 10.1111/j.1541-0420.2009.01242.x.

Barin, F., Meyer, L., Lancar, R., Deveau, C., Gharib, M., Laporte, A., Desenclos, J., and Costagliola, D. (2005). Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use on dried serum spots. *Journal of Clinical Microbiology* **43 (9)**, 4441–4447. Doi:10.1128/JCM.43.9.4441-4447.2005.

Barnighausen, T. A., McWalter, T. A., Rosner, Z., Newell, M. L., and Welte, A. (2010). HIV incidence estimation using the BED capture enzyme immunoassay: systematic review and sensitivity analysis. *Epidemiology* **21(5)**, 685–697.

Brookmeyer, R. (1991). Reconstruction and future trends of the AIDS epidemic in the United States. *Science* **253**, 37–42.

Brookmeyer, R. (2009). Should biomarker estimates of HIV incidence be adjusted? *AIDS* **23**, 485–491.

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 bimarker assays for HIV incidence. *Journal of Acquired Immune Deficiency Syndrome* **54 (4)**, 406–414.

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**, 166–172.

Brookmeyer, R., Quinn, T. C., Shepherd, M., Mehendale, S., Rodrigues, J., and Bollinger, R. (1995). The AIDS epidemic in India: a method for estimating current human immunodeficiency virus (HIV) incidence rates. *American Journal of Epidemiology* **142 (7)**, 709–713.

Chawla, A., Murphy, G., Donnelly, C., Booth, C. L., Johnson, M., Parry, J. V., Phillps, A., and Geretti, A. M. (2007). Human immunodeficiency virus (HIV) antibody avidity testing to identify recent infection in newly diagnosed HIV typpe 1 (HIV-1)-seropositive persons infected with diverse HIV-1 subtypes. *Journal of Clinical Microbiology* **45 (2)**, 415–420. Doi:10.1128/JCM.01879-06.

Chu, H. and Cole, S. R. (2006). Estimating bimarker-based HIV incidence using prevalence data in high risk groups with missing outcomes. *Biometrical Journal* **48 (5)**, 772–779.

Claggett, B., Lagakos, S. W., and Wang, R. (2012). Augmented Cross-Sectional Studies with Abbreviated Follow-up for Estimating HIV Incidence. *Biometrics* **68 (1)**, 6274. DOI: 10.1111/j.1541-0420.2011.01632.x.

Cole, S. R., Chu, H., and Brookmeyer, R. (2006). Confidence intervals for biomarker-based human immunodeficiency virus incidence estimates and differences using prevalence data. *American Journal of Epidemiology* **165 (1)**, 94–100.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34 (2)**, 187220.

Cox, D. R. and Hinkley, D. V. (1974). *Theoreticl statistics*. Chapman and Hall.

CSO (2008). BOTSWANA AIDS IMPACT SURVEY III. Technical report, Botswana Governement.

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

Gabaitiri, L. and Mwambi, H. G. (2012). Estimating HIV Incidence with adjustment for Sensitivity and Specificity. *Submitted* .

Gabaitiri, L., Mwambi, H. G., Lagakos, S. W., and Pagano, M. (2012). Estimation of HIV incidence under Linear incidence density function. *Submitted* .

Gabaitiri, L., Mwambi, H. G., Lagakos, S. W., and Pagano, M. (2013). A likelihood estimation of HIV incidence incorporating information on past prevalence. *South African Statistical Journal* **47**, 15 31.

Gregson, S., Machekano, R., Donnelly, C. A., Mbizvo, M. T., Anderson, R. M., and Katzenstein, D. A. (1998). Estimating HIV incidence from age-specific prevalence data: comparison with concurrent cohort estimates in a study of male factory workers, Harare, Zimbabwe. *AIDS* **12**, 2049–2058.

Guy, R., Gold, J., Calleja, J. M. G., Kim, A. A., Parekh, B., Busch, M., Rehle, T., Hargrove, J., Remis, R. S., and Kaldor, J. M. (2009). Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: a systematic review. *The Lancet Infectious Diseases* **9 (12)**, 747–759.

Hall, H. I., Song, R., Rhodes, P., Prejean, J., An, Q., Lee, L. M., Karon, J., Brookemeyer, R., Kaplan, E. H., McKenna, M. T., and Janssen, R. S. (2009). Estimation of HIV incidence in the United States. *Journal of American Medical Association* **200**, 520–529.

Hallett, T. B., Ghys, P., Brnighausen, T., Yan, P., and Garnett, G. (2009). Errors in 'BED'-derived estimates of HIV incidence will vary by place, time and age. *PLoS ONE* **4 (10)**, 1–10. DOI: 10.1371/journal.pone.0007368.

Hargrove, J. W., Humphrey, J. H., Mutasa, K., Parekh, B. S., McDougal, J. S., Ntozinie, R., Chidawaniyika, H., Moulton, L. H., Ward, B., Nathoo, K., Ili, P. J., and Kopp, E. (2008). Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. *AIDS* **22**, 511–518.

Hayashida, T., Gatanaga, H., Tanuma, J., and Oka, S. (2008). Effect of low HIV type I load and antiretroviral treatment on IgG-capture BED-enzyme immunoassay. *AIDS Research and Human Retroviruses* **24**, 495–498.

Janssen, R. S., Satten, G. A., Stramer, S. L., Rawal, B. D., O'Brien, T. R., Weiblen, B. J., Hecht, F. M., Jack, N., Cleghorn, J., Kahn, J. O., Chesney, M. A., and Busch, M. P. (1998). New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *Journal of the American Medical Association* **280**, 42–48.

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

Karita, E., Price, M., Hunter, E., Chomba, E., Allen, S., Fei, L., Kamali, A., Sunders, E. J., Anzala, O., Katende, M., Ketter, N., the IAVI Collaborative Seroprevalence, and Team, I. S. (2007). Investigating the utility of the HIV-1 BED capture enzyme immunoassay using cross sectional and longitudinal seroconverter specimens from Africa. *AIDS* **21**, 403–408.

Karon, J. M., Song, R., Brookemeyer, R., Kaplan, E. H., and Hall, H. I. (2008). Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Statistics in Medicine* **27**, 4617–4633.

Lagakos, S. W. and Gable, A. (2008). *Methodological challenges in biomedical HIV prevention trials*. Washington: Institute of Medicine, National Acdemy Press.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley and Sons, second edition. ISBN 0-471-18386-5.

Magder, L. S. and Hughes, J. P. (1997). Logistic Regression when the Outcome is Measured with Uncertainty. *American Journal of Epidemiology* **146 (2)**, 195–203.

McDougal, J. S., Parekh, B. S., Peterson, M. L., Branson, B. M., Dobbs, T., Ackers, M., and Gurwith, M. (2006). Comparison of HIV type I incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the BED capture enzyme immunoassay. *AIDS Research and Human Retroviruses* **22**, 945–952.

McWalter, T. A. and Welte, A. (2010). Relating recent infection prevalence to incidence with a sub-population of assay non-progressors. *Journal of Mathematical Biology* **60**, 687–710. DOI: 10.1007/s00285-009-0282-7.

Mermin, J., Musinguzi, J., Opio, A., Kirungim, W., Ekwaru, J. P., Hladik, W., Kaharuza, F., Downing, R., and Bunnell, R. (2008). Risk Factors for Recent HIV Infection in Uganda. *Journal of American Medical Association* **300(5)**, 540–549. DOI: 10.1001/jama.300.5.540.

Novitsky, V., Wang, R., Kebaabetswe, L., Greenwald, J., Rossebkhan, R., Moyo, S., Musonda, R., Woldegabriel, E., Lagakos, S. W., and Essex, M. (2009). Better control of early viral replication is associated with slower rate of elicited antibodies in the Detuned EIA during primary HIV-1C infection. *Journal of Acquired Immune Deficiency Syndrome* **52(2)**, 265–272.

Parekh, B. S., Kennedy, M. S., Dobbs, T., Pau, C.-P., Byers, R., Green, T., Hu, D. J., Vanichseni, S., Young, N. L., Choopanya, K., Mastro, T. D., and McDougal, J. S. (2002). Quantitative detection of increasing HIV type 1 antibodies after seroconversion: A simple assay for detecting recent HIV infection and estimating incidence. *AIDS Research and Human Retroviruses* **18 (4)**, 295–307.

Parekh, B. S. and McDougal, J. S. (2005). Application of laboratory methods for estimation of HIV-1 incidence. *Indian Journal of Medical Research* **121**, 510–518.

Park, S. Y., Love, T. M. T., Nelson, J., Thurston, S. W., Perelson, A. S., and Lee, H. Y. (2011). Designing a Genome-Based HIV Incidence Assay with High Sensitivity and Specificity. *AIDS* **25(16)**, F13–F19. DOI: 10.1097/QAD.0b013e328349f089.

UNAIDS (2006). Report on the global AIDS epidemic. Available: http://www.unaids.org/en/HIV data/2006GlobalReport/default.asp. Geneva: United Nations. Accessed 8 March 2008.

Walker, N., Stanecki, K. A., Brown, T., Stover, J., Lazzari, S., Garcia-Calleja, J. M., Schwartländer, B., and Ghys, P. D. (2003). Methods and procedures for estimating HIV/AIDS and its impact: UNAIDS/WHO estimates for the end of 2001. *AIDS* **17**, 2215–2225.

Wang, R. and Lagakos, S. W. (2009). On the use of adjusted cross-sectional estimators of HIV incidence. *Journal of Acquired Immune Deficiency Syndrome* **52(5)**, 538–547.

Wang, R. and Lagakos, S. W. (2010). Augmented cross-sectional prevalence testing for estimating HIV incidence. *Biometrics* **66 (3)**, 864–874. DOI: 10.1111/j.1541-0420.2009.01356.x.

Welte, A., McWalter, T. A., and Brnighausen, T. A. (2009). A simplified formula for inferring HIV incidence from cross sectional surveys using a test for recent infection. *AIDS Research and Human Retroviruses* **25**, 125–126.

WHO (2009). WHO Technical Working Group on HIV Incidence Assays. Available: http://www.who.int/diagnostics_laboratory/links/hiviwg_capetown_07_09.pdf. Accessed 28 May 2012.

WHO (2011). When and how to use assays for recent infection to estimate HIV incidence at a population level. Available: http://www.who.int/diagnostics_laboratory/hiv_incidence_may13_final.pdf. Accessed 13 June 2011.