Characterisation and Application of Tests for Recent Infection for HIV Incidence Surveillance

Reshma Kassanjee

A thesis submitted to the Faculty of Science, University of the Witwatersrand,
in fulfilment of the requirements for the degree of Doctor of Philosophy
21 October 2014

School of Computational and Applied Mathematics
University of the Witwatersrand
Johannesburg

Declaration

I declare that this thesis is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.

_____

21 October 2014

Abstract

Three decades ago, the discovery of the Human Immunodeficiency Virus (HIV) was announced. The subsequent HIV pandemic has continued to devastate the global community, and many countries have set ambitious HIV reduction targets over the years. Reliable methods for measuring incidence, the rate of new infections, are essential for monitoring the virus, allocating resources, and assessing interventions. The estimation of incidence from single cross-sectional surveys using tests that distinguish between 'recent' and 'non-recent' infection has therefore attracted much interest. The approach provides a promising alternative to traditional estimation methods which often require more complex survey designs, rely on poorly known inputs, and are prone to bias. More specifically, the prevalence of HIV and 'recent' HIV infection, as measured in a survey, are used together with relevant test properties to infer incidence. However, there has been a lack of methodological consensus in the field, caused by limited applicability of proposed estimators, inconsistent test characterisation (or estimation of test properties) and uncertain test performance. This work aims to address these key obstacles. A general theoretical framework for incidence estimation is developed, relaxing unrealistic assumptions used in earlier estimators. Completely general definitions of the required test properties emerge from the analysis. The characterisation of tests is then explored: a new approach, that utilises specimens from subjects observed only once after infection, is demonstrated; and currently-used approaches, that require that subjects are followed-up over time after infection, are systematically benchmarked. The first independent and consistent characterisation of multiple candidate tests is presented, and was performed on behalf of the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA), which was established to provide guidance and foster consensus in the field. Finally, the precision of the incidence estimator is presented as an appropriate metric for evaluating, optimising and comparing tests, and the framework serves to counter existing misconceptions about test performance. The contributions together provide sound theoretical and methodological foundations for the application, characterisation and optimisation of recent infection tests for HIV incidence surveillance, allowing the focus to now shift towards practical application.

Acknowledgements

I am grateful to my supervisor, Alex Welte, for all his time and support over the years. It was a privilege to grow under his mentorship and be given opportunities to collaborate with leading researchers in the field.

Thank you to Thomas McWalter for guiding my early learning about incidence estimation and easing my entry into postgraduate research.

Thank you to L Paul Fatti for his assistance with statistical aspects of this work, in its early planning stages.

Over the last four years, I have had the pleasure of working closely with an exceptional team comprising the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA). I am thankful to the principal investigators, Gary Murphy, Michael Busch, Christopher Pilcher and Alex Welte, and the many others whom I have had the opportunity to meet and work with through CEPHIA, including Christine Rousseau, Elaine McKinney, Sheila Keating, Shelley Facente and David Matten. I am also thankful to the co-authors of the papers that are presented, whose contributions made this work possible. A large number of individuals, too many to list, developed my interests in and refined my understanding of the field over the course of this research – and I am grateful to each one of them.

Funding for the research undertaken, both as a student and while employed by the *South African DST/NRF Centre of Excellence in Epidemiological Modelling and Analysis* (SACEMA), was provided by the *Bill and Melinda Gates Foundation*, *Canadian International Development Agency* (CIDA), SACEMA and the *University of the Witwatersrand*.

Finally, I appreciate the encouragement and support of those most dear to me. Thank you to all my colleagues at SACEMA, and especially to Gavin Hitchcock for his guidance. Thank you to all my family and friends, and especially to Hilmarié Brand for being both of these to me while far from home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Overview

This body of work focuses on how to appropriately use tests designed to distinguish between ‘recently’ and ‘non-recently’ acquired HIV infections to estimate HIV incidence from cross-sectional surveys. This topic and the contributions made within this thesis are introduced in three parts below. Firstly, the significance of this research is broadly summarised, providing a context for why this work is of interest in epidemiology. Secondly, a brief, critical review of past developments in this surveillance approach is presented. The review aims to both provide pertinent background information to the reader, as well as highlight key shortcomings in the field that this research aims to address. Lastly, the contributions made, and the organisation of these in this thesis, are outlined.

1.1 Significance

The early 1980s proved to be a period of rapid medical advance as research groups aimed to identify the retrovirus that was causing Acquired Immunodeficiency Syndrome (AIDS), an illness which was then beginning to sweep across nations [1-3]. Now called the Human Immunodeficiency Virus (HIV), the reach of the virus, particularly in the developing world, is alarming [4]. In South Africa, an estimated 29.5% of pregnant women attending public antenatal care clinics are HIV-positive [5] and 17.9% of the general adult population, aged 15 to 49 years old, is living with HIV [4].

These frightening statistics highlight the need for an effective response to the epidemic. There is thus an urgent need for reliable epidemiological measures, which are essential to monitor viral spread, optimally allocate limited resources, and plan and assess interventions. Incidence, the rate of new infections in a population, although more difficult to measure, provides a more direct and current measure of the state of the epidemic than prevalence, the proportion of the population infected at a point in time. Incidence has therefore become the focus of 'second-generation HIV surveillance' [6] in South Africa and around the world.

Time-honoured methods of estimating incidence typically require either longitudinal studies or surveys at multiple points in time. These methods have a number of drawbacks, such as being prone to capture unrepresentative samples, requiring prolonged study periods and relying on highly uncertain external model inputs (such as mortality rates). Therefore, the prospect of measuring incidence from a single cross-sectional survey, using a test that distinguishes 'recent' from 'non-recent' infection, has attracted much interest [7-14]. However, the widespread use of this approach has been hindered by a lack of methodological consensus [7, 15-27] and the poor or uncertain performance of candidate tests for recent infection [8-10, 12-14, 28].

This work has therefore aimed to address these obstacles to the application of cross-sectional incidence surveillance. A number of closely related original contributions have been made; and these have been shared through international journal articles [29-32], conference presentations [33-44], local newsletters [45, 46], project websites and online

tools [47, 48], training courses [49, 50], and various collaborations[1]. These contributions can be broadly categorised as:

- The development of a general theoretical framework for incidence estimation, relaxing the unrealistic assumptions made in earlier work – assumptions that led to confusion about the validity of this surveillance approach
- Methodological advancement in the measurement of test properties required for incidence estimation, both by formalising new ideas for the characterisation of tests, and by building consensus through the rigorous benchmarking of widely used approaches
- The first independent, large-scale and consistent characterisation of multiple candidate tests for recent infection, aimed at promoting standardisation in the field and guiding ongoing 'recent' infection 'biomarker discovery' projects
- The provision of guidance on the appropriate metric for the evaluation, optimisation and comparison of tests for recent infection, serving to counter existing misconceptions that could lead to spurious test assessments

1.2 Review of Incidence Estimation Using Tests for Recent Infection

The history and state of HIV incidence estimation, from cross-sectional surveys using tests for recent infection, excluding the contributions made in this work, are briefly reviewed below.

Traditionally, disease incidence is measured by directly counting the number new infections that occur while following a cohort of initially uninfected subjects. However, the required longitudinal surveillance is costly and difficult to administer, both practically and ethically, while limiting bias in results [12, 54, 55]. Another familiar approach is the calculation of historic incidence values that are consistent with prevalence data for multiple time points and age groups [56-60]. For example, in South Africa, prevalence

[1] For example, with the *Centers for Disease Control and Prevention* (CDC) [51], with the *World Health Organization* (WHO) *Technical Working Group on HIV Incidence Assays* [52], as a member of the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA) [47], and as a member of the *HIV Modelling Consortium Task Team for Incidence Assay Characterization* [53].

data from both general household surveys [61] and the sentinel surveillance of pregnant women attending antenatal clinics [5] are used to estimate incidence. This approach also underlies the widely-used UNAIDS Estimation and Projection Package (EPP) [62-64]. However, post-infection survival, a requisite external input, is typically poorly known and evolving (for example, as the effectiveness of treatment, and its coverage, grows). In countries such as the USA, case reporting data are used instead to 'back-calculate' incidence from reported HIV and AIDS cases [65-68]. However, often only incidence far into the past can be estimated, and assumptions about HIV progression and testing behaviour are required. Alternatively, microscopically descriptive models, such as the UNAIDS Modes of Transmission Model [69, 70], could be used. While these models attempt to capture the mechanisms of transmission in detail (for example, by estimating counts of sexual transmissions in various risk groups), the sophistication is often at odds with realistically available data to inform input parameters.

Consequently, the estimation of incidence using prevalence measurements obtained in a *single cross-sectional survey*, and involving only a *few well-estimable parameters*, holds much appeal [7-27]. This is straightforward for conditions with short and well-characterised durations, such as influenza: the prevalence of having the condition, relative to that of being susceptible to it, is equal to the product of incidence and the average duration of the condition (assuming equilibrium of constant incidence and susceptible population size over the maximum duration of the condition) [71]. Heuristically, the number of persons currently infected is the number of persons infected in, for example, the preceding week – and so incidence and prevalence closely track each other.

However, for an enduring condition, such as HIV, time in the infected state is long, difficult to characterise and ever-evolving, and so prevalence becomes uninformative of incidence. In 1995, Brookmeyer and Quinn [7] demonstrated that a cross-sectional approach to incidence estimation could nevertheless be used, by considering instead an early phase of HIV infection, termed 'recent' infection, that could be better characterised.

In the pioneering work of Brookmeyer and Quinn [7], the detection of p24 antigens in subjects with undetectable HIV antibodies identified putatively 'recently' infected subjects. However, this produced a state of 'recent' infection that lasted for only a few weeks, and therefore unrealistically large samples would be required to obtain precise incidence estimates [7, 8, 10, 11]. This triggered the development of further tests for recent infection, seeking more enduring post-infection 'recent' states [8-10, 12].

However, it quickly became evident that there is substantial inter-subject variability, with the result that some subjects who have been infected for many years nevertheless return 'recent' results [8-10, 12-14, 72-81] – typically referred to as 'false-recent' results. Accounting for this phenomenon has led to the emergence of two schools of thought [18-22].

In the first school of thought, additional parameters are used to discount the observed 'recent' results in the survey by the 'false-recent' results. This was introduced in 2006 by McDougal et al [16], who proposed an incidence estimator containing a test 'sensitivity', 'short-term specificity' and 'long-term specificity', which was further studied by Hargrove et al [17]. However, in 2009, McWalter and Welte demonstrated redundancy in these parameters [23, 82]. Furthermore, relaxing the assumption of uniform infection times made in these earlier incidence estimators [23], McWalter and Welte showed that one can instead obtain a particular weighted average of recent, and potentially changing, incidence. Two parameters naturally occur in their incidence estimator: a 'false-recent rate' (FRR) and 'mean duration of recent infection' (MDRI) [25]. A comparison of incidence estimators revealed that even under steady state conditions, only this last estimator produced unbiased results [23]. Wang and Lagakos derived the same estimator, however using additional assumptions [24].

Despite these advances, problematic assumptions remained: namely that (i) 'false-recent' results are caused (solely) by test 'non-progressors', defined as individuals who never transition out of the 'recent' state, while all remaining individuals transition (once) out of this state within some relatively short period after infection (for example, a year); and (ii) post-infection survival and test classifications are independent [23-25]. Under these assumptions, the above-mentioned FRR is the proportion of (infected) individuals who are 'non-progressors', and MDRI is the average time spent alive and in the 'recent' state for all other individuals [23-25]. However, it is known that subjects may transition (multiple times) between the 'recent' and 'non-recent' states over many years post infection (for example, patients who have AIDS or who are on antiretroviral treatment may return to the 'recent' state) [72-80] and that survival and test results are dependent (for example, survival is related to viral load and treatment [83-85], which are in turn related to biomarker results [72-81, 86]).

In the second school of thought, the original one-parameter incidence estimator (that is, the estimator containing a single test property) of Brookmeyer and Quinn [7] has been

retained [15, 27, 87-90]. All inter-subject variability (including that elsewhere called 'false-recent' results) is captured in a consistently-defined MDRI, which, in this paradigm, is defined as the average time spent alive and in the 'recent' state, considering all subjects over their full post-infection lifetimes [7, 15].

However, as the maximum post-infection time at which a person can return a 'true-recent' result can now be decades, this approach has a number of limitations. For one, the assumption of a constant susceptible population size over this duration preceding the survey, implicit in all estimators [7, 23, 25], becomes more likely to be substantially violated. Also, the larger this maximum post-infection time, the less informative is the measured incidence of recent incidence, as the implied weighting of historic incidence extends further into the past – thus limiting the ability to detect changes in incidence. Researchers have summarised this complexity into a 'shadow' or lag of the incidence estimator: by assuming a demographic and epidemiological history for the population, and particular dynamics of the test for recent infection, the cross-sectional study result can be interpreted as an estimate of incidence *some time* into the past [15, 27].

Furthermore, the MDRI is more difficult to estimate practically, as it requires a detailed understanding of survival and test dynamics even at very large times post infection. The MDRI would also be prone to varying by time and place – for example, by treatment coverage and as post-infection survival evolves. The estimation of the MDRI is burdensome, requiring the extensive longitudinal studies [16, 17, 91-103] that this cross-sectional surveillance approach aims to avoid. It is therefore key that a single MDRI estimate can be used in a number of cross-sectional incidence studies (across study populations and time).

Given these limitations of operating in a one-parameter world, and the current non-ideal performance of tests (occurrence of 'recent' results at large times after infection), this approach is not pursued in this work.

Alongside the refinement of the theoretical framework for incidence estimation, there has been substantial development of *Tests for Recent Infection* (TRIs), or *Recent Infection Testing Algorithms* (RITAs), over the years [8-10, 12]. These tests are typically based on measuring some viral process or host response, with classification rules used to map quantitative biomarker measurements onto dichotomous 'recent' and 'non-recent' classifications.

A couple of years after the initial work of Brookmeyer and Quinn [7], Janssen et al [91] coined the term *Serological Testing Algorithm for Recent HIV Seroconversion* (STARHS) when he introduced a test for recent infection based on a modification of an existing diagnostic test. By tailoring laboratory procedures so that antibody measurements grew more slowly over time after infection than they did in the original diagnostic test, readings below a chosen threshold were interpreted as being indicative of 'recent' infection. This concept has since been used to develop 'less-sensitive' or 'detuned' versions of other existing diagnostic tests [100, 104-106].

In 2002, the CDC attempted to overcome the HIV subtype-dependence of earlier tests with the creation of the BED assay, specifically for this surveillance application [92, 96, 107]. Low BED measurements of HIV-specific Immunoglobulin G (IgG) in total IgG indicate 'recent' infection. However, subtype differences remain [8, 9, 12, 108-110] and a significant proportion of subjects return 'recent' results years after infection [9, 10, 12, 16, 17, 26, 28, 73, 75-78, 111, 112]. For this surveillance application, an FRR of even a few percent critically undermines the potential of a test. In an attempt to address the high FRR, the CDC consequently developed the Limiting Antigen (LAg) assay, described in 2010 [98, 113]. Antibody avidity, the strength of binding between virus antigens and host antibodies, is expected to increase over time since infection as the host's immune response matures [9, 107]. LAg seeks to quantify this avidity, with measurements below a chosen threshold producing 'recent' classifications [98, 113]. The assay has since become widely used, despite remaining uncertainty about its performance [79, 81, 114].

A number of other candidate tests for recent infection have been developed, founded on concepts similar to those introduced above [8-10, 12, 107, 115-121]. Over the last few years, further innovative tests have been proposed: some based on measuring genetic diversity [122-125]; some using platforms that produce a number of related serological markers (such as measures of titre and avidity of antibodies to various HIV proteins), which are then together used to produce a classification [39, 40, 44, 86, 102]; some combining multiple independently-developed biomarkers, including clinical indicators such as viral load and CD4 cell counts [13, 61, 88, 89, 95, 101, 126] (sometimes termed *Multi Assay Algorithms*); and some even of the form of rapid tests [39, 40, 44, 127-129].

However, all of this development of new tests for recent infection has continued to be undermined by a number of obstacles, discussed below.

For application to incidence surveillance, tests must first be *characterised* – a term used in this work to describe the estimation of test properties appearing in the incidence estimator, namely the FRR and MDRI. This process of test characterisation has presented a bottleneck to test development, as developers and researchers have grappled with the questions of which test properties to estimate, how best to estimate them, and where to get the relevant data.

Estimation of the MDRI has typically relied on longitudinal data that describes the trajectories of biomarkers as a function of time since infection [16, 17, 91-103]. Capturing such data requires the regular follow-up and testing of (initially HIV-negative) subjects, and is therefore costly and logistically burdensome. Even given suitable data, confusion about the applicability of the theoretical framework for incidence inference has led to inconsistencies in, or unclear definitions of, test properties estimated in the literature. Furthermore, various analytical methods for characterising tests have been applied, and understanding methodological differences and identifying best practices has been an ongoing topic among experts in the field over recent years [130].

In principle, given a general incidence inference framework, any test can be consistently characterised and applied in a cross-sectional study to obtain valid incidence estimates. However, this mathematical process alone reveals little about the practical utility of a test for this surveillance application. Further challenges in test development therefore arise from the lack of standard measures to assess test performance, which are required to optimise tests and understand the relative performances of proposed tests. Within the research community, a 'Target Product Profile' [14] has been circulated and provides criteria candidate tests must meet. These criteria include a minimum MDRI (of 4 months) and maximum FRR (of 2%), and various conditions that support the viable transferability of technology. However, there remain no standards to further discriminate among tests, and various analyses have been published on candidate tests. These analyses often present estimates of test 'sensitivity' and 'specificity' (to detect infection within some specified period post infection) [92, 109, 111, 112, 115, 116, 118, 120, 124, 125] – metrics familiar in diagnostics settings, rather than performance characteristics which can be directly interpreted as indicating a test's utility for this unique population-level surveillance application. Alternatively, analyses compare incidence estimates obtained using various tests for recent infection to one another or to incidence measured longitudinally, studying the same population or even using the same sample of subjects (often termed 'field validation') [16, 90, 95, 103, 131-133]. However, this comparison is

uninformative as any consistent characterisation and application of tests will lead to incidence estimators with similar expected values (although potentially different variances).

In the light of the challenges noted above, there has been a call, in meetings convened by the *World Health Organisation* (WHO) *Technical Working Group on HIV Incidence Assays*, for a statistically sound and consistent comparative analysis of existing tests for recent infection as a logical next step to move the field forward [10, 14]. In response to such calls, in 2010, the *Bill and Melinda Gates Foundation* (BMGF) awarded a grant to establish the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA) [47] – a collaboration between *Public Health England* (PHE) in the UK; *Blood Systems Research Institute* (BSRI) in the USA; *University of California, San Francisco* (UCSF) in the USA; and the *South African Centre for Epidemiological Modelling and Analysis* (SACEMA) in South Africa.

The broad mandate of CEPHIA is to foster consensus in the scientific community. More specifically, CEPHIA was tasked with developing a specimen repository, and rigorously evaluating existing and new tests for recent infection. Given the successes of CEPHIA, in 2013, the BMGF awarded funding for a second phase of CEPHIA: the repository is being expanded to include non-plasma specimens (such as dried blood spots, oral fluids, stool samples, and urine), and the consortium is providing specimens and analytical support to a number of 'biomarker discovery' groups currently developing novel biomarkers to identify 'recent' infection. It is as a member of the core CEPHIA team that much of the work presented in this thesis has been performed.

Interest in tests for recent infection extends beyond cross-sectional incidence surveillance, but the framework for each application is unique and nuanced. In particular, in countries such as France and the USA, HIV incidence is estimated by testing patients newly diagnosed with HIV in health care settings for 'recent' infection [134-138]. By applying tests for recent infection to only patients who actively seek HIV testing, rather than those identified as HIV-positive in a cross-sectional survey of the population, testing behaviour needs to be accounted for when estimating incidence for the general population. Another application that has attracted considerable interest is the estimation of the durations of infections in (newly diagnosed) individuals and the reporting of these results to clinicians and patients, for example, for purposes of contact tracing and tailoring treatment plans. Tests for recent infection have been routinely utilised in clinical settings in areas of the

UK since 2009 [139-141], and this application has been piloted elsewhere [142]. Inter-subject variability in test dynamics should ideally be very small for such individual-level diagnostic applications.

Furthermore, the discourse on incidence estimation is moving rapidly towards methodologies that utilise multiple datasets to estimate incidence – combining measurements of the prevalence of 'recent' HIV infection with, for example, sentinel HIV prevalence surveillance data, national household survey data, life tables and treatment coverage statistics. Given the substantial subtleties around each application of tests for recent infection, the scope of this work is restricted to incidence estimation from cross-sectional surveys.

1.3 Contributions and Structure of Thesis

This work has focused on addressing key shortcomings in the field, which are highlighted in the review provided in Section 1.2. The contributions made, and the presentation of these in this thesis, are described below.

A key aspect of this research has been the dissemination of the work undertaken: in international journals [29-32] and local newsletters [45, 46], through presentations and posters at local and international conferences [33-44], through collaborations with various working groups and organisations,[2] as a member of the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA) [47], by providing training [49, 50] and by developing an online resource for guidance and analysis tools [48]. A selection of these referenced items is presented in this thesis.

Within each chapter, the core (published or prepared) journal article is presented as the first section, and remaining sections provide ancillary analysis details, related ideas, and applications of the concepts described. All articles are reproduced with permission of the publishers – with only minor modifications to text and notation, and so there is some

[2] For example, the *Centers for Disease Control and Prevention* (CDC) [50], *World Health Organization* (WHO) *Technical Working Group on HIV Incidence Assays* [51], and *HIV Modelling Consortium Task Team for Incidence Assay Characterization* [52].

repetition of background information. Because accepted terms evolved over the years during which this research was conducted, terminology varies slightly across sections.

Before trying to answer specific questions about its application, a sound theoretical foundation for incidence inference is required. A general framework for incidence inference was therefore derived, relaxing assumptions, made in previously proposed incidence estimators, of demographic and epidemiological equilibrium and about test dynamics. This derivation is presented as the opening work of this thesis in Chapter 2 [29]. The general framework was constructed by introducing a post-infection time cut-off, T, separating 'true-recent' and 'false-recent' results, with an appropriately defined 'false-recent rate' (FRR) and 'mean duration of recent infection' (MDRI) emerging from the analysis as the required test properties. Any unavoidable residual bias, such as from a varying susceptible population size, is systematically defined. Bias terms are considered in detail as part of ancillary explorations of the incidence estimator, which also include discussions on test characterisation, the moments of the incidence estimator, and the prospects of estimators containing additional test property parameters. An online resource and set of analysis tools [48], produced to support the application of the framework presented, are also described.

Having developed a general theoretical framework for incidence inference, focus shifts to the estimation of the required test properties – namely the FRR and MDRI. These properties are typically estimated in separate studies, prior to the incidence survey. The estimated test properties would ideally be recycled across multiple incidence surveys.

A bottleneck to the estimation of the MDRI has been the scarcity of specimen sets needed to generate relevant data – traditionally understood to contain multiple specimens per subject, drawn over time from soon after some (well-estimable) infection date. Furthermore, any access to such precious specimens usually requires that preliminary estimates of test properties suggest suitable promise of the test. A novel idea for obtaining preliminary estimates of the MDRI, using more widely available specimens, is therefore formalised, and its utility demonstrated, in Chapter 3 [31].

Under the proposed methodology, only a single specimen for each subject in the sample is required, drawn at the time of the first HIV-positive test, as well as knowledge of the time of the last HIV-negative test [143]. Moreover, if the period between HIV tests is approximately equal to the post-infection time cut-off T (defined above), the proposed method of analysis becomes non-parametric, safeguarding against bias from poor

parametric assumptions (at the expense of reduced precision). The framework in which these concepts are presented assumes that infection times are uniformly distributed between HIV-negative and HIV-positive tests, and this is reasonable in settings where testing schedules can be considered to be independent of subject behaviour. The methodology is explored by characterising two tests for recent infection (Less-sensitive Vitros [100] and Less-sensitive Vironostika [104]), using data from blood donors in South Africa and the USA. Ancillary analysis details are provided, and two further applications are summarised – a first characterisation of a new candidate test (based on SMARTube^{TM} technology) [35], and a local characterisation of an already widely used test (the BED assay) [33]. At the time of performing the work, the framework of McWalter and Welte [25] was in use, and therefore all analyses for estimating the MDRI were founded on the definitions of test characteristics presented in their work. A discussion of the application of the approach, under the general inference framework provided in Chapter 2, shows how the ideas presented remain just as relevant and valuable.

Even when the more precious panels of specimens are available, obtained by following subjects over time, various methods of estimating the MDRI have been used. There had been little exploration of the methodological differences, and of any artefacts in test characterisation which would in turn bias incidence estimates. An extensive, benchmarking exercise was therefore performed, and is presented in Chapter 4 [53]. Prominent researchers in the field were invited to participate in this project, which was commissioned by the *HIV Modelling Consortium* [144]. A platform for stochastically simulating data panels was developed. A large number of characterisation methods were implemented, and each used to estimate the MDRI. By simulating data, the experiment could be replicated many times, and the true MDRI calculated. This allowed the performance of the methods to be rigorously evaluated for the first time, in scenarios which systematically varied the extent of challenges encountered in reality.

A key obstacle to accurate MDRI estimation, as highlighted by the results of the benchmarking exercise, is the unknown infection times of subjects (which are only known to lie between last HIV-negative and first HIV-positive tests). An innovative idea is presented for limiting the bias that this could cause: subjects only enter the 'HIV-positive' (and 'recent') state some short time *after* detectable infection, where the time of entry into this deferred state could be more accurately estimated (than the time of entry into the otherwise-defined 'HIV-positive' state) given typically available data [34]. The

procedure for incidence estimation would remain unchanged, with the understanding that a short time lag for incidence estimation is introduced. Also, a framework is outlined for the estimation of infection times from subjects' diagnostic testing histories. Published analyses have paid little attention to this aspect of MDRI estimation, and any consequent artefacts in the estimation (or inconsistent definitions) of infection times would bias MDRI estimates.

Results from the characterisation of five prominent tests for recent infection are presented in Chapter 5 [32], and demonstrate the application of the methodologies developed in preceding chapters. This is the first of a series of planned publications by CEPHIA, and is the culmination of over three years of intense specimen collection and testing, and data gathering and analysis – all applying stringent quality control measures which it is hoped will guide standards in the field. Potential users of cross-sectional incidence surveillance have looked to CEPHIA for independent and careful guidance on its application, and therefore this work represents a particularly important milestone. Results indicate that each assay, used according to developers' guidelines, performs inadequately in isolation, and therefore further optimisation of these assays is required. Since viral suppression appears to drive FRRs, optimal use of assays is likely to include supplemental viral load tests. CEPHIA, having produced an invaluable data resource, and other groups are currently exploring this and more.

By consistently applying the concepts presented above, any test can in principle be characterised and used to produce valid incidence estimates. With the current surge in the development of tests for recent infection, the need for standard measures of a test's performance for this surveillance application has become increasingly urgent. A guidance article, targeted at test developers and analysts, is therefore presented in Chapter 6, and argues against the widespread use of sensitivity and specificity measures, which can produce spurious assessments of a test's utility. Since bias in incidence estimation is demonstrably small in the relevant contexts, the precision of the incidence estimator provides the only remaining important metric of performance. Although precision is context-specific, it provides a standard measure to assess, optimise and compare tests for recent infection.

As general and standard frameworks for incidence estimation and test evaluation are adopted, discourse in the field is inevitably moving towards application. Practical aspects of test optimisation are therefore briefly discussed, and include the scope of the

optimisation, the context-dependence of test performance, and the consideration of other test design criteria. The intention of this final section of the work is to touch on some of the topics that are likely to move towards the forefront of future discussions, given recent developments in the field, including those provided in this work.

In closing, the contributions presented in this thesis are summarised in Chapter 7, and some perspectives on future directions are provided.

Appendix A provides a detailed account of the calculation of outputs in each of the online analysis tools. All programming was performed in Matlab (The MathWorks, Inc., R2013b, 8.2.0.701), and selected code is provided in Appendix B.

Chapter 2

A General Theoretical Framework for Incidence Inference

A general theoretical framework for the estimation of incidence, from cross-sectional surveys using tests for recent infection, is presented.

In Section 2.1, a general incidence estimator is derived, relaxing assumptions contained in earlier proposed estimators. This work is a reproduction of a published journal article [29].[3] The framework is further explored, and ancillary analysis details provided, in Section 2.2, which was published as an appendix to the article. An online resource was developed over the course of this work, providing theoretical background information and practical analysis tools to users of this surveillance approach [48], and is described in Section 2.3.

[3] The contents of Section 2.1 have been published as: 'Kassanjee R, McWalter TA, Bärnighausen T, Welte A. A new general biomarker-based incidence estimator. *Epidemiology*. 2012; 23(5):721-728'. The article was reproduced with permission from *Wolters Kluwer Health*, *Lippincott Williams & Wilkins*. The manuscript was primarily written by RK, who performed the analysis, both formal, and of simulated data. TAM generated the simulated data. AW and TAM helped conceive the analysis and assisted in writing the manuscript. TB supported the project by reviewing results and text.

2.1 A New General Biomarker-Based Incidence Estimator

2.1.1 Introduction

The measurement of disease incidence — the rate of new cases in a population — is essential for effectively monitoring the spread of disease, and for targeting and assessing interventions. Longitudinal studies which directly count new infections are costly, time-consuming, and prone to capturing unrepresentative behaviour. Estimating incidence by modelling multiple prevalence values requires knowledge of the survival of those affected and unaffected by the condition. For incurable conditions such as HIV, prevalence emerges as a slow convolution (averaging) of historic incidence with survival and the variation in the size of the susceptible population. Thus, changes in prevalence over time are poor proxies for recent incidence. On the other hand, it has long been noted that prevalence of recent infection can be a very good proxy for recent incidence. Deriving an incidence estimate from a single cross-sectional survey has enormous practical advantages. There has consequently been considerable interest in developing recent infection tests based on host or viral biomarkers. This approach has been explored particularly in the context of HIV incidence [8-11]. A number of methodologies have been proposed [7, 15-17, 24, 25, 87, 91, 107, 143, 145, 146], and these have been reviewed and critiqued elsewhere [8-11, 18-21, 23, 26].

The limitations of current methodologies (using biomarkers for recent infection to estimate incidence) are hindering consensus, the development of test technology, and field implementation. There is no simple solution to the problem of estimating incidence, a rate, from a single cross-sectional survey, because there is an unavoidable loss of information when a population history is summarised into an instantaneous population state. Previously proposed estimators have been derived under very specific assumptions (known to be substantially violated) concerning both the epidemiological and demographic context, as well as the behaviour of the recent infection tests (as will be described below) [7, 15-17, 23-25, 87, 91, 107, 143, 145, 146].

Methodological background

Most tests for recent HIV infection classify persons as 'recently' infected based on a 'below-threshold' immune response such as antibody titre, avidity, or HIV-specific IgG proportion [8-10]. There is some evidence, for all tests proposed to date, that a small minority of persons remain classified as 'recently' infected long after infection [16, 17]. Additionally, late-stage HIV disease or treatment leading to viral suppression may diminish the host immune response, returning long-infected persons to the 'recent' infection state [8-11, 73-76]. This is the physiological basis for the introduction of the notion of 'false-recent' results, the effects of which are encoded into a population-level parameter widely called the 'false-recent rate' [16, 17, 24, 25]. This is not a 'rate' in the conventional sense, but the proportion of persons not 'truly' recently infected, who nevertheless produce a 'recent' result with the biomarker. In earlier analyses, dynamics were summarised into only one parameter – a mean duration of recent infection [7, 15].

Much of the analytical complexity and methodological contention arises from the difficulty of formally defining 'true-recent' and 'false-recent' results. Initially, attempts to account for 'false-recent' results were inspired not just by the biological variability noted above, but by a pattern of cross-sectional incidence estimates that were higher than prospectively obtained estimates in the same populations. However, as pointed out by Brookmeyer [18], subtracting 'false-recent' results is not the only way to obtain consistency – one can simply account for all times spent in the 'recent' state when defining the mean duration of recent infection. In practice, however, this creates other problems, one being that the development of a new test, a process that includes the estimation of a mean duration of recent infection, cannot feasibly wait for a decade or two of follow-up of a seroconverter cohort. Also, a long duration of recent infection (as defined by existing tests) can cause problematic temporal bias or blurring of incidence estimates – the extreme case being the use of prevalence as a proxy for incidence. Indeed, the notion of recent infection does not provide a totally unbiased estimate of the instantaneous incidence, but, at best, a weighted average of recent incidence, which in principle can be very close to a uniformly-weighted average over recent times. This statistical weighting may be understood by noting that, as people can persist in the 'recent' state for some time, a range of past values of incidence contributes to the current population count of 'recently' infected individuals.

Kaplan and Brookmeyer [15] and, more recently, Brookmeyer [27] explored this in the special case of incidence varying linearly with time, in which case the temporal statistical weighting can be summarised as a time lag of the incidence estimator, which they refer to as the 'shadow'. The longer this 'shadow', the less informative the estimate, with less power to detect changes in incidence over short periods of time. A key benefit of a rigorous notion of 'false-recent' results (in addition to a complementary notion of the mean duration of recent infection that is more practical to measure) is the reduction of this temporal bias.

In order to implement a formally consistent definition of both a 'false-recent rate' and mean duration of recent infection, an explicit time cut-off, T, is introduced to separate 'true-recent' and 'false-recent' results. To lead to an informative estimator, this cut-off, though theoretically arbitrary, must be chosen to reflect the temporal dynamic range of the test for recent infection; that is, at a time T post infection, the overwhelming majority of infected people should no longer be testing 'recent', and furthermore, T should not be larger than necessary to achieve this criterion. Although this time cut-off is reminiscent of a cut-off in previous analyses [16, 17], the present work dispenses with problematic assumptions of past analyses that have prevented the widespread use of cross-sectional incidences estimation from data on recent infection.

2.1.2 Analysis

The exposition proceeds in four key steps:

1. The derivation of a simple, general expression for a weighted mean recent incidence, which can be constructed without any particular assumptions about the demographic or epidemiological history or the dynamics of the biomarker used to classify persons with a disease (such as HIV infection) as 'recently' or 'non-recently' infected.
2. The derivation of an incidence estimator by expressing the general weighted incidence in terms of (i) quantities that can be known by an experimenter, and (ii) a bias term, the size of which can be approximately estimated in terms of a number of dimensionless parameters that characterise the failure of the test and context to conform to certain idealisations.
3. The estimation of the test characteristics.
4. The application of the methodology to estimate incidence in simulated scenarios, to confirm the consistency of confidence intervals.

A general expression for weighted incidence

A test for recent infection may employ an arbitrarily complex combination of criteria to classify infected persons as 'recently' or 'non-recently' infected [10]. It is understood that there will be natural inter-subject variability in progression through these categories after becoming infected. This range of responses may be captured in a function $P_R(\tau)$, which is the probability of still being alive and 'recently' infected at time τ post infection. Let $P_A(\tau)$ denote the probability of being alive at time τ post infection.

Throughout this work, 'infection' refers to the detectable infection of HIV, which depends on the diagnostic test being used. Practically, the delay between actual infection and detectable infection merely implies an epidemiologically inconsequential delay in entering the operationally HIV-positive state. This point is explored in more detail in Section 2.2.1.

Assuming a continuous population dynamic, and using the reference time $t = 0$ as the time at which a survey is conducted, consider the following explicit weighted averaging of incidence over a period of duration T:

$$I_T = \frac{\int_{-T}^{0} I(t) N_S(t) P_R(-t)\, \mathrm{d}t}{\int_{-T}^{0} N_S(t) P_R(-t)\, \mathrm{d}t}, \tag{2.1}$$

where the (possibly time-dependent) incidence is denoted by $I(t)$ and the susceptible population by $N_S(t)$. The incidence at time t contributes to I_T with weight $N_S(t)P_R(-t)$, that is, with a weight proportional to (i) the susceptible population vulnerable to being infected at time t, $N_S(t)$, and (ii) the probability of a person infected at time t still being alive and 'recent' at the time of the survey, $P_R(-t)$. It will emerge that, for practical purposes, this is very close to a uniformly-weighted average of incidence over a period preceding the survey. There is no exact way to extract either a uniformly-weighted average or an estimate of incidence at one point in time, given the substantial compression of information from a population history into a cross-sectional survey.

If there is a critical value, T_{crit}, with the property that $N_S(-t)P_R(t) = 0$ for all $t > T_{crit}$, then all choices of $T > T_{crit}$ yield the same result – the one obtained by Brookmeyer and Quinn [7]. For any finite value of T, the 'shadow', or temporal bias, is strictly less than T, and it is less than $T/2$ if $N_S(-t)P_R(t)$ is a strictly decreasing function of time in the interval $[0, T]$.

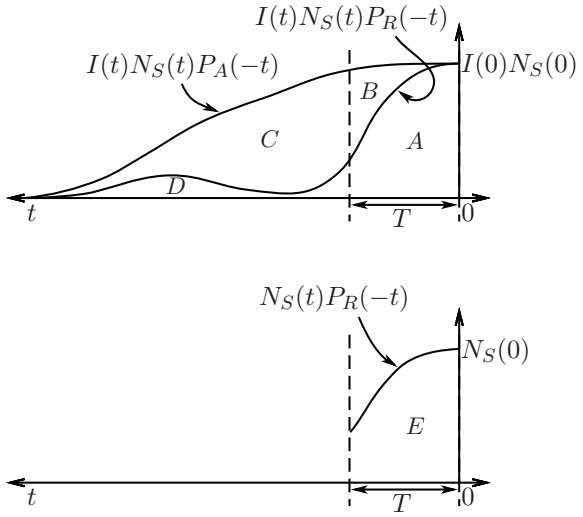


Figure 2.1: Epidemiological, demographic, and recent infection test dynamics
The figure portrays the functions (and areas under functions) of relevance for interpreting the weighted average of incidence that is measured in the cross-sectional survey, performed at time $t = 0$. The functions $I(t)$ and $N_S(t)$ are the incidence and susceptible population size at time t, respectively; $P_A(\tau)$ is the probability that a subject is alive at time τ after infection; and $P_R(\tau)$ is the probability that a subject is alive and 'recently' infected at time τ after infection.

In the interpretation of Equation (2.1), it is useful to consider the labelled areas in Figure 2.1:

- Areas A and D (the areas under the curve $I(t)N_S(t)P_R(-t)$) represent the 'recently' infected population at $t = 0$, infected for times less than and greater than T, respectively. A is the numerator of Equation (2.1).
- Areas B and C (the areas between the curves $I(t)N_S(t)P_A(-t)$ and $I(t)N_S(t)P_R(-t)$, that is, the infected population excluding the 'recently' infected population) represent the 'non-recently' infected population at $t = 0$, infected for times less than and greater than T, respectively.
- Area E (the area under the curve $N_S(t)P_R(-t)$) is the denominator of Equation (2.1).

Hence, Equation (2.1) can be rewritten as

$$I_T = \frac{A}{E}. \tag{2.2}$$

Previous derivations of incidence estimators have relied on assumptions of (at least recent) epidemiological and demographic equilibrium, some simplifications of the post-infection dynamics of tests for recent infection, or both. To elucidate the impact of general non-equilibrium conditions, it is useful to express the crucial time-dependent quantities in terms of time-dependent relative deviations from conveniently chosen constants:

$$I(t) = I_T \cdot \big(1 + f_I(t)\big) \tag{2.3}$$

$$N_S(t) = N_S(0) \cdot \Big(1 + f_{N_S}(t)\Big) \tag{2.4}$$

$$P_A(\tau) = 1 + f_{P_A}(\tau). \tag{2.5}$$

These equations do not represent any approximations or truncated power series, but are general, exact decompositions of $I(t)$, $N_S(t)$ and $P_A(\tau)$, the point of which is to characterise, rather than assume away, the non-ideal aspects of the population and test dynamics. More specifically: $f_I(t)$ is defined to capture the time dependence of the fractional deviation of incidence relative to the weighted incidence as defined in the period preceding $t = 0$; $f_{N_S}(t)$ is defined to capture the time dependence of the fractional deviation of the susceptible population from its instantaneous value at $t = 0$; and $-f_{P_A}(\tau)$ is defined as the probability of not surviving for at least a time τ after infection. The averages, over the period of duration T preceding the survey, of products of $f_I(.)$, $f_{N_S}(.)$ and $f_{P_A}(.)$, will be shown to summarise the effect of these deviations on incidence estimates.

Particular forms for an incidence estimator

Incidence estimation involves expressing Equation (2.2) as a function of test characteristic parameters, and population states. The key link between the numerator and the population states is

$$A = N_R - D, \tag{2.6}$$

where N_R is the size of the 'recently' infected population at $t = 0$ (Figure 2.1). Various authors have used different terms for the situation that $D \neq 0$ (Figure 2.1). These terms have included 'misclassification', 'false-positivity', 'false-recency', 'imperfect long-term specificity', the 'long tail' or 'non-progression' of the test for recent infection [16-21, 23, 25, 146]. Formally, the increasingly used term **false-recent rate**, β_T, given a cut-off T, is defined in this work as the probability that a randomly chosen person infected for longer than time T will be classified as 'recently' infected by the recent infection test.

A variety of approaches may be used for estimating the area D. For example,

$$\begin{aligned} D &= \frac{D}{C+D} \cdot (C+D) \\ &= \beta_T \cdot \big(N_+ - (A+B)\big), \end{aligned} \tag{2.7}$$

where

$$\beta_T = \frac{D}{C+D}, \tag{2.8}$$

and the areas $A+B$ and $C+D$ represent the current population infected for times less than and greater than T, respectively, so that $A+B+C+D=N_+$ is the size of the HIV-positive population.

By inspection of Figure 2.1,

$$A+B = \int_{-T}^{0} I(t) N_S(t) P_A(-t)\, \mathrm{d}t. \tag{2.9}$$

Using the parameterisation in terms of the dimensionless $f_I(.)$, $f_{N_S}(.)$, and $f_{P_A}(.)$ introduced earlier, this gives:

$$\begin{aligned} A+B &= \int_{-T}^{0} I_T N_S(0)\big(1+f_I(t)\big)\Big(1+f_{N_S}(t)\Big)\Big(1+f_{P_A}(-t)\Big)\,\mathrm{d}t \\ &= I_T N_S(0) T \cdot (1+\gamma_1+\gamma_2+\gamma_3+\gamma_4+\gamma_5+\gamma_6+\gamma_7), \end{aligned} \tag{2.10}$$

where the (also dimensionless) γ terms capture the consequences of the time-dependence of $I(t)$, $N_S(t)$ and $P_A(-t)$:

$$\gamma_1 = \frac{1}{T}\int_{-T}^{0} f_I(t)\,\mathrm{d}t \tag{2.11}$$

$$\gamma_2 = \frac{1}{T}\int_{-T}^{0} f_{N_S}(t)\,\mathrm{d}t \tag{2.12}$$

$$\gamma_3 = \frac{1}{T}\int_{-T}^{0} f_{P_A}(-t)\,\mathrm{d}t \tag{2.13}$$

$$\gamma_4 = \frac{1}{T}\int_{-T}^{0} f_I(t) f_{N_S}(t)\,\mathrm{d}t \tag{2.14}$$

$$\gamma_5 = \frac{1}{T}\int_{-T}^{0} f_I(t) f_{P_A}(-t)\,\mathrm{d}t \tag{2.15}$$

$$\gamma_6 = \frac{1}{T}\int_{-T}^{0} f_{N_S}(t) f_{P_A}(-t)\,\mathrm{d}t \tag{2.16}$$

$$\gamma_7 = \frac{1}{T}\int_{-T}^{0} f_I(t) f_{N_S}(t) f_{P_A}(-t)\,\mathrm{d}t. \tag{2.17}$$

These γ corrections may be positive or negative, but γ_3 is always non-positive.

Substituting into Equation (2.6) the expression for area D given by Equation (2.7), and then the expression for area $A + B$ given by Equation (2.10), the numerator becomes:

$$A = N_R - \beta_T N_+ + \beta_T I_T N_S(0) T \cdot \left(1 + \sum_{k=1}^{7} \gamma_k\right). \tag{2.18}$$

In the denominator of Equation (2.2),

$$\begin{aligned} E &= \int_{-T}^{0} N_S(t) P_R(-t)\,\mathrm{d}t \\ &= \int_{-T}^{0} N_S(0)\left(1 + f_{N_S}(t)\right) P_R(-t)\,\mathrm{d}t \\ &= N_S(0)\Omega_T \cdot (1 + \gamma_8), \end{aligned} \tag{2.19}$$

where

$$\begin{aligned} \Omega_T &= \int_{-T}^{0} P_R(-t)\,\mathrm{d}t \\ &= \int_{0}^{T} P_R(t)\,\mathrm{d}t, \end{aligned} \tag{2.20}$$

and

$$\gamma_8 = \frac{1}{\Omega_T}\int_{-T}^{0} f_{N_S}(t) P_R(-t)\,\mathrm{d}t. \tag{2.21}$$

The **mean duration of recent infection**, Ω_T, thus defined, given a cut-off T, is the average time spent both alive and 'recently' infected, within a time T post infection.

Substituting into Equation (2.2) the expressions for the numerator, given by Equation (2.18), and the denominator, given by Equation (2.19), gives:

$$I_T = \frac{N_R - \beta_T N_+ + \beta_T I_T N_S T \cdot (1 + \sum_{k=1}^{7} \gamma_k)}{N_S \Omega_T \cdot (1 + \gamma_8)}. \tag{2.22}$$

The right-hand side of this expression contains not only the mean duration of recent infection, Ω_T; the false-recent rate, β_T; and the uninfected, 'recently' infected and 'non-recently' infected populations at $t = 0$, $N_S = N_S(0)$, N_R and N_{NR}, respectively, where

$N_+ = N_R + N_{NR}$; but also the weighted incidence, I_T, itself. Rearranging and solving for I_T yields

$$I_T = \frac{N_R - \beta_T N_+}{N_S \cdot (\Omega_T - \beta_T T)} \cdot (1 + e)^{-1}, \tag{2.23}$$

where the equation

$$e = \left(\frac{\Omega_T}{\Omega_T - \beta_T T}\right)\gamma_8 - \beta_T \cdot \left(\frac{T}{\Omega_T - \beta_T T}\right)\sum_{k=1}^{7}\gamma_k \tag{2.24}$$

contains all the details that cannot be directly evaluated from an experimenter's point of view.

Using the sample counts of uninfected, 'recently' infected and 'non-recently' infected subjects at $t = 0$, n_S, n_R and n_{NR}, respectively, where $n_+ = n_R + n_{NR}$, a simple estimator of weighted incidence, with relative (fractional) error e, is obtained:

$$\hat{I}_T = \frac{n_R - \beta_T n_+}{n_S \cdot (\Omega_T - \beta_T T)}. \tag{2.25}$$

By using definitions of the test characteristics (Ω_T and β_T), which are subtly different from those used previously [11, 16, 17, 24, 25, 146], an incidence estimator is thus obtained in which multiple transitions between 'recently' and 'non-recently' infected states are allowed, and no assumption is required about the independence of progression through the 'recent'/'non-recent' states and post-infection survival (see [23] for a comparison of previously proposed incidence estimators). This estimator caters to completely general recent infection test dynamics. Bias arising from a non-constant incidence or susceptible population (in the period T before the incidence study) or imperfect survival (for T after infection) is fully described by e, and further discussed below.

The functional form for the estimator in Equation (2.25) can be obtained directly by assuming the system is in demographic and epidemiological equilibrium [146]. The present analysis shows that, when the system is away from equilibrium, in particular when incidence is not close to constant, this functional form provides an estimate of a particular weighted average of recent incidence, with a fractional bias e. In Section 2.2.1, the structure and meaning of the terms in Equation (2.24) for e are discussed, and the bias is computed in model scenarios.

The γ_8 term, closely related to a bias implicit in all previously proposed estimators [23, 25], is zero when the susceptible population is constant for T preceding the survey, but a time-dependent susceptible population imposes a fundamental limitation to cross-sectional incidence estimation. This highlights a key motivation for introducing T – namely to decouple the short-term dynamics of the test for recent infection from any long-term dynamics (which become convolved with the epidemiology and demography).

The remaining γ terms appear only in conjunction with two further multiplicative factors: (i) the fraction, dominated by T and Ω_T, which would perhaps typically have a value close to two, and (ii) a factor of β_T. Therefore the estimator can yield a weighted incidence as accurate as desired if β_T is sufficiently small, even when incidence and survival are varying substantially over the timescale set by T. It is already well known that informative incidence estimation requires that β_T be small [10, 11, 25, 147], and developers of tests for recent infection are seeking new technologies and algorithms to achieve this [86, 95].

Ultimately, the utility of a test for recent infection lies in its ability to produce accurate and precise incidence estimates. The expectation value and variance of the incidence estimator are approximated in Section 2.2.2. The uncertainty in the estimator, and its dependence on the test characteristics, is context-specific, depending on the history of HIV incidence and prevalence in the study population. The precision of the estimator improves with increasing mean duration of recent infection and with decreasing false-recent rate. This trade-off has been previously noted [11], with the additional subtlety in the present analysis that the choice of T and the test characteristics are intrinsically related.[4]

As noted in the introduction, the choice of T is theoretically arbitrary, but given the dynamics of available and foreseeable tests for recent infection, it will need to be around a year (or more). This implies a 'shadow' [15, 27] of about (but probably comfortably less than) half a year. This is considerably smaller than 'shadows' substantially exceeding one year, as obtained for a number of realistic scenarios considered by Brookmeyer [27]

[4] The use of the precision of the incidence estimator, as a standard metric for assessing the trade-off between the false-recent rate and mean duration of recent infection, was formally outlined in later work, and is discussed in Chapter 6.

when using the original one-parameter incidence estimator [7], and indeed implies less temporal bias or blurring than incurred in a cohort followed up for one year.

Estimation of test characteristics

As with any method aiming to infer incidence from the cross-sectional application of a recent infection test, use of the newly derived estimator requires measuring some characteristics of the test ahead of its application in the surveillance context. This test characterisation should be performed as locally as feasible, because test performance may be context-specific [131]. The false-recent rate, β_T, and the mean duration of recent infection, Ω_T, are intuitively close to previously proposed definitions [11, 16, 17, 24, 25, 146]. However, the definitions of the test characteristics emerging from this work allow, for the first time, arbitrary and complex test dynamics to be exactly captured. The estimation of each of the characteristics is briefly discussed below, with a slightly more technical discussion provided in Section 2.2.3.

The false-recent rate, β_T, would ideally be estimated by the proportion of 'recently' infected subjects in a representative sample of individuals infected for longer than T. It is also conceivable that β_T could be estimated from a combination of convenience samples, knowledge of the dynamics of anomalous subpopulations (who persist in, or return to, the 'recent' state despite being infected for a time greater than T) and knowledge of the embedding demography and epidemiology.

The mean duration of recent infection emerges as naturally in longitudinal surveillance settings (where well-pedigreed biological specimens may be obtained repeatedly over time) as it does in the context of cross-sectional incidence estimation analysed above. An idealised experiment, which revisits initially HIV-negative persons after a time equal to the post-infection time cut-off T, and counts the frequency of 'recent' results in those who have become HIV-positive, provides a direct estimate for Ω_T, assuming a uniform distribution of infection times over the inter-test interval (and negligible mortality within T post infection).[5] Specifically, the ratio Ω_T/T is the probability that a seroconverter is 'recently' infected. This idea can be expanded to account for varying inter-test intervals,

[5] The methodology for estimating the mean duration of recent infection using only specimens drawn at subjects' first HIV-positive visits, and when inter-test intervals may be large, was formalised and demonstrated in earlier work, which is presented in Chapter 3.

depending on available data and knowledge of the dynamics of the test for recent infection, with an example of such an extension provided in Section 2.2.3.

More traditionally, measurement of the mean duration of recent infection has been based on the frequent follow-up and recent infection testing of seroconverters. A form of survival analysis or regression can then be used to characterise the time taken to exit the 'recent' state or the evolution of the biomarker over time after infection, respectively, thereby estimating the mean duration of recent infection.[6]

Demonstration of methodology using simulated data

Having derived this new, general incidence estimator, and having outlined potential approaches for estimating the required test characteristics, implementation of the full set of analyses to infer incidence is demonstrated using simulated data.

Assuming a particular epidemiological and demographic history, post-infection survival function, and dynamic for the test for recent infection, one thousand simulations were performed, each producing (independent) datasets to (i) estimate the false-recent rate, β_T; (ii) estimate the mean duration of recent infection, Ω_T; and (iii) provide sample counts to infer incidence, I_T, using the incidence estimator in Equation (2.25).

The generation of the datasets and the maximum likelihood estimation of the test characteristics are described in Section 2.2.4. Asymptotic normality of maximum likelihood estimators (using estimated characteristics as proxies for true values) was used to approximate distributions for the estimated parameters and to obtain confidence intervals. Confidence intervals for incidence were then based on these results, the approximate normality of the trinomial sample counts (with sample statistics approximating population parameters), and the approximate normality of the incidence estimator and its estimated variance as provided in Section 2.2.2. As summarised in Table 2.1, almost exactly 95% of the one thousand thus generated 95% confidence intervals (one for each of the datasets) contained the relevant population parameter used in the simulation, demonstrating the numerical consistency of the full set of analyses.

---

[6] Methods for estimating the mean duration of recent infection from longitudinal data obtained by following (initially HIV-negative) subjects over time were systematically explored in later work, and results are presented in Chapter 4.

Parameter	Input value[i]	Average point estimate	Average CI width[ii]	CI coverage[iii]
$\boldsymbol{\beta_T}$	2.5%	2.52%	1.93%	95.5%
$\boldsymbol{\Omega_T}$	160 days	160.56 days	17.96 days	93.6%
$\boldsymbol{I_T}$	2%	1.98%	1.38%	94.6%

[i] The true parameter value, as input into the data simulator
[ii] Average width of realised 95% confidence intervals
[iii] Percentage of realised 95% confidence intervals containing the true parameter value

Table 2.1: Observed 95% confidence interval (CI) coverage of parameters using simulated data
Simulated datasets were used to validate the methodologies presented for estimating test characteristics and incidence. For each parameter, namely the false-recent rate β_T (%), mean duration of recent infection Ω_T (days) and incidence I_T (% per annum), 1 000 point estimates and confidence intervals were obtained (the chosen modelled scenario and estimation methods are described in the main text above and Section 2.2.4), for $T = 450$ days. The true parameter value, average point estimate, average 95% confidence interval (CI) width and CI coverage (%) are tabulated.

2.1.3 Discussion

The use of tests for recent infection to infer incidence is of considerable and increasing interest, especially for HIV surveillance. It is a fundamental limitation that all currently available (and perhaps all conceivable) tests with a mean duration of recent infection that is long enough for statistical robustness also classify some individuals as 'recently' infected at arbitrarily large times after infection. If there were no such 'false-recent' results, the use of recent infection tests for incidence estimation would be straightforward, as shown, for example, by Brookmeyer and Quinn [7]. Various methodological advances to accommodate a non-zero 'false-recent rate' have attracted attention, but consensus has not emerged on the best approach. Previous derivations of incidence estimators have relied on strong assumptions: perhaps most crucially that the 'false-recent rate' is an innate property of the test, rather than a convolution of test properties with the demographic and epidemiological context. This assumption is known to be substantially violated.

A formal approach is presented above for summarising an arbitrarily complicated recent infection test dynamic into two parameters, namely a mean duration of recent infection and a false-recent rate. A crucial construct is the introduction of a timescale T, describing the dynamic range of 'recent' infection. The consequence of relaxing the assumptions made by the incidence estimators developed previously is that demography and epidemiology are no longer perfectly separated from test characteristics, reflecting fundamental limitations to the inference of rates from instantaneous population states. If the false-recent rate is very close to zero, the limitations imposed by a non-zero false-recent rate become minor and its variation over time and place is restricted.

The present analysis offers the opportunity to consistently account for imperfect accuracy and precision of the incidence estimator. The utility of the estimator may be assessed in terms of changes in incidence and the susceptible population over the preceding period of duration T, the probability of survival over T post infection, and the characteristics of the recent infection test. The cross-sectional incidence estimator will be informative at feasible sample sizes, in a given context, for a suitably well-behaved test, that is, a test with a suitably long mean duration of recent infection and low false-recent rate.

The approach presented here is broad enough to recover previously-proposed estimators, with minor modification. It also clarifies the use of estimators that do not account for 'false-recent' results at all. Setting T to a very large value forces the false-recent rate arbitrarily close to zero, and the one-parameter estimator of Brookmeyer and Quinn [7] is obtained. The properties of the test are then summarised by the mean duration of recent infection. However, this mean duration, which is now the average time spent alive and 'recently' infected, is considerably more difficult to measure and more likely to change over time than one based on individual durations that are each explicitly limited to T. Also, a large T (effectively infinite, if T is not explicitly introduced) leads to a weighting scheme for averaging incidence that extends far back into the past. For heuristic purposes, the weighted incidence that emerges from the use of a realistically available test and a judicious choice of T can be viewed as a good proxy for the uniformly-weighted mean incidence in the period of duration Ω_T preceding the survey. One may consider whether there is any benefit in using additional parameters to characterise the dynamics of the test for recent infection [97]. Incidence inference would then be based on a more complex distribution of test results than counts of 'recent' and 'non-recent' cases. Section 2.2.4 presents a brief argument that suggests this approach has limited prospects.

Relaxing all formal assumptions about the dynamics of a putative test for recent infection and the demographic and epidemiological context leads to an estimator that substantially increases the robustness of incidence estimation based on cross-sectional surveys using tests for recent infection. The general analysis leads to a clearer characterisation of the utility of the estimator than previously possible. While the analysis is fundamentally novel, the resulting estimator has similarities to some previously published estimators [24, 25]. These similarities imply that, intuitively, the crucial concepts of a false-recent rate and mean duration of recent infection are substantially retained. Numerically, the improvement in incidence estimates implied by the new estimator will vary with context.

While the motivation for this work has been to improve our capacity to estimate HIV incidence, the methodology is general, and the approach could be applied to estimate incidence of other incurable conditions, such as herpes simplex virus. Future studies should examine the application of the methodology to a wider range of diseases, with the practical challenge being the development of suitable tests for recent infection.

2.1.4 Conclusion

For incurable conditions such as HIV, where prevalence emerges as a slow convolution of historic incidence with survival and the dynamics of the susceptible population, changes in prevalence are a poor proxy for recent incidence. Estimating incidence from cross-sectional surveys has many potential advantages over using longitudinal studies, and has attracted much interest in recent years, particular in the HIV context. However, previously proposed HIV incidence estimators have been derived under conditions of epidemiological and demographic equilibrium, or specific assumptions about the recent infection test dynamics, or both. These assumptions are known to be violated in many settings, and this has diminished the practical utility of previous methodologies.

In this article, biomarker-based incidence estimation, which uses data obtained in cross-sectional surveys, is consistently adapted to a general context. The generalisation implies that the strong assumptions about epidemiological and demographic history and biomarker dynamics required by previous estimators are no longer necessary for valid incidence estimation. Our new estimator thus substantially improves and clarifies the utility of tests for recent infection for estimating disease incidence. The familiar practical challenge remains – to make available ever better, and better characterised, tests for recent infection.

2.2 Ancillary Explorations of the Incidence Estimator

2.2.1 Bias of the Incidence Estimator

Bias arising from epidemiological and demographic properties of the population

In Section 2.1, the following explicit weighted average of incidence over the preceding period T was defined:

$$I_T = \frac{\int_{-T}^{0} I(t)N_S(t)P_R(-t)\,\mathrm{d}t}{\int_{-T}^{0} N_S(t)P_R(-t)\,\mathrm{d}t}, \tag{2.26}$$

where the, possibly time-dependent, incidence is given by $I(t)$, the susceptible population by $N_S(t)$, and $P_R(\tau)$ is the probability of being alive and 'recent' at time τ post infection.

This weighted incidence was shown to be:

$$I_T = \frac{N_R - \beta_T N_+}{N_S \cdot (\Omega_T - \beta_T T)} \cdot \left(1 + \left(\frac{\Omega_T}{\Omega_T - \beta_T T}\right)\gamma_8 - \beta_T \cdot \left(\frac{T}{\Omega_T - \beta_T T}\right)\sum_{k=1}^{7}\gamma_k\right)^{-1}, \tag{2.27}$$

where the population at $t = 0$ is decomposed into

- N_R = 'recently' infected population,
- N_{NR} = 'non-recently' infected population, and
- N_S = uninfected or susceptible population;

the characteristics of the test for recent infection is captured by

- the false-recent rate, β_T, which is the proportion of 'recently' infected individuals among individuals infected for times greater than T, and
- the mean duration of recent infection, $\Omega_T = \int_0^T P_R(t)\,\mathrm{d}t$, which is the average time alive and returning a 'recent' result, while infected for times less than T;

the subscripted γ terms capture non-equilibrium epidemiological/demographic conditions for a period of duration T preceding the cross-sectional survey (performed at $t = 0$) and imperfect survival until T post infection,

- $\gamma_1 = \frac{1}{T}\int_{-T}^{0} f_I(t)\,\mathrm{d}t$,
- $\gamma_2 = \frac{1}{T}\int_{-T}^{0} f_{N_S}(t)\,\mathrm{d}t$,

- $\gamma_3 = \frac{1}{T}\int_{-T}^{0} f_{P_A}(-t)\,\mathrm{d}t$,
- $\gamma_4 = \frac{1}{T}\int_{-T}^{0} f_I(t) f_{N_S}(t)\,\mathrm{d}t$,
- $\gamma_5 = \frac{1}{T}\int_{-T}^{0} f_I(t) f_{P_A}(-t)\,\mathrm{d}t$,
- $\gamma_6 = \frac{1}{T}\int_{-T}^{0} f_{N_S}(t) f_{P_A}(-t)\,\mathrm{d}t$,
- $\gamma_7 = \frac{1}{T}\int_{-T}^{0} f_I(t) f_{N_S}(t) f_{P_A}(-t)\,\mathrm{d}t$, and
- $\gamma_8 = \frac{1}{\Omega_T}\int_{-T}^{0} f_{N_S}(t) P_R(-t)\,dt$,

where $P_A(\tau)$ is the probability of being alive at time τ post infection; and the time-dependencies of $I(t)$, $N_S(t)$ and $P_A(\tau)$ are captured by the deviations $f_I(t)$, $f_{N_S}(t)$ and $f_{P_A}(\tau)$, respectively,

- $I(t) = I_T \cdot \big(1 + f_I(t)\big)$
- $N_S(t) = N_S(0) \cdot \Big(1 + f_{N_S}(t)\Big)$, and
- $P_A(\tau) = 1 + f_{P_A}(\tau)$.

The relative bias of the estimator from the exact weighted incidence is therefore

$$e = \left(\frac{\Omega_T}{\Omega_T - \beta_T T}\right)\gamma_8 - \beta_T \cdot \left(\frac{T}{\Omega_T - \beta_T T}\right)\sum_{k=1}^{7}\gamma_k. \tag{2.28}$$

A general discussion of the structure and the meaning of the terms that make up the relative bias, e, is now provided (elaborating on the discussion in the previous section), and the bias is quantified and explored in some model scenarios.

The γ_8 term is zero when the susceptible population is constant for a period of duration T preceding the survey. If the susceptible population is varying considerably, the incidence estimator would be substantially biased by the γ_8 term. For example, if the susceptible population is increasing (decreasing) over the preceding period T, the weighted incidence is expected to be underestimated (overestimated) by this demographic dynamic, although the overall net bias will also depend on changes in incidence, post-infection survival and the characteristics of the test for recent infection.

This potential for bias highlights a key motivation for introducing T, rather than defining recent duration simply through the function $P_R(t)$ – namely to decouple the short-term dynamics of the test for recent infection from any long-term dynamics which become convolved with the epidemiology/demography. Reducing T, and therefore averaging over a much shorter period, would tend to reduce the magnitude of the γ correction terms, but

ultimately erode the statistical power of incidence estimation by making the test-defined 'recent' state more transient and difficult to measure with confidence at realistic sample sizes.

The γ_8 and γ_2 terms are closely related, both representing (variously weighted) average deviations of the susceptible population over the preceding period of duration T from its current size, with γ_8 typically expected to be smaller in magnitude than γ_2.

The γ_k terms, $k = 1, 2, \ldots, 7$, appear only in conjunction with two further multiplicative factors: (i) the fraction, dominated by T and Ω_T, which perhaps typically has a value close to two, and (ii) a factor of β_T. Therefore, the estimator can yield a weighted incidence as accurate as desired if β_T is sufficiently small. For an ideal test for recent infection – that is, a test for which $P_R(t)$ reaches and remains at zero for t larger than the time over which incidence is to be averaged (a time much shorter than post-infection survival) – there is no need to explicitly introduce T into the analysis. In that case, the false-recent rate is effectively zero and a one-parameter incidence estimator is obtained. Since the γ_k terms ($k = 1, 2, \ldots, 7$) are then multiplied by the factor $\beta_T = 0$, a consistent estimate (unbiased in the limit of large sample size) of the weighted incidence is obtained in the case of a constant susceptible population.

Therefore, bias introduced by a varying incidence (for example, in a population with an outbreak of HIV cases, or experiencing a successful prevention intervention) or imperfect survival until T post infection (although likely to be negligible) may be suppressed by forcing the false-recent rate close to zero.

Under typical epidemiological conditions and in the applicable regimes of utility, the deviations $f_I(.)$, $f_{N_S}(.)$ and $f_{P_A}(.)$ are expected to be much less than 1. Therefore, γ_4 to γ_7 are expected to be smaller in *magnitude* than γ_1 to γ_3.

It is worth noting that the deviations, $f_I(.)$, $f_{N_S}(.)$ and $f_{P_A}(.)$, and resulting γ correction terms, playoff against one another in a complex way, and could together compound the bias or bring it cumulatively closer to zero (compared to the bias arising from any one deviation or γ term considered in isolation).

In the model scenarios described below, the susceptible population and incidence vary over the preceding period of duration T, and the exact relative bias was calculated. The estimation of the approximate magnitude of the bias, from the point of view of the

experimenter, is also briefly considered. The Matlab code that was produced for the investigation below is provided in Appendix B.1.

Two model scenarios (termed Scenarios 1 and 2) were considered, corresponding to increasing and decreasing incidence, respectively, during the period preceding the cross-sectional incidence estimation at $t = 0$ (with *t* measured in years):

1. Incidence increased linearly over the preceding year, from 1% to $I(0)$, where $I(0)$ was varied from 1% to 5% (corresponding to a percentage increase in incidence of 0% to 400% over the year).
2. Incidence decreased linearly over the preceding year, from $I(-1)$ to 1%, where $I(-1)$ was varied from 1% to 5% (corresponding to a percentage decrease in incidence of 0% to 80% over the year).

Incidence was constant prior to the specific changes noted.

In each scenario, the susceptible population changed exponentially over time, with the annual growth rate ranging from -10% to 10%, and a susceptible population of one million individuals at $t = 0$.

The timescale T was set to 1 year, and the following survival and test dynamics were assumed:

- Post-infection survival times followed a Weibull distribution, with a coefficient of variation of 50% and mean post-infection survival of 8 years.
- The probability of being 'recently' infected, conditional on being alive, linearly decreased from 100% to some *constant*, between 0.25 years and 0.75 years post infection, and remained at that *constant*.

The probability of being both alive *and* 'recently' infected, at time τ post infection, $P_R(\tau)$, equals the product of the probabilities of being (i) alive; and (ii) 'recently' infected, conditional on being alive; at time τ post infection. By construct, the false-recent rate, β_T, equals the *constant* probability of being 'recently' infected, conditional on being alive, for times post infection greater than $T = 1$ year. Scenarios 1 and 2 were each further split into two sub-scenarios, namely corresponding to a *low* false-recent rate of $\beta_T = 1\%$ (Scenarios 1A and 2A) and *high* false-recent rate of $\beta_T = 5\%$ (Scenarios 1B and 2B) respectively. Under the assumptions above, the mean duration of recent infection, Ω_T, is 184 days for Scenarios 1A and 2A, and 192 days for Scenarios 1B and 2B.

The population states at $t = 0$ (sizes of the uninfected, 'recently' infected and 'non-recently' infected populations, N_S, N_R and N_{NR}, respectively, where $N_+ = N_R + N_{NR}$) were calculated based on the described epidemiological/demographic history, survival, and test dynamics.

The relative bias, e, of the estimator from the true weighted incidence, is shown as a percentage in the contour plots in Figure 2.2, for each of the Scenarios 1A, 1B, 2A and 2B:

- Even in the extreme cases, of atypically large changes in the susceptible population and/or incidence, the expected value of the incidence estimator remains within about 5% of the weighted incidence (for example, the expected value lies within 1.9% and 2.1% per annum when the true weighted incidence is 2% per annum).
- The probability of dying within T post infection is less than 1%, and so imperfect survival contributes negligible bias, and, when the percentage annual growth in the susceptible population is 0% and the change in incidence is 0%, the bias is undetectable.
- When the susceptible population and/or incidence experience non-zero change, the γ corrections become non-zero and bias is introduced. For example, with a false-recent rate of 5% and increasing incidence (Scenario 1B), for a 200% growth in incidence (from 1% to 3%) and 2% annual growth in the susceptible population, the fractional error is 1.25%. While the exact weighted incidence is 2.42%, the expected estimated incidence is 2.45%.
- Since β_T is a factor that suppresses the bias from the non-zero γ terms, the biases in Scenarios 1A and 2A (the *low* false-recent rate scenarios) are expected to be generally smaller than those in Scenarios 1B and 2B (the *high* false-recent rate scenarios), respectively.
- The γ corrections playoff against one another to either build-up or reduce the cumulative bias, just as the deviations $f_I(.)$, $f_{N_S}(.)$ and $f_{P_A}(.)$ or their products average out in different ways to increase or decrease the magnitude of each γ term. Also, perhaps more subtly, $f_I(.)$ describes the time-dependent deviation of incidence from the exact weighted incidence, which itself depends on how incidence, the susceptible population, and the probability of being 'recently' infected and alive over time post infection *all* vary. This complicated interaction between all the crucial functions to produce the overall bias is evident in

Figure 2.2, where, even though the susceptible population and incidence are varying, there are contours of zero error.

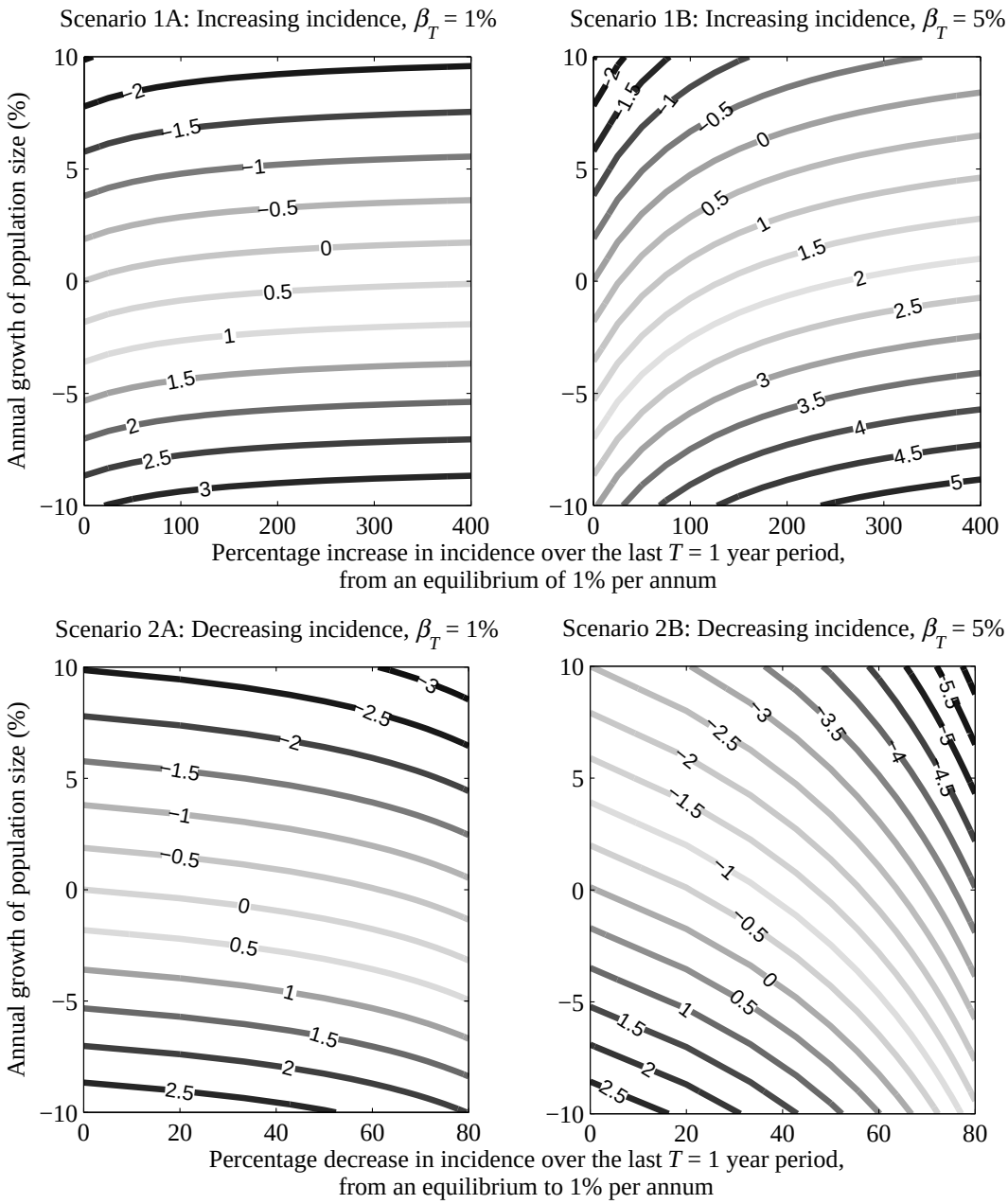


Figure 2.2: Relative bias of estimator from weighted incidence in modelled scenarios
The exact relative bias (%) of the incidence estimator is calculated for each of four model scenarios, as a function of susceptible population growth rate and change in incidence. In Scenarios 1A and 1B, incidence increases linearly over the preceding $T = 1$ year, beginning at 1% per annum, while in Scenarios 2A and 2B, incidence decreases linearly over that time to 1%. Post-infection survival follows a Weibull distribution (mean of 8 years, coefficient of variation of 50%), and the probability of testing 'recently' infected, conditional on being alive, decreases linearly from 100% to either 1 % (Scenarios 1A and 2A) or 5% (Scenarios 1B and 2B) between 0.25 years and 0.75 years post infection.

The experimenter may use more limited knowledge of the study population and its history to estimate the magnitude of bias of the incidence estimator, and to understand whether a study is in a regime of utility. There should be some data on whether there is potentially considerable time-dependence in the susceptible population or incidence, or notable post-infection mortality, in the study population, for the relevant timescale set by T:

- Should the experimenter believe that the susceptible population is varying substantially over the preceding period of duration T, the magnitude of γ_8 could be estimated, and this correction would likely dominate the bias (due to the remaining correction terms carrying a factor of β_T). The γ_2 term, also describing the average deviation of the susceptible population from its current size, would typically be larger in *magnitude* than γ_8, and therefore provides an indication of the size of the γ_8 term. For example, for $T = 1$ year, estimating a 2% annual growth rate in the population, the experimenter may consider 0.01 to be a conservative estimate of the magnitude of γ_2 ($\gamma_2 = -0.0098$ for a population growing exponentially at 2% per annum).
- The γ_1 correction would typically be small as it measures the uniformly averaged deviation of incidence, over the preceding T, from an alternatively weighted (by $N_S(t)P_R(-t)$) average incidence over the same period. Estimating a large 200% increase in incidence over the last year, the experimenter may consider 0.2 to be a conservative estimate of the magnitude of γ_1 ($\gamma_1 = -0.1725$ for a 200% increase in incidence, from 1% to 3%, and a 2% annual growth in the susceptible population, in Scenario 1B).
- Given multiplication of the γ_8 term by a factor of around unity in the relative error, e, and multiplication of the γ_1 term by two factors, one close in value to two and the other equalling β_T, an initial estimate of the magnitude of relative bias, assuming negligible $T = 1$ year post-infection mortality, could be $0.01 + 2 \times 5\% \times 0.2 = 0.03$ for a false-recent rate of approximately 5%. This is conservative compared to the exact relative error calculated in Scenario 1B, of magnitude 0.0125.

The estimation of the magnitude of bias, by the experimenter, could be tackled with increasing sophistication and detail, depending on the available data – although probably with diminishing returns beyond an elementary calculation such as outlined. The ad-hoc example demonstrates that the experimenter should reasonably be in a position to understand the magnitude of bias and relate the regimes of utility to the current analysis,

keeping in mind the inherently imperfect reproducibility in the incidence estimator (approximated in Section 2.2.2). In short, whenever cross-sectional incidence estimation is at all informative, bias will be small compared to variance.

Other factors impacting the incidence weighting function

The exact weighting function, $N_S(t)P_R(-t)$, is never exactly known, as it involves complete knowledge of the dynamics of the test for recent infection, rather than merely a mean duration of recent infection. For practical purposes, the estimator of this work is a useful proxy for a uniformly-weighted mean incidence over the preceding period Ω_T.

Sensitivity and specificity of the 'diagnostic' test used to identify HIV-positive individuals may be imperfect. In particular, sensitivity will change rapidly from zero to a value close to one over a short period post exposure, and thus be highly correlated with status on the recent infection test. In the limit that sensitivity approaches unity at some time post exposure, even when incidence is varying, the analysis remains consistent if the weighting $N_S(t)P_R(-t)$ is understood to apply to the rate of detectable infection events – which differs from the weighting of underlying exposure events by an epidemiologically irrelevant delay – and requires only consistent estimation of the mean duration of recent infection.[7] If there is correlation between diagnostic delay and the duration of 'recent' infection, this delay may involve a slight time-dependent blurring, rather than a pure time translation, but none of this changes the previously noted heuristic that the estimator is an excellent proxy for the uniformly-weighted mean incidence over the period Ω_T preceding the survey.

In HIV diagnostics, levels of sensitivity and specificity are exceptionally high, but if indeed this analysis were to be contemplated for a context where there is substantially imperfect diagnostic performance over the full lifetime post infection, and non-trivial correlation with status on the recent infection test, it may warrant further investigation to see whether the diagnostic is preferentially misclassifying individuals in the 'recent' versus 'non-recent' categories.

---

[7] Estimation of the mean duration of recent infection, using a definition of 'infection time' that is consistent with the HIV diagnostic algorithm used in the surveillance survey, is further discussed in Section 4.3.

If the field work for a nominally 'cross-sectional' survey takes an extended period of time, comparable to Ω_T, this would introduce additional temporal blurring of the estimate. In the limit of fieldwork carried out over a much longer time than Ω_T, the resulting incidence estimate essentially averages incidence over the period of fieldwork rather than over the duration of 'recent' infection.

2.2.2 Moments of the Incidence Estimator

Using the delta method [148] together with assumptions of Gaussian uncertainty in the sample counts and estimated characteristics of the test for recent infection, the moments of the incidence estimator are approximated (see Appendix A.1 for a summary of the delta method and its application in this work). The key results are listed, and the assumptions and derivation provided thereafter.

By Equation (2.25), the incidence estimator is:

$$\hat{I}_T = \frac{n_R - \beta_T n_+}{n_S \cdot (\Omega_T - \beta_T T)}, \tag{2.29}$$

where n_S, n_R and n_{NR} are the counts of uninfected, 'recently' infected and 'non-recently' infected individuals in the cross-sectional sample at $t = 0$, $n_+ = n_R + n_{NR}$, and the characteristics of the test for recent infection, namely the false-recent rate, β_T, and mean duration of recent infection, Ω_T, would need to be estimated. The test characteristics are typically estimated in separate studies conducted prior to the cross-sectional incidence study, though a combined study design, applicable under more restrictive assumptions, has been proposed [93, 149].

Consider the problem of calculating the mean and the variance of the incidence estimator, for a cross-sectional survey of sample size $n = n_S + n_R + n_{NR}$. Let P_S, P_R and P_{NR} be the true population proportions of uninfected, 'recently' infected and 'non-recently' infected individuals. Let $\sigma_{\hat{\Omega}_T}$ be the standard deviation of the unbiased estimator for the mean duration of recent infection Ω_T, and $\sigma_{\hat{\beta}_T}$ be the standard deviation of the unbiased estimator for the false-recent rate β_T. Then the expected value of the incidence estimator is

$$E\left[\hat{I}_T\right] \approx \frac{P_R - \beta_T P_+}{P_S \cdot (\Omega_T - \beta_T T)} \tag{2.30}$$

and the variance is

$$
\begin{aligned}
var(\hat{I}_T) \approx & \left(\frac{P_R - \beta_T P_+}{P_S \cdot (\Omega_T - \beta_T T)}\right)^2 \times \\
& \left[\frac{1}{n \cdot (P_R + P_{NR})}\left(\frac{1}{P_S} + \frac{P_R P_{NR}}{(P_R - \beta_T \cdot (P_R + P_{NR}))^2}\right)\right. \\
& + \sigma^2_{\hat{\Omega}_T} \cdot \left(\frac{1}{\Omega_T - \beta_T T}\right)^2 \\
& \left. + \sigma^2_{\hat{\beta}_T} \cdot \left(\frac{P_{NR}\Omega_T - P_R \cdot (T - \Omega_T)}{\big(P_R - \beta_T \cdot (P_R + P_{NR})\big)(\Omega_T - \beta_T T)}\right)^2\right].
\end{aligned}
\tag{2.31}
$$

The coefficient of variation (ratio of standard deviation to mean) of the incidence estimator is c, where:

$$
\begin{aligned}
c^2 \approx & \frac{1}{n \cdot (P_R + P_{NR})}\left(\frac{1}{P_S} + \frac{P_R P_{NR}}{(P_R - \beta_T \cdot (P_R + P_{NR}))^2}\right) \\
& + \sigma^2_{\hat{\Omega}_T} \cdot \left(\frac{1}{\Omega_T - \beta_T T}\right)^2 \\
& + \sigma^2_{\hat{\beta}_T} \cdot \left(\frac{P_{NR}\Omega_T - P_R \cdot (T - \Omega_T)}{\big(P_R - \beta_T \cdot (P_R + P_{NR})\big)(\Omega_T - \beta_T T)}\right)^2.
\end{aligned}
\tag{2.32}
$$

From this, it is possible to formally explore the trade-off between test characteristics: test developers need to find biomarkers for recent infection which have a suitably large mean duration of recent infection and a suitably small false-recent rate.[8] Note that the variability of the incidence estimator inflates as the terms $(\Omega_T - \beta_T T)$ and $\big(P_R - \beta_T \cdot (P_R + P_{NR})\big)$, which occur in denominators above, become small. Hence, definitions of recent infection, such as being antibody negative while viral RNA or p24 antigen positive (commonly referred to as 'acute infection') are of little use for incidence estimation at the population level as the mean duration of recent infection is too short.

[8] The use of the precision of the incidence estimator, as a summary metric for formally assessing the trade-off between the false-recent rate and mean duration of recent infection, was further explored in later work, and is presented in Chapter 6.

The above results were derived by first assuming Gaussian uncertainty in both the sample counts (of uninfected, 'recently' infected and 'non-recently' infected subjects, n_S, n_R and n_{NR}, respectively) and estimated test characteristics (estimated mean duration of recent infection, $\widehat{\Omega}_T$, and estimated false-recent rate, $\hat{\beta}_T$). This allows the counts and estimated parameters to be expressed as:

- $n_S = n_s(\alpha_1) = nP_S + \sigma_S\alpha_1$,
- $n_R = n_R(\alpha_1, \alpha_2) = nP_R - \sigma_R\alpha_1 + \sigma_{R,NR}(\alpha_1)\alpha_2$,
- $n_{NR} = n_{NR}(\alpha_1, \alpha_2) = nP_{NR} - \sigma_{NR}\alpha_1 - \sigma_{R,NR}(\alpha_1)\alpha_2$,
- $\widehat{\Omega}_T = \widehat{\Omega}_T(\alpha_3) = \Omega_T + \sigma_{\widehat{\Omega}_T}\alpha_3$, and
- $\hat{\beta}_T = \hat{\beta}_T(\alpha_4) = \beta_T + \sigma_{\hat{\beta}_T}\alpha_4$,

where

- $\sigma_S = \sqrt{nP_S \cdot (1 - P_S)}$,
- $\sigma_R = \frac{P_R}{P_R + P_{NR}}\sigma_S$,
- $\sigma_{NR} = \frac{P_{NR}}{P_R + P_{NR}}\sigma_S$,
- $\sigma_{R,NR}(\alpha) = \frac{\sqrt{(n - nP_S - \sigma_S\alpha)P_RP_{NR}}}{P_R + P_{NR}}$,

and α_1, α_2, α_3 and α_4 are identically and independently distributed standard normal random variables.

The incidence estimator was then expressed as a function of the α_i ($i = 1,2,3,4$), and a multivariate Taylor series expansion around $\alpha_i = 0$ ($i = 1,2,3,4$) was constructed. Taking the expected value and variance of this series, retaining powers of α_i of up to 1, led to the results in Equations (2.30) to (2.32).

The approximation and distributional assumptions are highly accurate in their handling of counting error when characteristics of the test for recent infection are known. This was verified numerically: assuming a constant historical incidence and a constant prevalence to incidence ratio of HIV, as well as known test characteristics, the population proportions were calculated and the coefficient of variation of the incidence estimator computed by directly enumerating all possible trinomially distributed survey counts. Even with a small sample of $n = 100$, the maximum error in the coefficient of variation obtained using the delta method (that is, the absolute difference between the approximated and actual coefficient of variation) is 0.017 for all combinations of incidence in [0.1%,3%], prevalence to incidence ratio in [2,10], Ω_T in [100,300] days and

β_T in [0%,15%]. Matlab code to perform this investigation is provided in Appendix B.2, and could be used to explore other regimes of interest.

The assumption of Gaussian uncertainty in the estimated test characteristics is heuristic, and it is difficult to ascertain the accuracy of the approximation provided. When the false-recent rate is very close to zero, it is likely that a normal distribution will provide a poor approximation for the distribution of its estimator, and more sophisticated methods of error propagation should be investigated.

2.2.3 Estimation of Test Characteristics

The parameters that have been identified to describe the characteristics of the test for recent infection required to infer incidence (namely, the false-recent rate, β_T, and the mean duration of recent infection, Ω_T) are intuitively close to those previously proposed. The estimation of the characteristics, demonstrating their emergence in longitudinal surveillance settings, is discussed below.

The false-recent rate

The false-recent rate, β_T, is the proportion of 'recently' infected individuals among individuals infected for a time greater than T.

Therefore, the binomial maximum likelihood estimator for β_T is:

$$\hat{\beta}_T = \frac{m_R}{m}, \tag{2.33}$$

where m_R is the number of 'recently' infected individuals in a representative sample of m individuals infected for longer than T.

It would also be possible, although probably more challenging, to estimate β_T from a combination of convenience samples, knowledge of the dynamics of anomalous subpopulations (who persist in or return to the 'recent' state despite being infected for a time greater than T) and knowledge of the embedding demography/epidemiology.

The mean duration of recent infection

The mean duration of recent infection, Ω_T, is the average time spent alive and 'recently' infected while infected for times less than T.

In an idealised experiment, which revisits initially HIV-negative individuals after a time equal to the post-infection time cut-off T and counts the frequency of 'recent' results in those who have become HIV-positive, a direct estimate for Ω_T is provided, assuming a uniform distribution of infection times over the inter-test interval of duration T (and negligible mortality within T post infection).[9] More specifically, the ratio Ω_T/T is the probability of a seroconverter providing a 'recent' result on the first HIV-positive test, and therefore the binomial maximum likelihood estimator for Ω_T is:

$$\widehat{\Omega}_T = T\frac{k_R}{k}, \tag{2.34}$$

where k_R is the number of 'recently' infected subjects in the group of k subjects who are HIV-positive at follow-up. Such a study would probably need to be prohibitively large to capture a reasonably large sample of seroconverters, but it is worth noting that no additional input parameters are needed in the estimation, overcoming a key obstacle of unknown input parameters to the estimation of previously-defined mean durations of recent infection [31].

This idea can be further developed to account for *varying* inter-test intervals, depending on available data and knowledge of the dynamics of the test for recent infection. As an example, a method of maximum likelihood is outlined below for estimating Ω_T from data capturing recent infection test classifications at the times of the first HIV-positive tests (that is, when there is no follow-up of HIV-positive subjects), and there are varying but large intervals between last HIV-negative and first HIV-positive tests, utilising assumptions about test dynamics at times post infection in the vicinity of T.

[9] The estimation of the mean duration of recent infection using only specimens drawn at subjects' first HIV-positive visits, and when inter-test intervals may be large, was formalised and demonstrated in earlier work, which is presented in Chapter 3.

In general, the probability, p_i, that the i^{th} seroconverter, with inter-test interval Δ_i, returns a 'recent' result at the time of the first HIV-positive test is:

$$p_i = \int_{-\Delta_i}^{0} f(-t)P_R(-t)\,\mathrm{d}t, \tag{2.35}$$

where $f(\tau)$ is the probability density of getting infected at time τ before the first HIV-positive test and $P_R(\tau)$ becomes the probability of being 'recently' infected when tested at time τ ($< \Delta_i$) after infection, assuming no mortality for at least time Δ_i post infection. More specifically, for a sample of subjects who have inter-test intervals between L and U ($L \leq \Delta_i \leq U$), where L and U ($L \leq T \leq U$) are chosen so that the probability of testing 'recently' infected is constant at θ for times post infection between L and U, and assuming that infection times are uniformly distributed in inter-test intervals (that is, $f(-t) = 1/\Delta_i$ for all t such that $t \in [-\Delta_i, 0]$), p_i becomes:

$$\begin{aligned} p_i &= \frac{1}{\Delta_i} \cdot \left(\int_{-T}^{0} P_R(-t)\,\mathrm{d}t + \int_{-\Delta_i}^{-T} P_R(-t)\,\mathrm{d}t \right) \\ &= \frac{\Omega_T + \theta \cdot (\Delta_i - T)}{\Delta_i}. \end{aligned} \tag{2.36}$$

The likelihood of the entire set of classifications (for all seroconverters) can then be maximised to estimate Ω_T.

Traditionally, studies aimed at estimating the mean duration of recent infection often capture seroconversion panels, obtained from the frequent follow-up and recent infection testing of a relatively small sample of seroconverters. A form of survival analysis, or regression, can then be used to characterise the time taken to exit the 'recent' state, or the evolution of the biomarker over time after infection, respectively, thereby estimating the mean duration of recent infection.[10] For a biomarker that monotonically increases over time after infection, what is often measured is the average time from infection to the biomarker response crossing a selected threshold (defining the transition from the

[10] Methods for estimating the mean duration of recent infection from longitudinal data were systematically explored in later work, and results are presented in Chapter 4. The dangers of neglecting noise in a measured biomarker, by assuming that subjects have single continuous sojourns in the 'recent' state, implicit in all survival analysis approaches and methods that estimate times at which biomarkers 'cross' thresholds, are discussed in Chapter 4. However, such analyses have been used to estimate the MDRI, and are therefore described here for completeness.

'recently' infected to 'non-recently' infected state). Such data and methods may be related to estimation of Ω_T by decomposing the parameter into two parameters, ε and ω:

$$\Omega_T = \varepsilon T + (1-\varepsilon)\omega, \tag{2.37}$$

where ω is the estimated time from seroconversion to threshold-crossing for those individuals who do so within T post infection, and ε is the proportion of seroconverters whose responses are still below the selected threshold at time T post infection.

In all approaches, the recent infection test classifications of subjects at follow-up are used to estimate the mean duration of recent infection. These classifications describe the probabilities of being in the 'recent' state at times post infection *conditional on being alive at follow-up*, which introduces a relative error in the estimation of Ω_T related to the size of γ_3 (see Section 2.2.1).

2.2.4 Simulation and Analysis of Test Datasets

As a further demonstration of the methods presented, 1 000 datasets were simulated and used to infer the characteristics of the test for recent infection and incidence, and results were presented in Section 2.1.2. The simulation of the datasets is outlined below.

The model population and test dynamics were constructed so that, at $t = 0$ and for a time cut-off of $T = 450$ days,

- HIV prevalence was 15%;
- weighted HIV incidence, I_T, was 2%;
- the false-recent rate, β_T, was 2.5%; and
- the mean duration of recent infection, Ω_T, was 160 days.

The probability of being alive and in the state of 'recent' infection at time τ post infection (measured in years) took the form:

$$P_R(\tau) = (1-q)e^{-(\tau/\lambda)^k} + q \text{ for all } \tau \text{ such that } \tau \in \left[0, \frac{600}{365}\right], \tag{2.38}$$

where $q = 0.0167$, $\lambda = 0.4707$ and $k = 3.7183$. This set of parameters ensures that $\Omega_T = 160$ days. It was assumed that there is no death prior to 600 days after infection.

For each of 1 000 simulations, the following datasets were generated:

1. A dataset for estimating β_T, which consisted of binomially generated classifications of 1 000 subjects, representing a random sample of individuals infected for longer than $T = 450$ days, as 'recently' or 'non-recently' infected. In

the estimation, β_T was measured using the maximum likelihood method described by Equation (2.33).

2. A dataset for estimating Ω_T, which consisted of the recent infection test classifications for 2 000 sampled subjects at their first HIV-positive tests, as well as the subjects' inter-test intervals (times between last HIV-negative and first HIV-positive tests). Infection times for subjects were uniformly distributed in their inter-test intervals. Inter-test intervals were generated from a uniform distribution, with a support of 300 to 600 days. The parameter, Ω_T, was estimated by maximising a likelihood function based on Equation (2.36), using $L = 300$ days, $U = 600$ days and the estimated false-recent rate as the input θ. Although θ and the false-recent rate are different in reality, this approximation is likely to be made in practice. Uncertainty in the input θ, which has a true value of 1.67%, was neglected in estimation of Ω_T.
3. A dataset providing the sample counts for inferring incidence, that is, containing the categorisations of 10 000 subjects, captured in a cross-sectional survey at $t = 0$, as uninfected, 'recently' infected or 'non-recently' infected. This dataset was produced by calculating the proportion of the population in each of these three states, based on being consistent with the constructions described above, and generating classifications for the 10 000 subjects from a trinomial distribution. The incidence estimator provided in Equation (2.25) was used to infer incidence.

2.2.5 Prospects for Tests Characterised by More Than Two Parameters

There are a number of ways one may contemplate summarising test dynamics into a greater number of parameters. More complex characterisations of tests for recent infection are briefly considered below, and appear to offer limited prospects.

If the objective of a survey is to provide a single weighted average of recent incidence, then the number of infections, in a specified recent period, is formally a 'sufficient statistic' [150] of this rate – that is, there is no additional benefit in knowing the times of occurrence of these infections. The estimation of the numerator in Equation (2.2) is very nearly the estimation of the number of infections in the last period Ω_T. This suggests that a finer breakdown, beyond counts of 'recent' and 'non-recent' cases, of the survey results

adds no significant accuracy for obtaining a single point estimate of incidence, though it may add significant imprecision.

The characterisation of an ideal test for recent infection (one with a false-recent rate of zero), for the purpose of estimating incidence, requires no parameters other than the mean duration of recent infection. When the test exhibits 'recent' results at large times post infection, an additional parameter, namely a false-recent rate, may be introduced (together with the relevant timescale, T, of 'false-recent' results) to describe this tail. No assumptions about the dynamics of the 'false-recent' results are required. If the 'false-recent' results are distributed over the dynamic range of values for the measured biomarker in a known, non-uniform manner, there may be a benefit to characterising this dynamic with additional parameters. However, it seems unlikely that the characterisation of this dynamic, or its statistically detectable manifestation in a survey, will be feasible, given the need for the false-recent rate to be very low.

One final point that bears mentioning is that data from *a single* survey for a test for recent infection with a long dynamic range could *in principle* be used to yield *multiple* incidence estimates, using different recent/non-recent thresholds, and perhaps different values of T. This would provide, from a single survey, multiple estimates of incidence, each with a subtly different weighting scheme. Given how difficult it is to provide even single incidence estimates, or to detect differences in incidence using data from separate surveys, the effective estimation of an incidence trend from a single survey would require daunting sample sizes and tests for recent infection of currently unrealised performance.

2.3 Online Resource and Analysis Tools for Practical Application

An online resource, maintained by the *South African DST/NRF Centre of Excellence in Epidemiological Modelling and Analysis* (SACEMA), was developed over the course of this project. The website, www.incidence-estimation.com, aims to provide users with both relevant theoretical background information and supporting analysis tools for HIV incidence estimation. Two approaches for estimating incidence are currently supported: (1) from cross-sectional surveys using test for recent infection, which is the focus of this work as further discussed below, and (2) from age-stratified HIV prevalence data

measured at multiple time points, for which methodology and tools were developed by other researchers [60].

The website's landing page is captured in Figure 2.3. Users can then navigate to a theoretical review of the methodology, a set of analysis tools and an archive of related documents. Contact details are also supplied, and investigators from around the world have enquired about the applicability and tailoring of the methodology to their particular studies, and the choice of inputs for analysis tools and interpretation of outputs. A section of the website is dedicated to the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA), facilitating both the consortium's communication with the public, and the internal sharing of documents and data files (access is restricted to CEPHIA members). Updates are also regularly posted, describing latest developments in the field.

The current suite of analysis tools, called *Assay-Based Incidence Estimation* (ABIE) v2.0, was developed to support application of the incidence estimation framework presented in Section 2.1. The suite provides users with various 'calculators', each performing a task-specific statistical analysis. The toolset replaces and offers some advancements on ABIE v1.0, which was both based on the less general incidence inference framework of, and created by, McWalter and Welte [25].

The objective of ABIE v2.0 was increasing accessibility to the work, and therefore the tools were designed to be uncomplicated and user-friendly, allowing a range of potential users (such as test developers, individuals designing studies or HIV programme managers, and data analysts) to more easily move from the theoretical discourse to practical application. The calculators are in the form of Microsoft Excel spreadsheets, providing an interface that is familiar to many. All calculations and statistical tests are performed using simple closed-form approximations (if not exact solutions) and the most straightforward test statistics (differences and ratios of incidences estimates) respectively. While the spreadsheets are protected to prevent any accidental modifications of formulae, users can unlock the spreadsheets to view all underlying calculations or even purposefully tailor the tools to better suit their specific needs. A diverse range of warning and error messages appear when inputs suggest inconsistencies or problematic regimes for application of this surveillance approach. In a number of the calculators, multiple sets of inputs can be provided, enabling the user to consider a range of contexts and analyse the sensitivity of outputs to context.

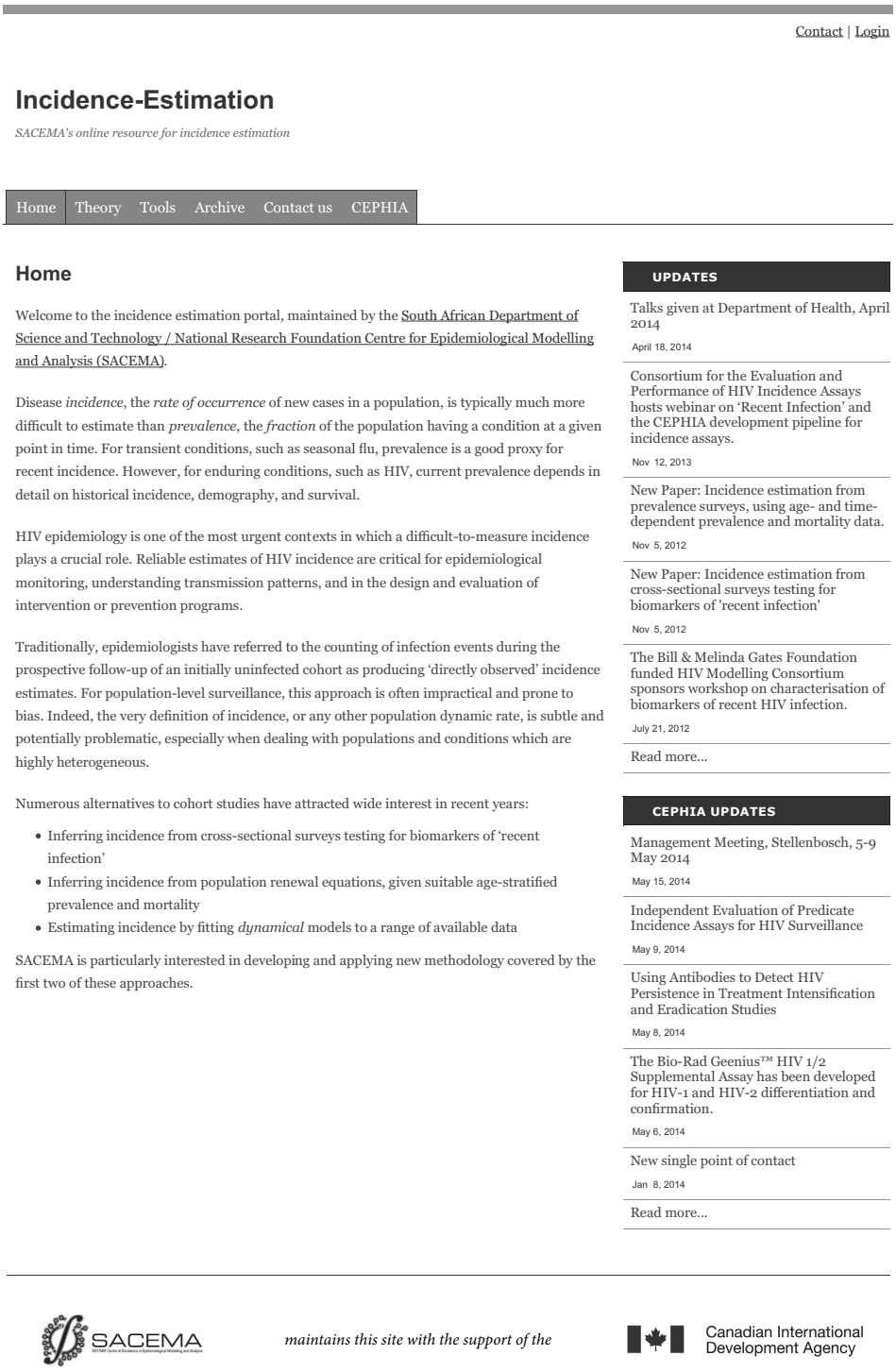


Figure 2.3: Online resource for incidence estimation
The landing page of the incidence estimation website (www.incidence-estimation.com) is shown (as it appeared on 16 July 2014). The SACEMA-maintained website was developed to provide users with methodological background information and analysis tools for HIV incidence estimation.

A detailed description of the inputs, outputs and analyses underlying each of the calculators is provided in Appendix A. In brief, the following calculators are currently available in ABIE v2.0:

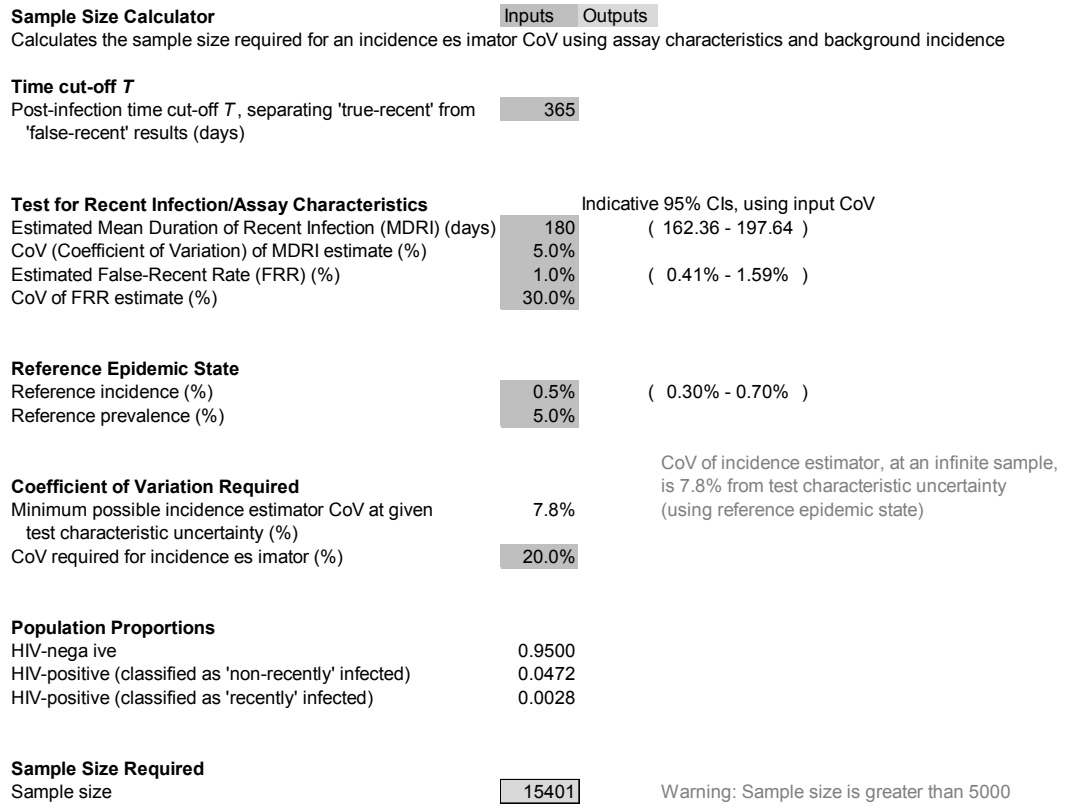
- *Incidence and Prevalence Calculator* – estimates HIV incidence and prevalence, from (i) counts of uninfected, 'recently' infected and 'non-recently' infected subjects observed in a cross-sectional survey, and (ii) properties of the test for recent infection (the test's MDRI and FRR, and the uncertainties with which they are measured).
- *Sample Size Calculator* – calculates the number of subjects required in a cross-sectional incidence study to obtain a specified precision of the incidence estimator, given the (i) desired precision, (ii) assumed epidemiological context (HIV incidence and prevalence), and (iii) test properties.
- *Incidence Ratio Calculator* – estimates the ratio of two incidence values, from (i) counts of uninfected, 'recently' infected and 'non-recently' infected subjects observed in each of two cross-sectional surveys, and (ii) test properties (presumed to be the same in both populations).
- *P-value for Difference Calculator* – calculates a conventional p-value to summarise the discrepancy between two incidence estimates (more specifically, the probability of obtaining as large a difference between two incidence estimates as that actually observed), under the null hypothesis of equal incidence in the two study populations, assuming equal HIV prevalence. Inputs are the (i) counts of uninfected, 'recently' infected and 'non-recently' infected subjects observed in each of the two cross-sectional surveys, and (ii) test properties (presumed to be the same in both populations).
- *Power to Detect Difference Calculator* – calculates the probability of inferring a difference between two incidence values, in the correct direction, when considering the difference between two incidence estimates, assuming equal HIV prevalence. Inputs are the (i) assumed epidemiological context, for each study population, (ii) incidence survey size, for each population, (iii) test properties (equal in both populations), and (iv) statistical significance level for the test for difference in incidence.

- *Test Performance Calculator*[11] – calculates the precision of the incidence estimator, given the (i) assumed epidemiological context, (ii) incidence survey size, and (iii) test properties.

A weakness of cross-sectional incidence surveillance, drawing some criticism, is the potentially large samples required to obtain suitably precise incidence estimates in low HIV incidence settings, given the current non-ideal performance of tests for recent infection. This important topic of sample size considerations is used to illustrate an application of the calculators. In Figures 2.4 and 2.5, the *Sample Size Calculator* is used to explore the relationship between sample size and incidence estimation precision, for a test for recent infection that has properties closely meeting the 'Target Product Profile' [14]: an MDRI of 180 days and an FRR of 1%, for $T = 1$ year. Based on experiences from analysing data produced by CEPHIA, the coefficient of variation (CoV) for MDRI estimation was taken to be 5%; and the CoV for FRR estimation was set to 30%, which corresponds to estimation of the FRR in a sample of 1 000 long-infected subjects.

In the screenshot of the *Sample Size Calculator* provided in Figure 2.4, the tool was used to explore an epidemiological setting where the HIV incidence rate is low at 0.5% per person year and HIV prevalence is 5%. The incidence value represents the weighted incidence that would be measured in the cross-sectional survey (that is, there are no assumptions of constant historical incidence). The input specifies that the required CoV of the incidence estimator is 20% – roughly equivalent to requiring a 95% probability of obtaining an HIV incidence point estimate between 0.3% and 0.7% per person year. The calculator output indicates that a sample of 15 400 subjects would be required in the incidence study to achieve this level of precision. This can be compared to the approximately 11 000 HIV diagnostic tests required to achieve the same precision of incidence estimation using a longitudinal study, which identifies HIV-negative subjects in a cross-sectional survey of the population and retests these subjects one year later, in the described context (5 500 subjects would be tested for HIV at the start of the study, and the enrolled HIV-negative subjects retested at the end of the study).

[11] The *Test Performance Calculator* was developed to accompany the article presented in Chapter 6, in which the precision of the incidence estimator is proposed as a standard metric for assessing test performance.



Sample Size Calculator	Inputs	Outputs	
Calculates the sample size required for an incidence es imator CoV using assay characteristics and background incidence			
Time cut-off *T*			
Post-infection time cut-off *T*, separating 'true-recent' from 'false-recent' results (days)	365		
Test for Recent Infection/Assay Characteristics		Indicative 95% CIs, using input CoV	
Estimated Mean Duration of Recent Infection (MDRI) (days)	180	(162.36 - 197.64)	
CoV (Coefficient of Variation) of MDRI estimate (%)	5.0%		
Estimated False-Recent Rate (FRR) (%)	1.0%	(0.41% - 1.59%)	
CoV of FRR estimate (%)	30.0%		
Reference Epidemic State			
Reference incidence (%)	0.5%	(0.30% - 0.70%)	
Reference prevalence (%)	5.0%		
Coefficient of Variation Required		CoV of incidence estimator, at an infinite sample, is 7.8% from test characteristic uncertainty (using reference epidemic state)	
Minimum possible incidence estimator CoV at given test characteristic uncertainty (%)	7.8%		
CoV required for incidence es imator (%)	20.0%		
Population Proportions			
HIV-nega ive	0.9500		
HIV-positive (classified as 'non-recently' infected)	0.0472		
HIV-positive (classified as 'recently' infected)	0.0028		
Sample Size Required			
Sample size	15401	Warning: Sample size is greater than 5000	

Figure 2.4: Example analysis using the *Sample Size Calculator* contained in the ABIE v2.0 tool suite

In the screenshot of the *Sample Size Calculator*, the analysis tool output indicates that 15 401 subjects are needed in a cross-sectional incidence study to obtain the desired precision of the incidence estimator, in the specified scenario. Inputs indicate that an incidence estimator with a 20% CoV is required, in a population where HIV incidence and prevalence are 0.5% per annum and 5% respectively, using a test for recent infection that has an MDRI of 180 days, estimated with a 5% CoV, and an FRR of 1%, estimated with a 30% CoV, for $T = 1$ year.

The required sample size is expected to vary greatly by epidemiological context. The calculator, which accommodates multiple sets of inputs, was therefore used to calculate the sample sizes required in populations with various HIV incidence and prevalence values. The tool output, provided in Figure 2.5, shows that the sample size required to obtain the same CoV of the incidence estimator (20%) decreases as HIV incidence increases or prevalence decreases. A much smaller survey, of 5 900 subjects, is required when incidence is higher, at 1% per annum, and prevalence is lower, at 2.5% (this context may capture, for example, an emerging epidemic among teenagers).

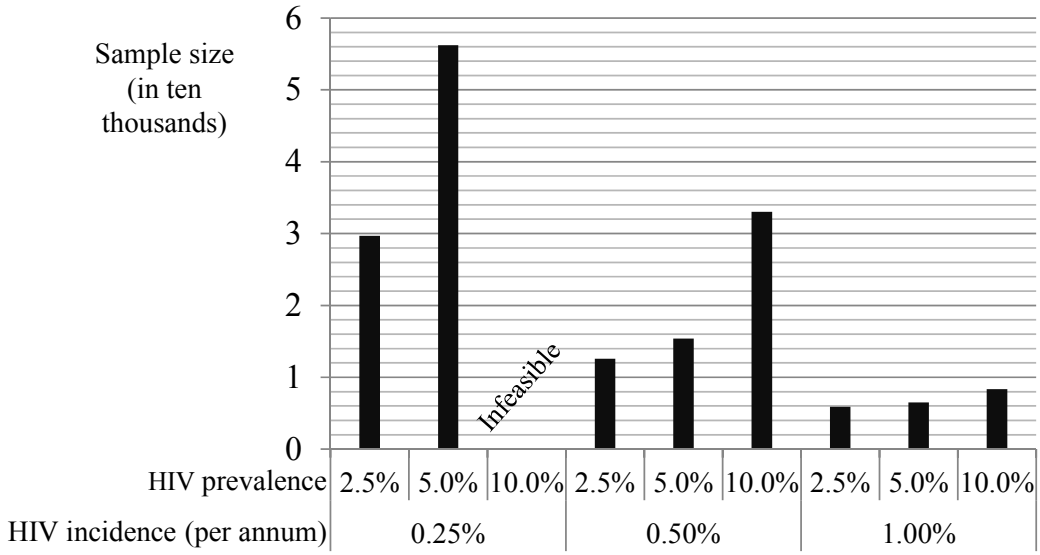


Figure 2.5: Sample size required to obtain a specified CoV for the incidence estimator by HIV incidence and prevalence in the study population, using the *Sample Size Calculator*

The chart, produced by the Sample Size Calculator contained in the toolset ABIE v2.0, shows the sample size (in tens of thousands of subjects) required in a cross-sectional surveillance study to obtain an incidence estimator with a 20% CoV in each epidemiological context considered. The epidemiological contexts capture different combinations of HIV incidence (0.25%, 0.5% and 1% per annum) and prevalence (2.5%, 5% and 10%). The test for recent infection has an MDRI of 180 days, estimated with a 5% CoV, and an FRR of 1%, estimated with a 30% CoV, for $T = 1$ year. When HIV incidence is 0.25% per annum, and HIV prevalence is at least 7.5%, the uncertainty from the estimated test properties already implies a CoV of the incidence estimator greater than 20%.

The ABIE analysis tools have been endorsed by the *World Health Organisation* (WHO) *HIV Incidence Assays Working Group* [13, 130], and are used as part of training delivered by the *Centers for Disease Control and Prevention* (CDC). Locally, the ABIE tools are used to analyse data from the South African national household surveys [61]. We, at SACEMA, also directly hosted a training workshop, at the request of the WHO, in South Africa in 2012 [49, 50]. Participants were primarily epidemiologists and programme officers from national health departments and CDC divisions in countries around Africa. The theoretical foundations were presented, and the use of each analysis tool demonstrated. The WHO has requested that the training is continued, and a second workshop was conducted in September 2014.

In addition to the standard analysis tools provided online, numerous ad-hoc calculators have been created in response to specific requests. For example, the *Incidence and Prevalence Calculator* was tailored to consistently account for the scenario in which not all of the HIV-positive subjects identified in the cross-sectional survey were tested for 'recent' infection.[12] As another example, a tool was developed to calculate the sample size required to achieve a desired statistical power when testing for a difference in incidence in two populations, where incidence is estimated from a cross-sectional study using a test for recent infection in one population, and by a conventional longitudinal study in the other.[13]

The next version of the tool suite, ABIE v3.0, is envisioned as providing much greater flexibility and more rigorous statistical calculations. The intention is to develop the tools in R, which is open-source software that is suited to statistical programming [152]. More accurate calculations, requiring more computationally-expensive numerical methods, could then be efficiently implemented. For example, in power calculations, the full distribution of possible survey counts could be considered (rather than utilising closed-form approximations that are based on expected counts), and bootstrap resampling could be used to estimate uncertainties. A broader range of scenarios, better capturing those that may be encountered in reality, could be accommodated, for example by relaxing assumptions about equal HIV prevalence and equal test properties in the populations being compared; and optimally-powered statistical tests should be developed for hypothesis testing. A broader range of input types could also be accommodated (such as plausible minimum and maximum values of parameters rather than specifications of single values) and more diverse outputs produced (such as files summarising outputs in tables or visually through various plots). While such a tool suite would not present the simple interface and calculations contained in ABIE v2.0, progressively advancing the analytical tools would support the expanding application of this surveillance approach.

---

[12]The analysis was requested by *Medicins Sans Frontiers*, for data from an incidence study in Kenya, intended to provide a baseline incidence measure for monitoring transmission over time [151].

[13]The tool was used by the South African *Perinatal HIV Research Unit* to design a study intended to evaluate the impact of medical male circumcision in Soweto, South Africa.

Chapter 3

Estimating the Mean Duration of Recent Infection I: Observing Subjects Once after Infection

This chapter formalises and demonstrates the utility of an approach for obtaining preliminary estimates of a test's mean duration of recent infection (MDRI) [143] using previously overlooked data. The procedure requires only a single application of the test for recent infection per subject, on specimen drawn at the time of the first HIV-positive test. No subsequent follow-up of subjects is necessary, and the methodology is suited to contexts where times between HIV diagnostic tests may be large (such as a year or two).

The methodology is explored in Section 3.1, and used to perform initial characterisations of two detuned assays using the previously untapped source of specimens provided by blood donors. This analysis is a reproduction of a published journal article [31].[14] Ancillary details of the analysis, which were published in an appendix to the article, are provided in Section 3.2.

[14]The contents of Section 3.1 have been published as: 'Kassanjee R, Welte A, McWalter TA, Keating SM, Vermeulen M, Stramer SL, Busch MP. Seroconverting blood donors as a resource for characterising and optimising recent infection testing algorithms for incidence estimation. *PLOS ONE*. 2011; 6(6):e20027'. PLOS applies the Creative Commons Attribution (CC BY) license to all works published, and therefore no permission was required to reproduce the work. The manuscript was primarily written by RK, who performed all analyses. AW and TAM critically reviewed results and assisted in writing the manuscript. Specimen collection and laboratory work were performed by MPB, SMK, MV and SLS, and these were led by MPB who conceived the design.

Two further applications of this approach are presented in Section 3.3, namely the characterisation of a newly proposed test for recent infection, based on SMARTube™ technology, and the characterisation of the widely used BED assay for the South African context. These applications were presented at HIV conferences [33, 35].[15]

At the time of the work described above, the incidence inference framework of McWalter and Welte [25] was in use. The more general inference scheme, derived in Chapter 2, had not yet been adopted. However, the ideas presented in this chapter are as applicable and valuable under the more general framework of Chapter 2, and in fact benefit from it, as shown in Section 3.4.

[15]The analysis described in Section 3.3.1 has been previously presented as 'Kassanjee R, Welte A, Jehuda-Cohen T. SMARTube as a test for recent infection. Poster 41 and presentation at the *2010 HIV Diagnostics Conference*. 24-26 March 2010, Florida, USA'. The analysis presented in Section 3.3.2 has appeared as 'Kassanjee R, Welte A, McWalter TA, Viljoen J, Bärnighausen T, Newell ML, Fatti, LP. Calibration of BED assay for use in incidence estimation. E-poster CDB018 at the *5th IAS Conference on HIV Pathogenesis, Treatment and Prevention*. 19-22 July 2009, South Africa'.

3.1 Seroconverting Blood Donors as a Resource for Characterising Tests for Recent Infection

3.1.1 Introduction

Incidence, the rate of new infections, provides a more direct and current indication of the spread of HIV than prevalence, the fraction of the population in an infected state. Incidence estimates are key to monitoring epidemics, assessing outbreaks, and targeting and evaluating interventions. Prospective longitudinal studies, which allow for the direct counting of new infections in cohorts of individuals, are costly, logistically difficult to set up and maintain, and prone to capturing unrepresentative behaviours. Consequently, estimation of incidence using cross-sectional surveys [7, 15, 91] has attracted much interest over recent years.

Tests for recent infection, also termed Recent Infection Testing Algorithms (RITAs) [10] or Serologic Testing Algorithms for Recent HIV Seroconversion (STARHS) [91], classify infections as 'recently' or 'non-recently' acquired. Incidence is then related to the prevalence of test-defined 'recent' infection [7, 15-17, 24-26, 91, 92, 143, 145] as estimated in a cross-sectional survey.

Tests for recent infection traditionally employ the laboratory measurement of HIV viral or host biomarkers which evolve over time after infection. Antibody avidity, titre, or HIV-specific proportion is typically considered, with a measurement below a chosen threshold indicative of 'recent' infection [8, 9, 107].

Immune responses vary for individuals, and so each individual experiences a unique evolution of the biomarker. There are two test characteristics of relevance for population-level incidence estimation.

1. The test-defined state of 'recent' infection should not be too transient. This ensures that the proportion of the population in this state may be estimated with good statistical power in surveys with feasible sample sizes. Therefore, the average time spent in the state of 'recent' infection, termed the **mean duration of recent infection (MDRI)**, ω, should be large (typically, at least six months [11]).

2. For many tests, there is evidence that some long-infected individuals are classified as 'recently' infected [8, 9]. Although the phenomenon of 'false-recent' results may, in principle, be accounted for without introducing bias, adjustments result in considerable loss of statistical precision of incidence estimates [11]. The proportion of long-standing infections classified by the test as 'recent', termed the **false-recent rate (FRR)**, ε, should therefore be as low as feasible.

Increasing the threshold (the biomarker cut-off used to discriminate 'recent' from 'non-recent' infection) increases the MDRI, but typically also results in a higher FRR. Therefore, as the threshold varies, there is a trade-off between the two test characteristics. Since population-level surveillance is of interest, rather than each individual's 'recent' or 'non-recent' infection diagnosis, a sensitivity-specificity trade-off (with truly recent infection defined by a fixed duration after infection) is not an appropriate threshold optimisation criterion.[16]

Both calibration data and cross-sectional survey data are required to estimate incidence. Calibration data is used to estimate the test characteristics, namely the MDRI, ω, and FRR, ε. Cross-sectional data is used to estimate the proportions of susceptible or uninfected, 'recently' infected and 'non-recently' infected individuals in the population, denoted by P_S, P_R and P_{NR}, respectively.

Population proportions and test characteristics are related to incidence, I, by the following expression:[17]

$$I = \frac{P_R - \frac{\varepsilon}{1-\varepsilon} P_{NR}}{\omega P_S}. \tag{3.1}$$

This has been derived in an analysis by McWalter and Welte [25], shown to be the maximum likelihood estimator by Wang and Lagakos [24], and informally generalised by Welte et al [11]. McWalter and Welte [23] compare this estimator to the previously proposed estimators of McDougal et al [16] and Hargrove et al [17].

---

[16]The optimisation of tests for recent infection, using a performance metric that is appropriate to their application in HIV incidence surveillance, is discussed in Chapter 6.

[17]The general incidence inference framework, which also provides definitions of the test properties that are free of assumptions about the test dynamics, as presented in Chapter 2, was still under development at the time of this analysis.

Ideally, the test should perform similarly in a number of populations, allowing for the reuse of test characteristic estimates. However, differences in the stage of the epidemic, or viral subtype or clade, may necessitate the estimation of these critical parameters in relevant populations for each study. For example, the proportions of individuals who are elite controllers (whose immune systems successfully suppress viraemia in the absence of treatment), have advanced immunodeficiency or are receiving antiretroviral therapy, may vary, and these individuals have a propensity to produce 'false-recent' classifications [8, 9].

Traditionally, methods of estimating the MDRI have relied on the testing of serial samples from acutely infected subjects [7, 9, 15-17, 24, 91-93]. This typically requires at least one pre-seroconversion and multiple post-seroconversion samples, with short intervals between follow-up so that the seroconversion and threshold-crossing times may be estimated with minimal uncertainty.[18] Such panels of data are costly and difficult to capture, requiring precisely the demanding longitudinal studies that cross-sectional incidence estimation seeks to circumvent.

Despite being more easily obtained, specimens from seroconverting subjects with relatively long intervals between follow-up have been largely overlooked. Obtaining such specimens from repeat blood donors provides unique efficiencies as the collection of blood for transfusions is ongoing in most countries, and therefore procuring specimens does not require the establishment of new surveillance. Although the prevalence and incidence of HIV are generally lower in blood donors than the general population, the large-scale collection of blood and routine testing of serial donations for HIV provide a relatively large sample of seroconverting donors. Furthermore, large volumes of plasma, derived from routinely prepared frozen plasma components, are obtained.

In this investigation, data captured on seroconverting blood donors in South Africa and the USA are used to demonstrate the characterisation and optimisation of tests for recent infection.

---

[18]In this chapter, the terms 'seroconversion' and 'infection' are used interchangeably, and both refer to HIV infection being detectable by the HIV diagnostic test (assumed to antibody-based here, although the work presented is completely general and any HIV diagnostic could be considered). Seroconversion describes the development of antibodies in the subject, in response to the virus, specifically to levels that are measurable by a chosen antibody-based HIV test.

3.1.2 Methods

Ethics statement

The research and the incidence testing were approved by the *University of California, San Francisco* (UCSF); *American Red Cross* (ARC) and *South African National Blood Service* (SANBS) institutional review boards or ethics committees.

Specimen collection and testing for 'recent' infection

Specimens were collected by the *South African National Blood Service* (SANBS) of South Africa and the *American Red Cross* (ARC) of the USA, and tested by the *Blood Systems Research Institute* (BSRI) of the USA. Repeat donors who were observed to seroconvert were tested for 'recent' infection using the specimens collected at the times of the first seropositive donations.

The investigation was performed for the less-sensitive Vironostika assay (LS-Vironostika) [104], the test for which more data are available, and thereafter, the currently-used less-sensitive Vitros assay (LS-Vitros) [153] was characterised. These tests are both based on 'less-sensitive' versions of diagnostic tests that measure antibody titre, a concept introduced by Janssen et al [91]. For each test, 'recent' infection is indicated by a measured biomarker, namely a standardised optical density (SOD), below a chosen threshold.

LS-Vironostika is a modification of the Vironostika HIV-1 microELISA diagnostic test (bioMérieux, Marcy l'Étoile, France) [104]. The laboratory procedures and threshold of 1 specified by Rawal et al [104] were used. Seroconverting blood donors were tested using LS-Vironostika until 2007, as production of the Vironostika assay ceased in the year thereafter [9]. Manufacturing of the assay has since been resumed by Avioq Inc (Rockville, MD) [154].

LS-Vitros is based on the Ortho Vitros ECi anti-HIV 1+2 instrument (Ortho-Clinical Diagnostics, Raritan, NJ) [153]. The laboratory conditions were established by BSRI in earlier work that sought the closest agreement of classifications by LS-Vitros, using a threshold of 20, to those provided by LS-Vironostika [153].

Data consist of the biomarker reading (SOD) at the time of the first seropositive donation, and the interval between the last seronegative and first seropositive donation, termed the inter-donation (ID) interval, for each seroconverting blood donor. Three datasets (plotted in Section 3.2.2) were used for the analysis: LS-Vironostika was applied to samples of South African donors (October 2005 to September 2007, sample size of $n = 485$) and North American donors (November 2001 to December 2005, $n = 176$), and LS-Vitros was applied to a sample of South African donors (October 2007 to December 2009, $n = 199$).

Data analysis

Test characteristics were estimated using a maximum likelihood method. Rather than fitting a curve describing the evolution of the biomarker over time after seroconversion, the overall probability of the test classifications at the first seropositive donations in the sample was maximised [33]. The likelihood function is derived below (and more detail is provided in Section 3.2.1).

Assuming that the time of seroconversion is uniformly distributed in the ID interval, the probability that a seroconverter with ID interval Δ is classified as 'recently' infected at the time of the first seropositive donation is

$$p(\Delta) = \frac{1}{\Delta}\int_0^{\Delta} S_R(t)\,\mathrm{d}t, \tag{3.2}$$

where $S_R(t)$ is the probability of being in the test-defined state of 'recent' infection when tested at time t after seroconversion. For the i^{th} seroconverter in the sample, with ID interval Δ_i, the probability of testing 'recent' is $p_i = p(\Delta_i)$.

The likelihood, L, to be maximised, of all test classifications in a sample of n seroconverters is

$$L = \prod_{i=1}^{n} (p_i)^{x_i}\,(1 - p_i)^{1-x_i}, \tag{3.3}$$

where x_i is the observed result for the i^{th} seroconverter, and equals 1 if the subject is 'recently' infected and 0 if the subject is 'non-recently' infected.

In the analyses of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24], individual biomarker curves either cross the threshold (distinguishing 'recent' from 'non-recent' infection) and readings remain above it thereafter, or else fail to reach the

threshold, and therefore $S_R(t)$ approaches some constant value, α, which is the proportion of biomarker curves that fail to reach the threshold for large t. $S_R(t)$ may then be expressed as

$$S_R(t) = \alpha + (1-\alpha)S_{R'}(t). \tag{3.4}$$

In the above-mentioned analyses, the MDRI, ω, is the mean time under the threshold for only those biomarkers curves that do cross the threshold, as described by $S_{R'}(t)$.

Substituting from Equation (3.4) into Equation (3.2), the probability that the i^{th} seroconverter is 'recently' infected becomes

$$p(\Delta_i) = p_i = \alpha + (1-\alpha)\frac{1}{\Delta_i}\int_0^{\Delta_i} S_{R'}(t)\,\mathrm{d}t. \tag{3.5}$$

This approach also facilitates non-parametric inference, by considering only subjects with large Δ_i. For a time cut-off, W, such that

$$S_{R'}(t) = 0 \ \ (\text{or } S_R(t) = \alpha) \ \ \forall\, t > W, \tag{3.6}$$

if $\Delta_i > W$, then

$$\int_0^{\Delta_i} S_{R'}(t)\,\mathrm{d}t \ = \int_0^{\infty} S_{R'}(t)\,\mathrm{d}t = \omega \tag{3.7}$$

is the MDRI. Substituting from Equation (3.7) into Equation (3.5), p_i becomes a function of the two test characteristics,

$$p_i = \alpha + (1-\alpha)\frac{\omega}{\Delta_i}, \tag{3.8}$$

and no assumptions are required about the shape with which the biomarker grows after seroconversion (that is, no full specification of $S_{R'}(t)$ is required). The estimated test characteristics maximise the likelihood, L, which is now a function of ω, and also of α if there is no input estimate of α.

McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24] additionally assumed that post-seroconversion survival is independent of the shape of the biomarker curves. When the above-mentioned assumptions are obeyed, $\alpha = \varepsilon$ in the incidence expression in Equation (3.1). More generally, $S_R(t)$ may not remain constant for $t > W$ for a value of W used in the analysis. An FRR may then be defined as the proportion of individuals, seropositive for longer than W, that are classified as 'recently' infected [11]. In this case, the above procedure that produces an estimate of α, then considered only a proxy 'FRR', probably overestimates the FRR if biomarker curves cross the threshold

after W or underestimates it if biomarker curves move back below the threshold at times since seroconversion greater than that captured in the dataset. The estimated test characteristics, α and ω, therefore provide unrefined estimates for the FRR and MDRI.

Uniformly distributed seroconversion times are reasonable when the timing of donations and exposures to HIV are independent. Test-seeking behaviour (the donation of blood soon after exposure specifically to receive HIV testing) or deferral of donations (the delay of donations soon after exposure) could therefore bias estimates. In the USA, an investigation, which highlighted test-seeking behaviour among homosexual men, noted little indication of test-seeking behaviour among blood donors [155], while evidence of deferred donations has been observed [156]. Behaviour in the South African donor population may vary due to the large scale of the epidemic and stigma associated with HIV.

In this work, various analyses involving the 'parametric' and 'non-parametric' inference of test characteristics by maximising the likelihood function in Equation (3.3) were performed for LS-Vironostika and LS-Vitros. In parametric inference, the probability of being 'recently' infected as expressed in Equation (3.5) was used, assuming forms for $S_{R'}(t)$ and including all data in the analysis; and in non-parametric inference, the probability in Equation (3.8) was used, including only data satisfying $\Delta_i > W$. Using simulated data, estimates obtained from the parametric and non-parametric approaches were compared. Differences in test characteristics for specific subpopulations were explored. The utility of the test for obtaining precise incidence estimates was also investigated.

Asymptotic maximum likelihood theory was used to estimate confidence intervals (CIs) and confidence regions (CRs) and to test the significance of parameters (based on the distribution of the deviance statistic and using the loglikelihood ratio test) [157]. Chi-squared goodness of fit tests were used to assess agreement between data and assumptions [158]. All tests used a significance level of 5%.

3.1.3 Results

Characterisation of LS-Vironostika

The estimated test characteristics for LS-Vironostika (using a threshold of 1) are shown in Figure 3.1, for both estimation of ω assuming a known α and simultaneous estimation of ω and α, and analysing South African and American blood donors separately. Non-parametric estimation was performed, using only observations with $\Delta_i > W = 1$ year (the maximum duration in the state of 'recent' infection has been estimated to be 200 days [159] and 1 year [93]). The sample sizes for the analyses of South African and American donors were $n = 282$ and $n = 106$ respectively. A comparison of the observed proportions of seroconverters who were 'recently' infected to the expected proportions (based on estimated test characteristics), as a function of ID interval, suggests good agreement under simultaneous estimation of the parameters (see Section 3.2.2). When exploring the sensitivity of results to W, by increasing W to values up to 2.5 years, estimates for South African donors varied by at most 10% ($n = 189$ when $W = 2.5$ years), while the large uncertainty in estimates based on the relatively small dataset for American donors ($n = 53$ when $W = 2.5$ years) did not support meaningful inference.

The estimated α is large, consistent with results from the application of this method to assess the BED assay [33 – results not shown]. Estimation of ω using an input estimate of α is preferable. In the extreme case of all ID intervals being equal, ω and α cannot be simultaneously estimated as the likelihood function may be kept at its maximum while arbitrarily increasing the estimate of α by appropriately decreasing the estimate of ω. Furthermore, using a value of W that is too low (biomarker curves cross the threshold after W) would bias estimates of α upwards and ω downwards, based on the framework of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24]. Note that larger values of W would be required at higher thresholds for discriminating between 'recent' and 'non-recent' results.

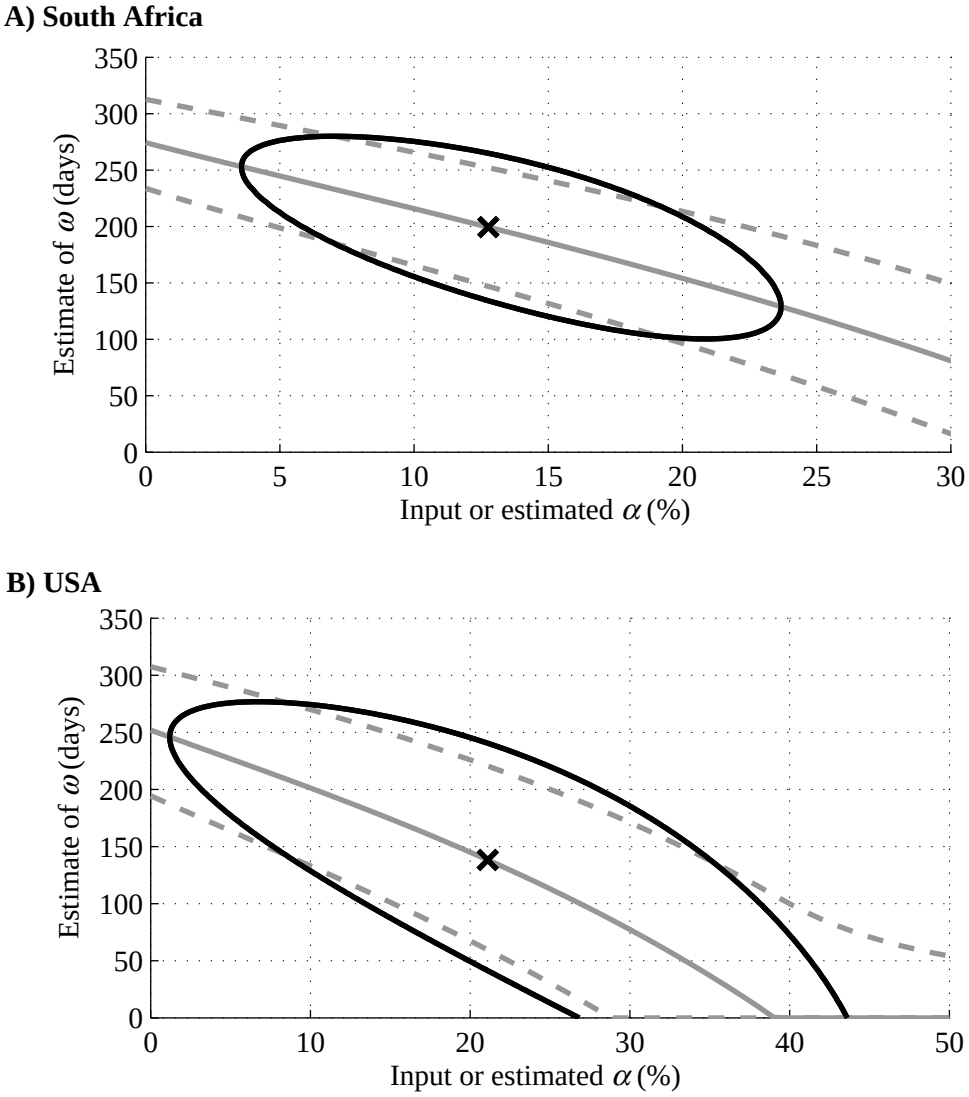


Figure 3.1: Estimated test characteristics for LS-Vironostika in the repeat donor population

Estimates of the MDRI, ω (days), for LS-Vironostika are shown, under both the simultaneous estimation of ω and α, and when using an input α, for $W = 1$ year. For the latter estimation, the estimated ω is plotted as a function of the assumed α. The 95% confidence regions (CRs) for ω and α (simultaneous estimation) and confidence intervals (CIs) for ω (assuming α, and not accounting for uncertainty in α) are displayed. Results are shown for repeat blood donors from A) South Africa and B) the USA.

Legend:
- ✖ Estimates of ω and α from simultaneous estimation of characteristics ($\hat{\alpha} = 13\%$, $\hat{\omega} = 199$ days for South Africa, $\hat{\alpha} = 21\%$, $\hat{\omega} = 138$ days for USA)
- ━ 95% CR boundary for characteristics from simultaneous estimation
- ─ Estimate of ω using an input α
- ┄ 95% CI limits for ω using an input α

The estimated MDRIs, for a number of thresholds (holding W at 1 year), are compared to published estimates in Figure 3.2:

1. Busch et al [160] utilised the directly measured incidence in the repeat donor population to estimate ω. Using 'known' incidence and proportions P_R, P_{NR} and P_S (measured in the repeat donor population), and assuming a zero FRR, a 'back-calculation' for ω using the expression for incidence in Equation (3.1) was performed. Since the possibility of 'false-recent' results was neglected, overestimation of ω is expected, with greater bias at higher thresholds. Methodologically, estimation of ω by back-calculation requires an existing estimate of the FRR, ε, for the same threshold, with such data currently unavailable. Furthermore, uncertainty in the estimate of ω arises from uncertainty in the estimated incidence; proportions and FRR.
2. The *Centers for Disease Control and Prevention* (CDC) utilised seroconversion panels to estimate ω in an American population [104, 160].

Parametric versus non-parametric approach

The need for parametric assumptions about the shape of the antibody titre response curve, summarised into parametric assumptions about the probability of testing 'recently' infected as a function of time since seroconversion, is circumvented by using only data with large ID intervals. Consequently, estimation of ω is no longer prone to bias arising from poor parametric assumptions, but the dataset used for the estimation is reduced in size, decreasing the precision (increasing the variability) of estimates of ω. The characterisation of LS-Vironostika in the South African repeat donor population was revisited, this time using all data and parametric assumptions.

The probability that a seroconverter is 'recently' infected at the first seropositive donation is given by Equation (3.5), which can be assessed once a form for $S_{R'}(t)$ is fully specified. For $S_{R'}(t) = S_{R'}\left(t|\underline{\phi}\right)$, where $\underline{\phi}$ is a vector of parameters, the likelihood of the data, given by Equation (3.3), becomes a function of $\underline{\phi}$. The MDRI estimator is then

$$\widehat{\omega} = \int_0^\infty S_{R'}\left(t|\underline{\hat{\phi}}\right)\,\mathrm{d}t. \tag{3.9}$$

where $\underline{\hat{\phi}}$ maximises the likelihood function (in Equation (3.3)).

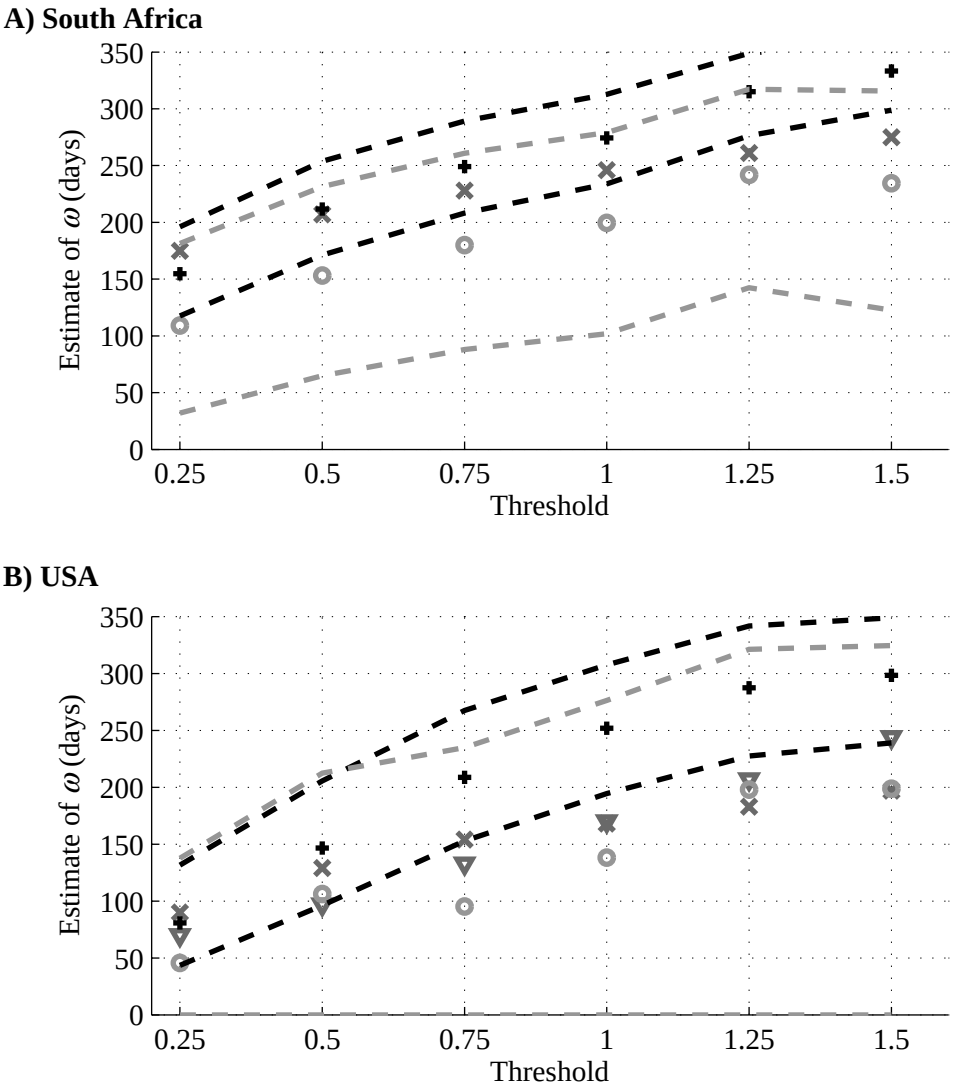


Figure 3.2: Comparison of mean duration of recent infection estimates for LS-Vironostika to previously published estimates

Estimates of the MDRI, ω (days), for LS-Vironostika, under both the simultaneous estimation of ω and α, and when assuming $\alpha = 0\%$, for $W = 1$ year, are compared to published estimates, as a function of test threshold. Published estimates were obtained by 'back-calculation' in the repeat donor population (Busch et al) [160] or using seroconversion panels (CDC) [104, 160]. The minimum and maximum ω occurring in the 95% confidence regions (CRs) for ω and α (simultaneous estimation) and 95% confidence interval (CI) limits for ω (assuming $\alpha = 0\%$, with no uncertainty) are also displayed. Estimates pertain to A) South African and B) USA populations.

Legend:
- Estimate of ω from simultaneous estimation of characteristics
- Minimum and maximum ω in 95% CR for characteristics from simultaneous estimation
- Estimate of ω using an input $\alpha = 0\%$
- 95% CI limits for ω using an input $\alpha = 0\%$
- Published back-calculation estimate of ω by Busch et al
- Published seroconversion panel estimate of ω by CDC

A number of parametric forms for $S_{R'}(t)$ were used in the estimation of the MDRI for LS-Vironostika, based on assumptions ranging from a fixed duration of 'recent' infection for all individuals to a fat-tailed Pareto distribution for the time spent in the state of 'recent' infection. Widely varying estimates of ω were obtained, even after excluding estimates for which assumptions and data did not agree (and all results are shown in Section 3.2.3). Since the true underlying dynamics of the data are unknown, the extent of bias from any incorrect parametric assumptions is unclear.

Simulated data was therefore used to explore the trade-off between precision and bias when moving between non-parametric and parametric approaches. Based on each of a number of forms for $S_{R'}(t)$, multiple datasets were generated. For each dataset, the MDRI was estimated parametrically using each of the forms for $S_{R'}(t)$ in turn (including all data) as well as non-parametrically (including only data with large ID intervals). Agreement between data and parametric assumptions was also assessed. The results of the investigation (shown in Section 3.2.3) suggest that power to reject 'incorrect' parametric assumptions is at times poor and that large bias in estimates may occur. When the assumptions leading to Equation (3.8) hold, estimates using the non-parametric approach are unbiased, although less precise.

In the estimation of ω for LS-Vironostika using the South African donor sample, by using a non-parametric approach, there is a 40% reduction in the sample size from excluding ID intervals smaller than $W = 1$ year. However, potential bias arising from poor parametric assumptions is then eliminated, noting that one cannot easily distinguish between appropriate and poor parametric assumptions using only the data at hand.

Population-specific test characteristics

Significant systematic bias could be introduced to incidence estimates if the test characteristics are not evaluated in a population representative of that in which incidence estimation is to occur [11, 26]. Since most HIV antibody assays are based primarily on clade B antigens, antibody-antigen reactivity may vary when applying assays in populations in which other clades occur [161], with differences in the characteristics of LS-Vironostika already observed [104, 105, 159, 161, 162]. Other factors, such as the association between viral RNA levels and clade, and seroconverters' genetic backgrounds, may also affect results [8, 9, 92, 161].

The significance of gender (male and female) and country (South Africa and USA) on the characteristics of LS-Vironostika was assessed. Country differences are likely to be largely representative of clade differences, as clade C infections are predominant in South Africa, and clade B in the USA [163]. Investigations by SANBS on a sample of donors (data made available to authors) and studies of North American donors [164, 165] indicate that a very small percentage (<5%) of infections are not of the predominant clade.

The null hypothesis, that the characteristics of LS-Vironostika are the same in all four groups (each pairing of gender and country), is not rejected with a p-value of 10.48%. Estimated test characteristics for the groups are shown in Figure 3.3 (using non-parametric estimation and $W = 1$ year). However, in this investigation, large uncertainty in estimates, arising from small samples of seroconverters, would result in little power to identify significant factors.

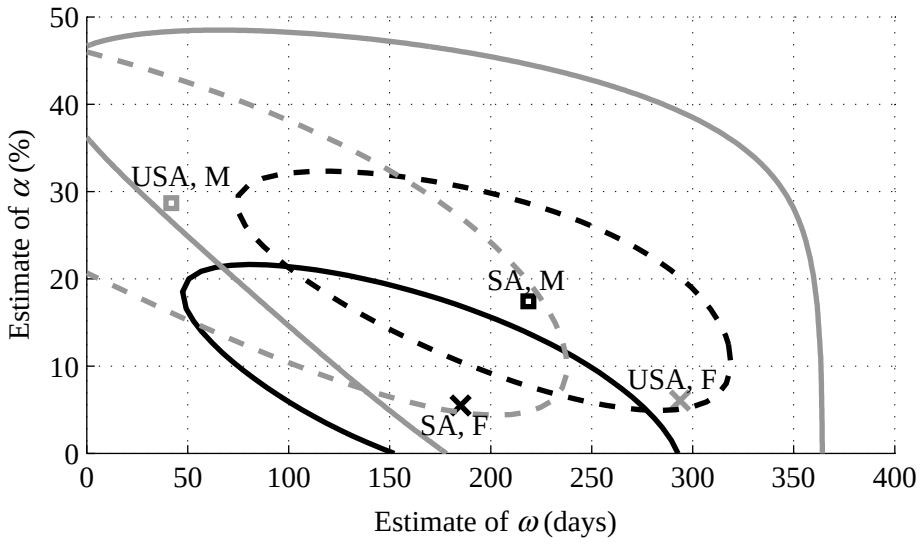


Figure 3.3: Estimated test characteristics for LS-Vironostika in the repeat donor population by gender and country

Estimates of ω (days) and α (%) for LS-Vironostika are shown for South African male donors, South African female donors, USA male donors and USA female donors. Parameters were estimated simultaneously, using $W = 1$ year, and 95% confidence regions (CRs) are indicated.

Legend
- South Africa, males (SA, M): $\hat{\omega} = 219$ days, $\hat{\alpha} = 17\%$, $n = 162$ (--- 95% CR boundary)
- South Africa, females (SA, F): $\hat{\omega} = 185$ days, $\hat{\alpha} = 5\%$, $n = 120$ (— 95% CR boundary)
- USA, males (USA, M): $\hat{\omega} = 42$ days, $\hat{\alpha} = 29\%$, $n = 77$ (--- 95% CR boundary)
- USA, females (USA, F): $\hat{\omega} = 294$ days, $\hat{\alpha} = 6\%$, $n = 29$ (— 95% CR boundary)

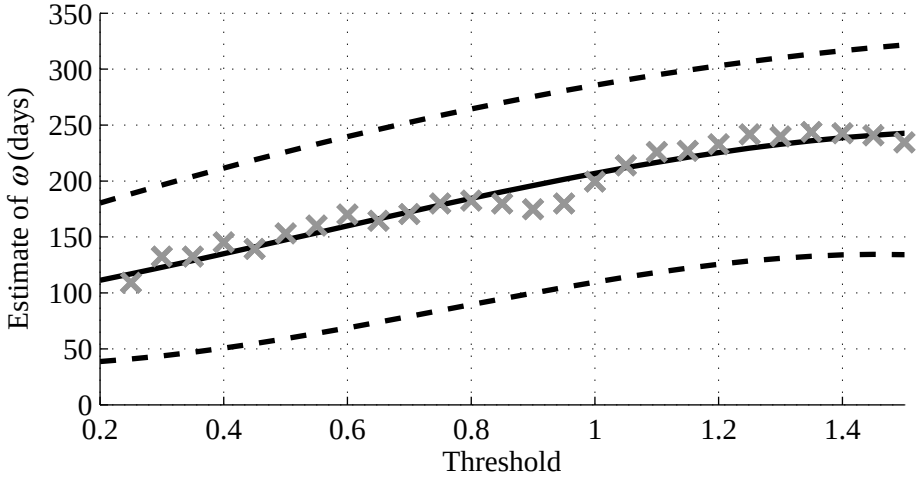
Optimisation and comparison of tests for recent infection

The ultimate objective is incidence estimation. The precision of the incidence estimator (and hence power to detect changes in incidence) increases with a larger MDRI and smaller FRR [11]. However, when optimising a biomarker-based test for recent infection by tuning the threshold distinguishing 'recent' from 'non-recent' infections, there is a fundamental trade-off between these two test characteristics as both increase with increasing threshold. Figure 3.4 shows the estimated test characteristics of LS-Vironostika, in the South African donor population, for a range of thresholds (using non-parametric estimation, and $W = 1$ year).

The precision of the incidence estimator, given the estimated test characteristics of LS-Vironostika, is compared to that achieved by a BED-like test for recent infection in Figure 3.5.[19] The BED-like test has an MDRI of $\omega = 155$ days and FRR of $\varepsilon = 5.6\%$ (with no uncertainty), as per BED package insert [166], and it is assumed that $\varepsilon = \alpha$ for LS-Vironostika. The coefficient of variation (CoV) of the incidence estimator [25] is calculated for a hypothetical population that has an HIV incidence of 1.5% per annum and HIV prevalence of 17.5%, loosely based on the South African adult population [167, 168]. Since the CoV ratio (LS-Vironostika to BED-like) is indistinguishable from 1, at all thresholds considered, the LS-Vironostika appears comparable to a BED-like test for recent infection. Additional data, such as captured during the follow-up of seropositive individuals awaiting treatment, could be used to explore whether systematic artefacts in the estimation occur (for example, from individuals progressing after $W = 1$ year).

[19]This early work, which captures the first article in this thesis that was published, touches on a number of ideas that were being further developed at the time. The use of the precision of the incidence estimator as a metric for optimising and comparing tests for recent infection is further discussed in Chapter 6.

A) MDRI, ω



B) Proportion of curves not reaching threshold, α

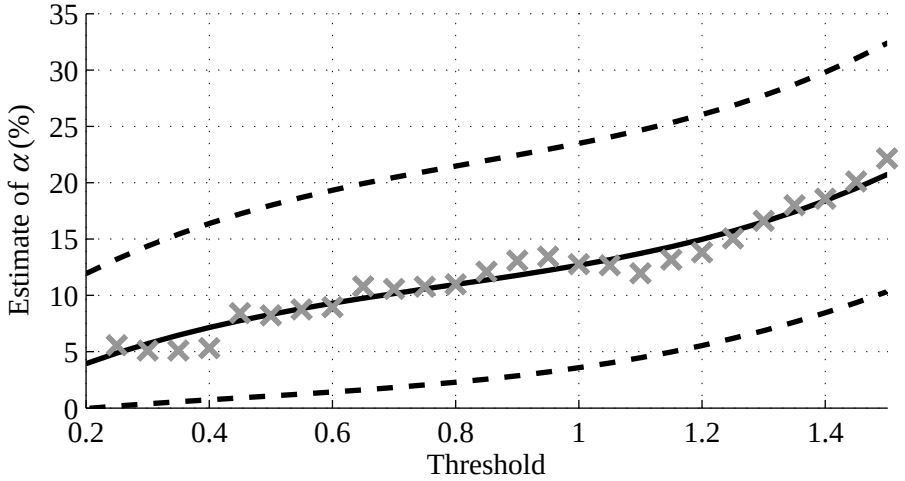


Figure 3.4: Estimated test characteristics for LS-Vironostika in the South African donor population as a function of test threshold

Estimates of ω and α for LS-Vironostika, based on the South African repeat donor sample, are shown for values of the threshold (discriminating 'recent' from 'non-recent' infection) between 0.2 and 1.5 (in SOD units). The parameters were estimated simultaneously, and estimates for A) ω (days) and B) α (%) are shown as functions of threshold, for $W = 1$ year. The minimum and maximum ω and α occurring in the 95% confidence region (CR) for these parameters are also displayed.

Legend:
- × Estimated test characteristic from simultaneous estimation
- — Estimated test characteristics from simultaneous estimation – smoothed (a cubic polynomial was fitted by least squares)
- ▪▪▪ Minimum and maximum value of test characteristic in 95% CR for characteristics from simultaneous estimation – smoothed

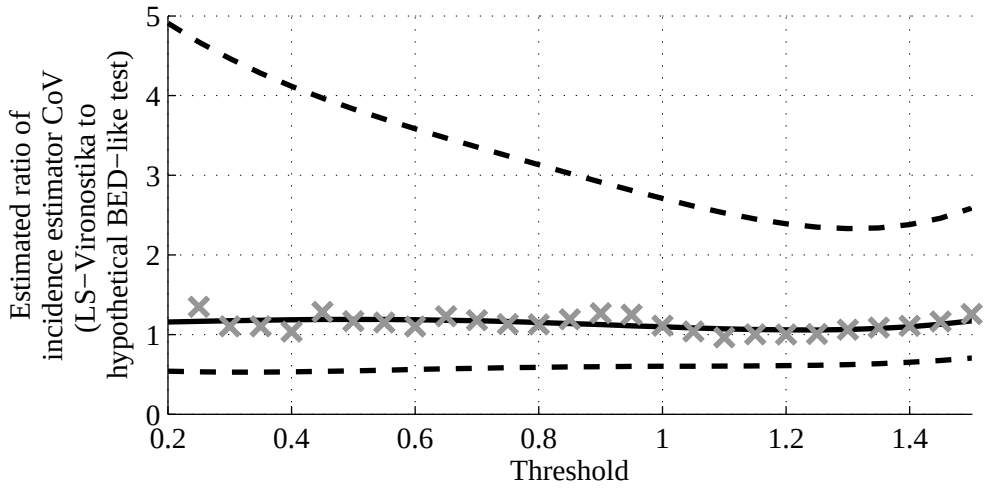


Figure 3.5: Relative performance of LS-Vironostika for incidence estimation
The ratio of the estimated CoV of the incidence estimator, for LS-Vironostika to that of a BED-like test for recent infection, is shown as a function of the LS-Vironostika test threshold. For LS-Vironostika, test characteristics estimated from the South African repeat donor sample were used to calculate the CoV, assuming $\varepsilon = \alpha$, while the BED-like test was assigned 'known' test characteristics of $\omega = 155$ days and $\varepsilon = 5.6\%$ [166]. In the hypothetical population considered, HIV incidence is 1.5% per annum and HIV prevalence is 17.5%.

Legend:
- × Estimated CoV ratio
- — Estimated CoV ratio – smoothed (a cubic polynomial was fitted by least squares)
- - - Minimum and maximum CoV ratio in 95% CR for test characteristics from simultaneous estimation – smoothed

Characterisation of LS-Vitros

Preliminary test characteristic estimates of the currently used LS-Vitros (using a threshold of 20), for the South African repeat donor population, are shown in Figure 3.6. Simultaneous (non-parametric) estimation of ω and α, for a range of W, was performed ($n = 108$, for $W = 1$ year, reduces to $n = 59$, for $W = 2.5$ years). Observed and expected proportions of seroconverters who were 'recently' infected were also compared (see Section 3.2.2).

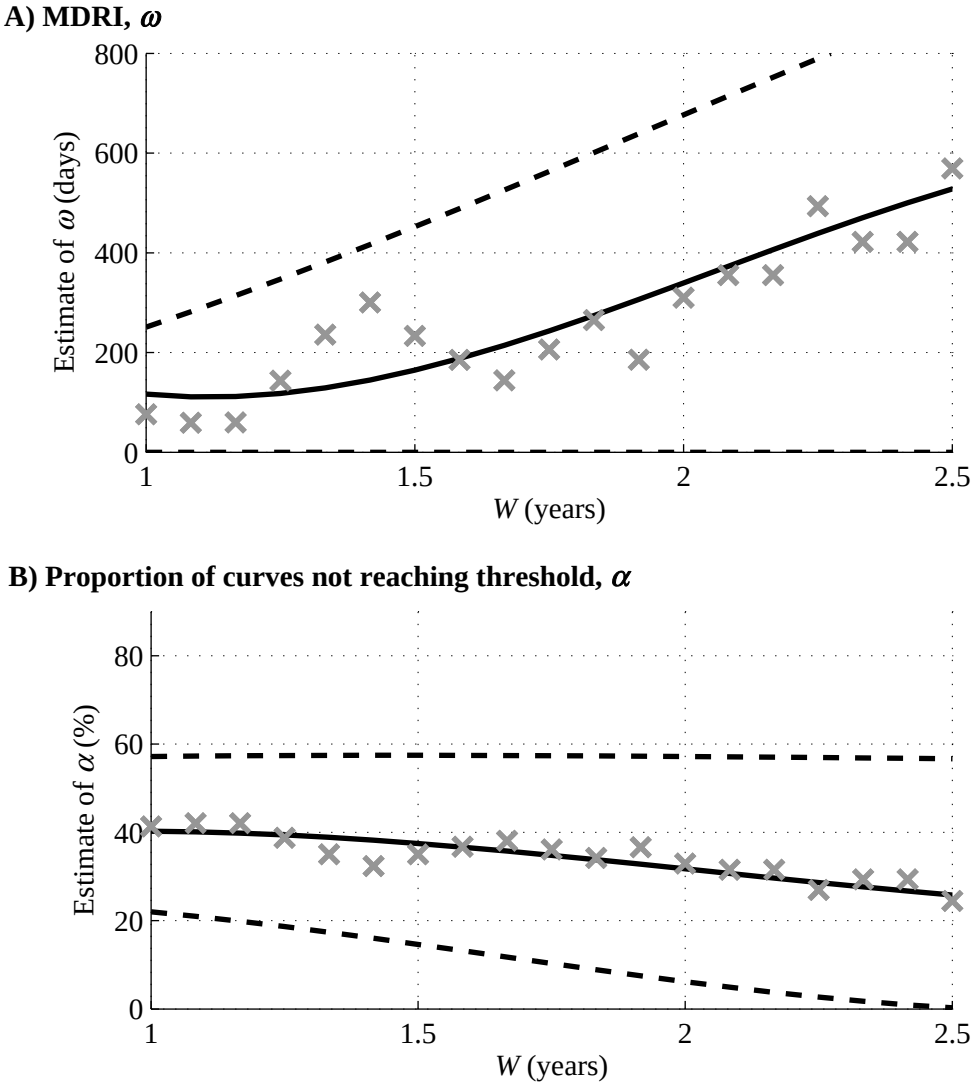


Figure 3.6: Estimated test characteristics for LS-Vitros in the South African repeat donor population as a function of W

Estimates of ω and α for LS-Vitros, based on the South African repeat donor sample, are shown for values of W between 1 and 2.5 years. The parameters were estimated simultaneously, and estimates for A) ω (days) and B) α (%) are shown as functions of W(years). The minimum and maximum ω and α occurring in the 95% confidence region (CR) for these parameters are also displayed

Legend:
- × Estimated test characteristic from simultaneous estimation
- — Estimated test characteristics from simultaneous estimation – smoothed (a cubic polynomial was fitted by least squares)
- --- Minimum and maximum value of test characteristic in 95% CR for characteristics from simultaneous estimation – smoothed

For LS-Vitros, the test characteristics were plotted as a function of the time cut-off W (Figure 3.6) as no estimates of W were found in the literature and the estimation appeared fairly sensitive to the choice of W for this dataset. Test characteristic estimates are highly uncertain and vary widely. This indicates the need for a larger dataset and an external estimate of α for a carefully selected time cut-off W (large enough for the biomarker curves to have crossed the threshold within a time W post infection) to get a more accurate and precise estimate of the MDRI. Such data and estimates are currently unavailable. A value of W that is too small would bias estimates of α upwards and ω downwards, under the above-mentioned analysis assumptions. However, as the value of W increases, the sample size reduces and ID intervals are more closely clustered together, decreasing the power to perform simultaneous estimation.

The large, albeit highly uncertain, estimates of α suggest that one should be cautious about the utility of LS-Vitros for incidence estimation, at this stage of the characterisation, while being mindful that α is not the FRR in Equation (3.1) if $S_R(t)$ is not (approximately) constant for $t > W$.[20]

3.1.4 Discussion

Traditionally, the characterisation of tests for recent infection (individual assays and multiple-test algorithms) has relied on the use of seroconversion panels. The scarcity of these panels is therefore an obstacle to the development of tests for incidence estimation. In this work, a source of more readily available specimens has been identified, and an approach for obtaining preliminary characterisations of tests using these specimens has been demonstrated. Further refinement of the characterisation of only the most promising tests may thereafter be performed, thus conserving precious longitudinal specimens (for MDRI estimation) and specimens from populations with known long-standing infections (for FRR estimation) for this purpose.

Utilising specimens from blood donors provides unique efficiencies as relatively large samples of seroconverters and high-volume specimens (125-250ml of plasma per seroconverter) are captured during routine blood collection procedures. Furthermore,

[20] In later work, as part of CEPHIA, a large dataset (which included longitudinal data) was produced and analysed to characterise LS-Vitros, and results are presented in Chapter 5.

specimens from seropositive subjects around the world are collected, thus providing data to investigate whether test characteristics are population-specific.

The method of estimating the test characteristics (MDRI and a proxy 'FRR' for parameter estimation purposes) does not require the follow-up of seroconverters. Moreover, by using data with large (pre-seroconversion) follow-up intervals, non-parametric estimation is supported. To obtain more accurate and precise estimates of the MDRI, an external estimate of the proportion of biomarker curves that do not reach the threshold is desirable, as well as insight into the maximum time seroconverters otherwise spend in the test-defined 'recent' state.

For incidence estimation, the utility of LS-Vironostika appears comparable to that of a BED-like test for recent infection, over the range of thresholds considered. The precision of the incidence estimator provides a criterion for both comparing tests and identifying optimal thresholds. While additional data is required for LS-Vitros, preliminary results suggest prudence when utilising the assay for incidence estimation.

The assumptions under which estimates are unbiased are strict. Potential for systematic bias in estimates, such as that arising from individuals remaining in the state of 'recent' infection for prolonged periods, or from non-uniformly distributed seroconversion times, should be explored using additional data. This method of estimating the test properties is not intended to provide final parameter estimates required for incidence estimation, but rather to provide cost-effective and efficient preliminary characterisations of tests using previously overlooked data. It is hoped that the concepts and tools demonstrated in this work will contribute to the resourceful characterisation, and subsequently focused development, of tests for recent infection for population-level incidence estimation.

3.2 Ancillary Analysis Details

3.2.1 The Test Characteristic Estimators

In the analyses presented in Section 3.1, two test characteristics were estimated, namely the MDRI, ω, and a proxy 'FRR', α, defined according to the analysis constructs of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24]. The maximum likelihood estimators for the test characteristics are derived below, and their distributional properties explored.

The function $S_R(t)$ denotes the probability that a seroconverter is in the state of 'recent' infection at time t after seroconversion, conditional on being alive.

For a given seroconverter, with inter-donation (ID) interval Δ between the last seronegative test and first seropositive test:

1. The random variable X denotes the result of the test for recent infection at the time of the first seropositive test, and has a probability mass function $f_X(x)$, where X equals 1 if the subject is 'recently infected and 0 if the subject is 'non-recently' infected.
2. The random variable Y captures the time since seroconversion at the time of the first seropositive donation, and has a probability density function $f_Y(y)$, where $0 \leq Y < \Delta$, and, in particular, is assumed to be uniformly distributed in the ID interval.

The joint probability function of X and Y is denoted by $f_{X,Y}(x, y)$, and the distribution of X conditional on Y by $f_{X|Y}(x|y)$.

The probability, p, that the seroconverter is classified as 'recently' infected at the time of the first seropositive donation is

$$\begin{aligned} p(\Delta) &= f_X(1) \\ &= \int_0^{\Delta} f_{X,Y}(1, \mathrm{y})\, \mathrm{d}y \\ &= \int_0^{\Delta} f_Y(y) f_{X|Y}(1|y)\, \mathrm{d}y \\ &= \frac{1}{\Delta}\int_0^{\Delta} S_R(t)\, \mathrm{d}t\,, \end{aligned} \tag{3.10}$$

since $f_Y(y) = 1/\Delta$ for all y such that $0 \leq y \leq \Delta$, and $f_{X|Y}(1|t) = S_R(t)$.

The likelihood, L, of all test classifications in a sample of n seroconverters is

$$L = \prod_{i=1}^{n} (p_i)^{x_i}(1 - p_i)^{1-x_i}, \tag{3.11}$$

where the subscript i denotes quantities relating to the i^{th} seroconverter in the sample and x denotes the observed values of X. The i^{th} seroconverter has ID interval Δ_i and therefore $p_i = p(\Delta_i)$.

The analyses of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24] assume that individual biomarker curves either cross the threshold (distinguishing 'recent' from 'non-recent' infection) and readings remain above it thereafter, or else fail to reach the threshold. Therefore $S_R(t)$ approaches some constant value, α, which is the proportion of biomarker curves that fail to cross the threshold, for t larger than some time cut-off W, and

$$S_R(t) = \alpha + (1-\alpha)S_{R'}(t)\ . \tag{3.12}$$

In the above-mentioned analyses, the MDRI, ω, is defined as the mean time under the threshold for those curves that do cross the threshold, which is described by $S_{R'}(t)$.

Substituting from Equation (3.12) into Equation (3.10), the probability that a seroconverter (with ID interval Δ) is 'recently' infected at the time of the first seropositive donation becomes

$$p(\Delta) = \alpha + (1-\alpha)\frac{1}{\Delta}\int_0^{\Delta} S_{R'}(t)\ \mathrm{d}t. \tag{3.13}$$

For $S_{R'}(t) = S_{R'}\left(t\middle|\underline{\phi}\right)$, L is a function of the unknown parameters $\underline{\phi}$ and α (if there is no input estimate for α), which are estimated to maximise L. The estimate of the MDRI is then

$$\widehat{\omega} = \int_0^{\infty} S_{R'}\left(t|\underline{\hat{\phi}}\right)\ \mathrm{d}t, \tag{3.14}$$

where $\underline{\hat{\phi}}$ is the estimate of $\underline{\phi}$.

This likelihood approach also facilitates non-parametric inference, by considering only individuals with large Δ. Since

$$S_{R'}(t) = 0\ (\text{since } S_R(t) = \alpha)\ \forall\ t > W, \tag{3.15}$$

if $\Delta > W$, then

$$\int_0^{\Delta} S_{R'}(t)\ \mathrm{d}t = \int_0^{\infty} S_{R'}(t)\ \mathrm{d}t = \omega \tag{3.16}$$

is the MDRI.

Substituting from Equation (3.16) into Equation (3.13), $p(\Delta)$ relies only on the test characteristics:

$$p(\Delta) = \alpha + (1-\alpha)\frac{\omega}{\Delta}, \tag{3.17}$$

and the likelihood function becomes

$$L = \prod_{i=1}^{n^*} (p_i)^{x_i} (1-p_i)^{1-x_i} \text{ where } p_i = p(\Delta_i) = \alpha + (1-\alpha)\frac{\omega}{\Delta_i} \tag{3.18}$$

and $n^* (\leq n)$ is the size of the sample consisting of only seroconverters with ID intervals larger than W (and the subscript i denotes quantities relating to the i^{th} subject in this smaller sample). The estimated test characteristics maximise the likelihood L, which is now a function of ω, and of α if there is no input estimate of α.

Simultaneous estimation of the test characteristics is less feasible in samples with closely clustered ID intervals. In the extreme case of $\Delta_i = \Delta^* > W$ for all i $(i = 1,2, \ldots n^*)$, simultaneous estimation is not possible as there are no unique estimates of ω and α which maximise the likelihood function. More specifically, the likelihood function,

$$L \propto \left(\alpha + (1-\alpha)\frac{\omega}{\Delta^*}\right)^{\sum_{i=1}^{n^*} x_i} \left(1 - \left(\alpha + (1-\alpha)\frac{\omega}{\Delta^*}\right)\right)^{\sum_{i=1}^{n^*}(1-x_i)}, \tag{3.19}$$

is maximised when

$$\frac{\sum_{i=1}^{n^*} x_i}{n^*} = \alpha + (1-\alpha)\frac{\omega}{\Delta^*}. \tag{3.20}$$

The left-hand side of the equation depends on the observed data. The maximum likelihood estimate for ω appearing in the right-hand side of the equation can be chosen arbitrarily by selecting a corresponding estimate of α that ensures the equation holds.

A maximum likelihood estimator, $\hat{\underline{\xi}}$, is asymptotically normally distributed around the true parameter value, $\underline{\xi}$, with variance equal to the inverse of the expected Fisher's Information Matrix, under regularity conditions [157]:

$$\hat{\underline{\xi}} \xrightarrow{d} N\left(\underline{\xi}, \left[-E\left(\frac{\partial^2 \ln\left(L\left(\underline{\xi}\right)\right)}{\partial \underline{\xi}^2}\right)\right]^{-1}\right) \text{ as } m \to \infty, \tag{3.21}$$

where m is the size of the sample used in the estimation procedure, $E(.)$ is the expected value and $L(.)$ is the likelihood function (and $\ln(L(.))$ the natural logarithm of the likelihood function). Using this property of maximum likelihood estimators, large sample

approximations for the distribution and moments of the test characteristic estimators follow.

When α is known, the distribution of the estimator for ω is

$$\widehat{\omega} \dot{\sim} N\left(\omega, \left[\sum_{i=1}^{n^*}\left(\frac{1-\alpha}{\Delta_i}\right)^2 \frac{1}{p_i(1-p_i)}\right]^{-1}\right). \tag{3.22}$$

When α is unknown, the bivariate distribution for the joint estimator for ω and α is

$$\begin{bmatrix}\widehat{\omega}\\ \widehat{\alpha}\end{bmatrix} \dot{\sim} N\left(\begin{bmatrix}\omega\\ \alpha\end{bmatrix}, \left[-E\begin{pmatrix}\frac{\partial^2 \ln(L)}{\partial\omega^2} & \frac{\partial^2 \ln(L)}{\partial\omega\partial\alpha}\\ \frac{\partial^2 \ln(L)}{\partial\omega\partial\alpha} & \frac{\partial^2 \ln(L)}{\partial\alpha^2}\end{pmatrix}\right]^{-1}\right), \tag{3.23}$$

where the covariance matrix is

$$\begin{bmatrix}\sum_{i=1}^{n^*}\left(\frac{1-\alpha}{\Delta_i}\right)^2 \frac{1}{p_i(1-p_i)} & \sum_{i=1}^{n^*}\frac{1}{\Delta_i p_i}\\ \sum_{i=1}^{n^*}\frac{1}{\Delta_i p_i} & \sum_{i=1}^{n^*}\left(1-\frac{\omega}{\Delta_i}\right)^2 \frac{1}{p_i(1-p_i)}\end{bmatrix}^{-1} \tag{3.24}$$

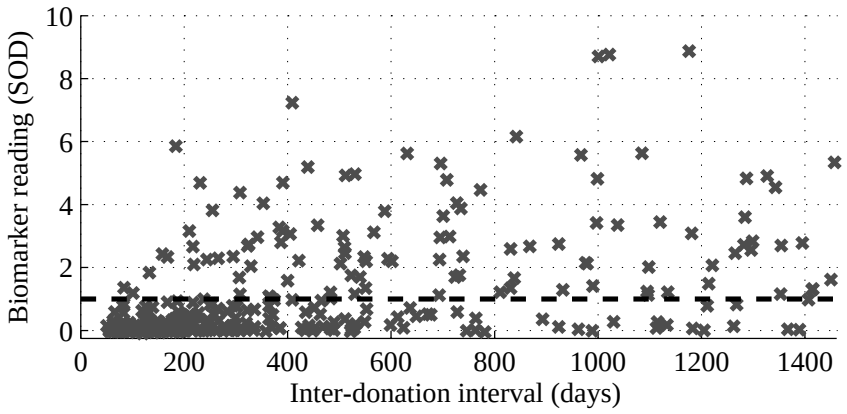
and $p_i = p(\Delta_i)$ and $L = L(\omega, \alpha)$ are given in Equation (3.18).

3.2.2 The Observed Data and Fitted Models

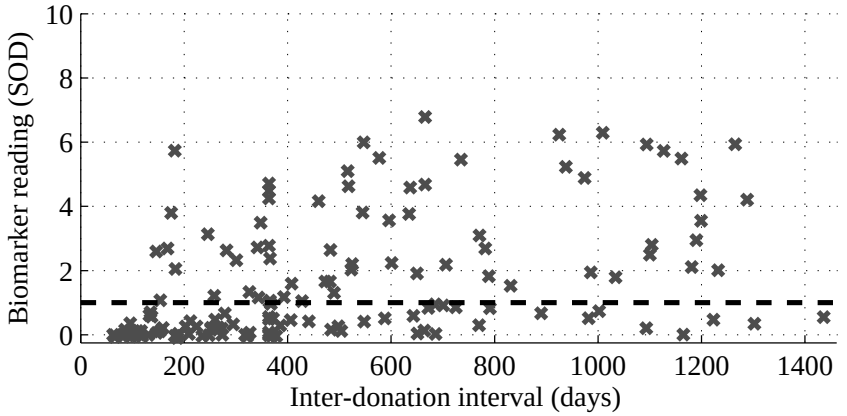
Plots of the data used to perform the analyses presented in Section 3.1 are provided, and the fitted models informally assessed by visually comparing observed and expected proportions of 'recent' results.

The data for each of the two tests for recent infection, LS-Vironostika and LS-Vitros, are shown in Figure 3.7, stratified by country (South Africa and the USA). Each seroconverting blood donor's biomarker reading at the time of the first seropositive donation and inter-donation interval (time between the last seronegative and first seropositive donation) are shown.

A) LS-Vironostika, South Africa



B) LS-Vironostika, USA



C) LS-Vitros, South Africa

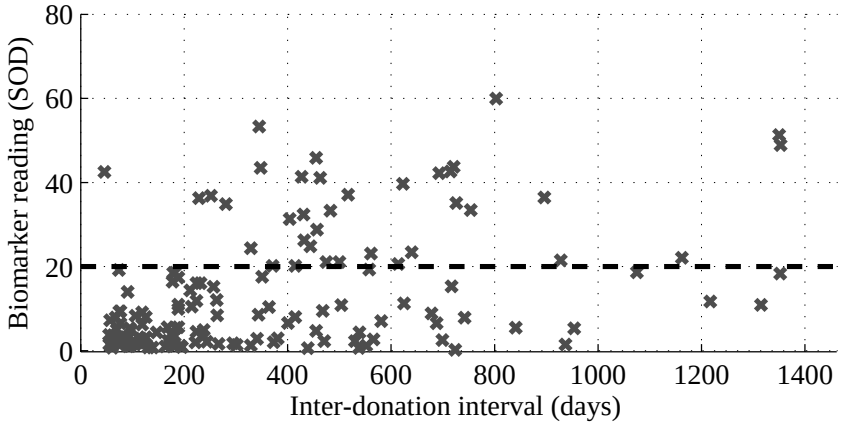


Figure 3.7: Observed biomarker readings for LS-Vironostika and LS-Vitros in seroconverting blood donors, and corresponding inter-donation intervals

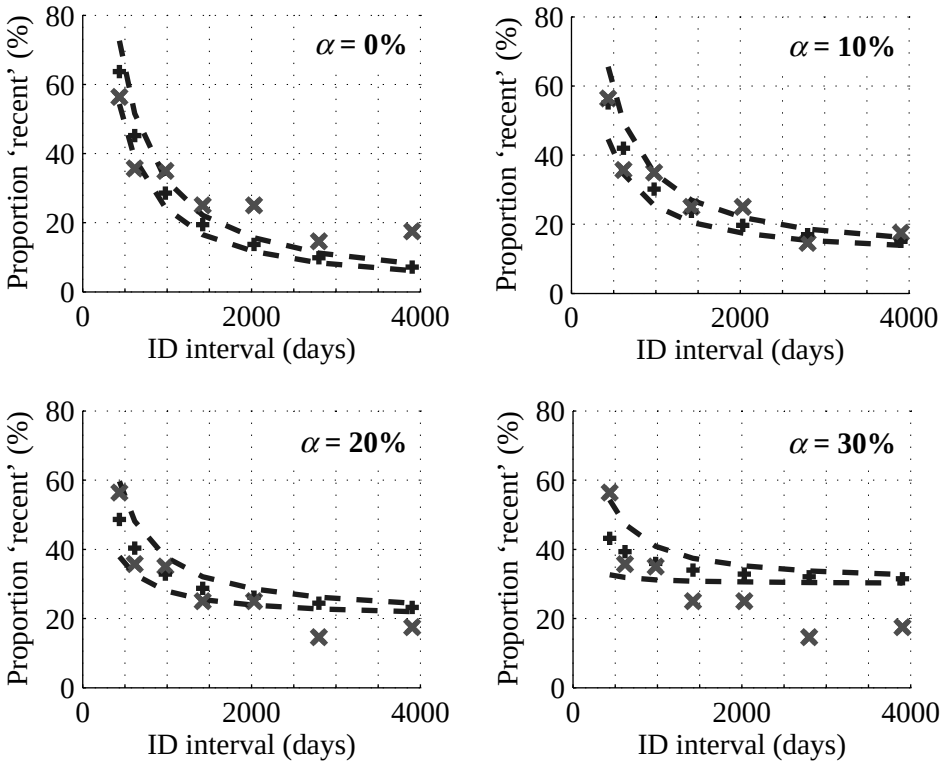
Each subject's biomarker reading at the first seropositive donation, expressed as a standardised optical density (SOD), and inter-donation interval (time between last seronegative and first seropositive donation, in days) are shown (for inter-donation intervals less than 4 years), for A) LS-Vironostika for donors in South Africa, B) LS-Vironostika for donors in the USA and C) LS-Vitros for donors in South Africa. Measurements below test thresholds (horizontal dashed lines) produce 'recent' classifications.

The method of maximum likelihood, outlined in Section 3.2.1 above, was used to characterise the tests. In Figures 3.8 to 3.11, observed proportions of seroconverters who were 'recently' infected at the first seropositive donations are compared to expected proportions, for each primary analysis presented in Section 3.1. For each analysis, firstly, the MDRI, ω, was estimated assuming a known α, and secondly, simultaneous estimation of ω and α was performed. Non-parametric estimation was applied, using data on seroconverters with inter-donation (ID) intervals larger than the chosen time cut-off W.

The probability of testing 'recently' infected declines with increasing ID interval. Subjects with similar ID intervals were therefore grouped together (at least 20 subjects per group), and, for each group, the observed and expected proportions were plotted against the average ID interval. Expected proportions were obtained by substituting estimated (or input) test characteristics into Equation (3.17), and averaging the probabilities obtained for subjects in the group. When assuming a known α (Part A of figures), the 95% confidence interval limits for the expected proportion were obtained by instead substituting the 95% confidence interval limits for ω into Equation (3.17). When simultaneously estimating ω and α (Part B of figures), the plotted limits for the expected proportion indicate the minimum and maximum values for this proportion that were obtained when considering all pairs of values for the test characteristics within the 95% confidence region for ω and α.

Figure 3.8 and Figure 3.9 capture results for LS-Vironostika, for the South African ($n = 282$) and American repeat donor ($n = 106$) samples, respectively, using $W = 1$ year. In Figure 3.10 and Figure 3.11, observed and expected proportions are compared for LS-Vitros in the South African donor sample for $W = 1$ year ($n = 108$) and $W = 2.5$ years ($n = 59$), respectively.

A) Estimation of ω using input α



B) Simultaneous estimation of ω and α ($\hat{\omega} = 199$ days and $\hat{\alpha} = 13\%$)

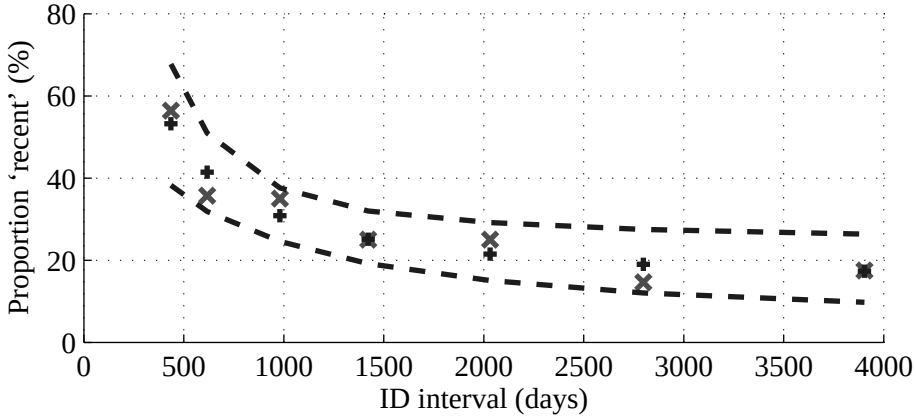


Figure 3.8: Comparison of observed and expected proportions of 'recently' infected seroconverters for LS-Vironostika, South Africa, $W = 1$ year

Observed and expected proportions of 'recently' infected subjects, at first seropositive donations, are shown as a function of inter-donation (ID) intervals (days) for LS-Vironostika, for the South African repeat blood donor sample. Expected proportions are based on estimated test characteristics when A) estimating only ω and assuming a known α, and B) simultaneously estimating ω and α, using $W = 1$ year. Proportions are plotted against average ID intervals for groups of at least 20 subjects. 95% CI limits or uncertainty bounds for expected proportions are also indicated.

Legend: × Observed proportion of 'recent' infection
+ Expected proportion of 'recent' infection
--- 95% CI limits / uncertainty bounds for expected proportion

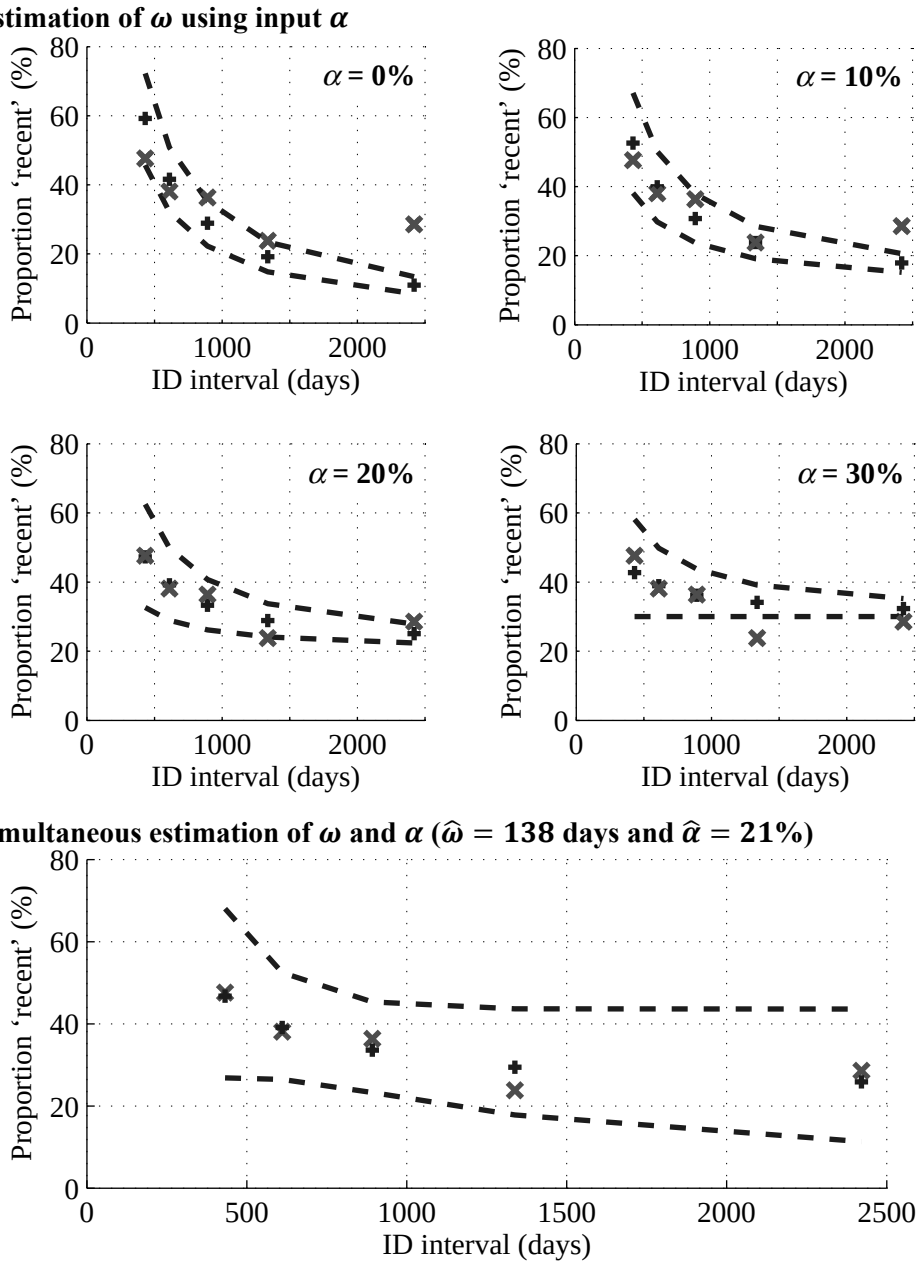
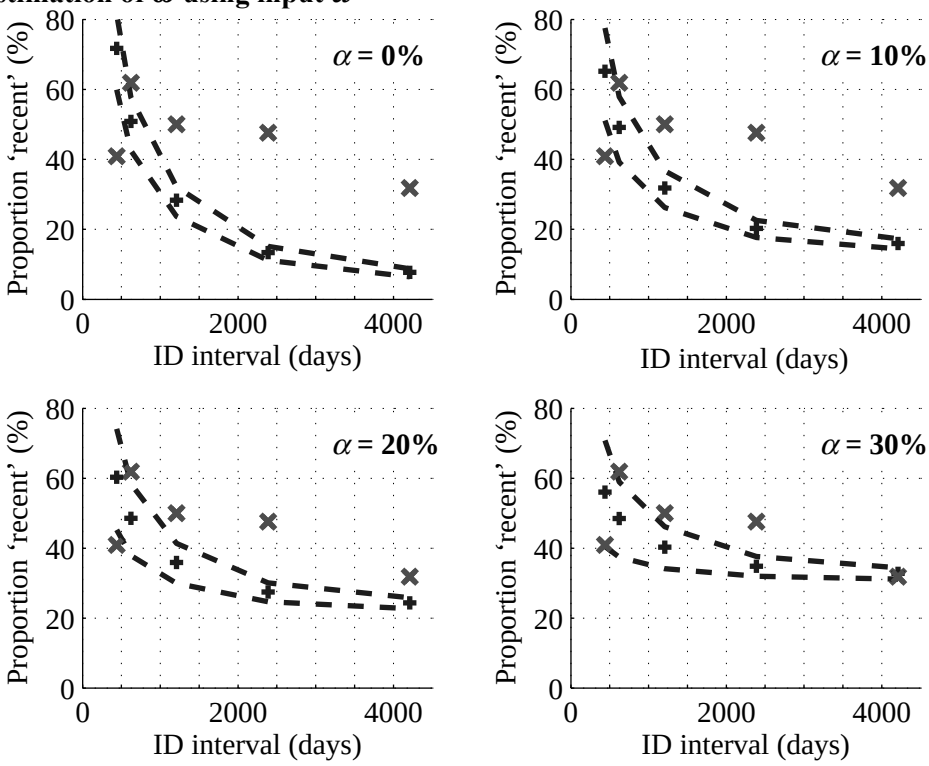


Figure 3.9: Comparison of observed and expected proportions of 'recently' infected seroconverters for LS-Vironostika, USA, $W = 1$ year

Observed and expected proportions of 'recently' infected subjects, at first seropositive donations, are shown as a function of inter-donation (ID) intervals (days) for LS-Vironostika, for the American repeat blood donor sample. Expected proportions are based on estimated test characteristics when A) estimating only ω and assuming a known α, and B) simultaneously estimating ω and α, using $W = 1$ year. Proportions are plotted against average ID intervals for groups of at least 20 subjects. 95% CI limits or uncertainty bounds for expected proportions are also indicated.

Legend: × Observed proportion of 'recent' infection
+ Expected proportion of 'recent' infection
--- 95% CI limits / uncertainty bounds for expected proportion

A) Estimation of ω using input α



B) Simultaneous estimation of ω and α ($\hat{\omega} = 76$ days and $\hat{\alpha} = 41\%$)

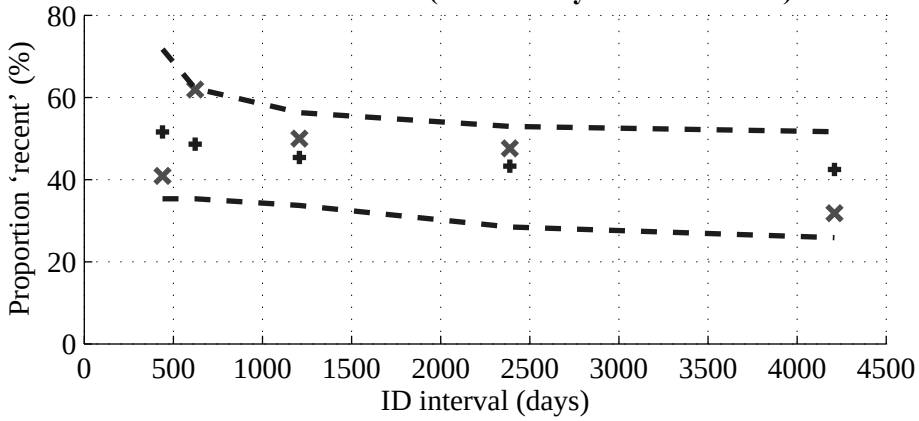


Figure 3.10: Comparison of observed and expected proportions of 'recently' infected seroconverters for LS-Vitros, South Africa, $W = 1$ year

Observed and expected proportions of 'recently' infected subjects, at first seropositive donations, are shown as a function of inter-donation (ID) intervals (days) for LS-Vitros, for the South African repeat blood donor sample. Expected proportions are based on estimated test characteristics when A) estimating only ω and assuming a known α, and B) simultaneously estimating ω and α, using $W = 1$ year. Proportions are plotted against average ID intervals for groups of at least 20 subjects. 95% CI limits or uncertainty bounds for expected proportions are also indicated.

Legend: × Observed proportion of 'recent' infection
+ Expected proportion of 'recent' infection
--- 95% CI limits / uncertainty bounds for expected proportion

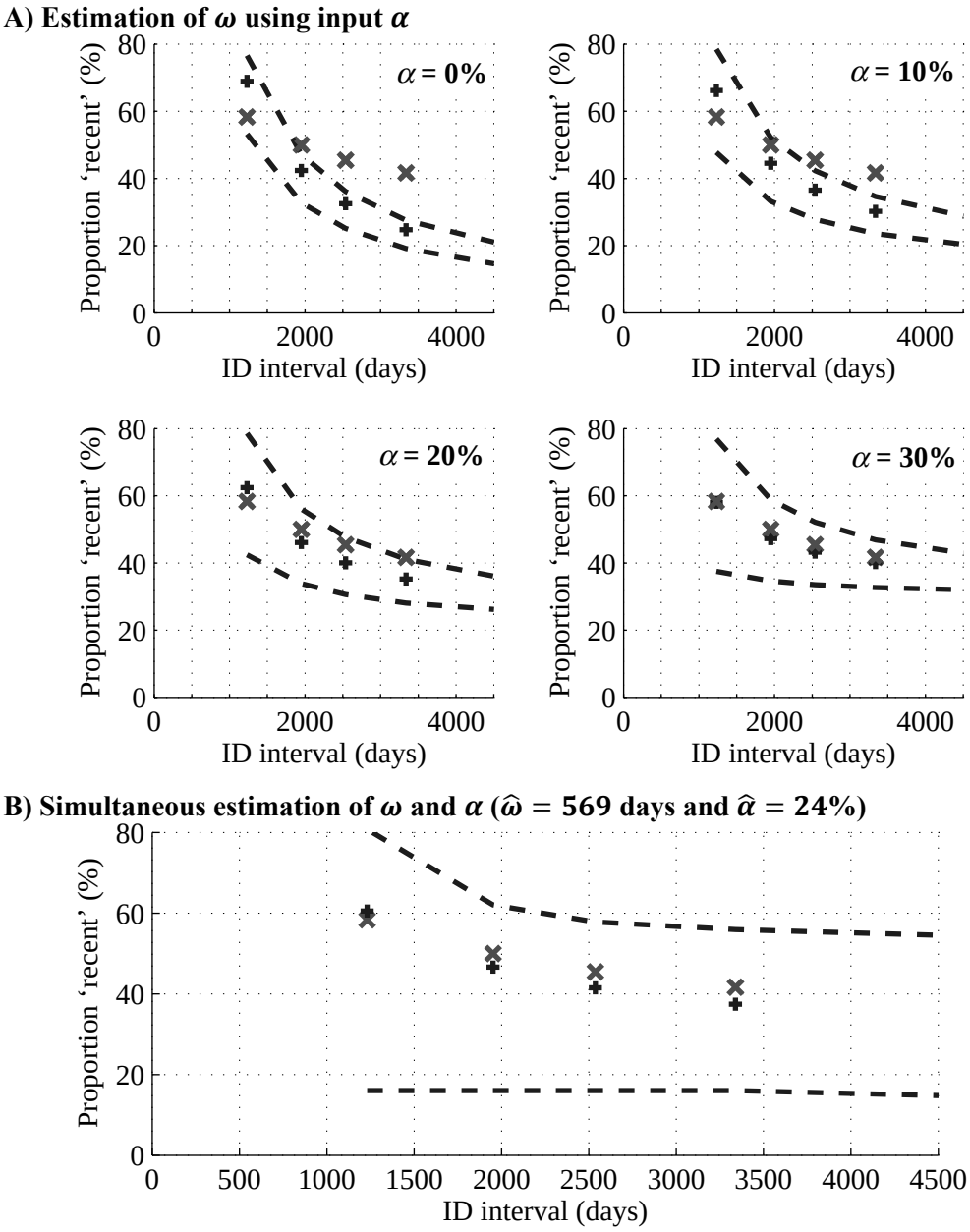


Figure 3.11: Comparison of observed and expected proportions of 'recently' infected seroconverters for LS-Vitros, South Africa, $W = 2.5$ years

Observed and expected proportions of 'recently' infected subjects, at first seropositive donations, are shown as a function of inter-donation (ID) intervals (days) for LS-Vitros, for the South African repeat blood donor sample. Expected proportions are based on estimated test characteristics when A) estimating only ω and assuming a known α, and B) simultaneously estimating ω and α, using $W = 2.5$ years. Proportions are plotted against average ID intervals for groups of at least 20 subjects. 95% CI limits or uncertainty bounds for expected proportions are also indicated.

Legend: × Observed proportion of 'recent' infection
+ Expected proportion of 'recent' infection
--- 95% CI limits / uncertainty bounds for expected proportion

The expected proportions of 'recent' results are sensitive to the input value for α as well as W. Expected proportions are relatively well aligned to observed proportions when α is estimated together with ω (rather than being provided as an input), but uncertainty in the estimation procedure becomes large when the parameters are jointly estimated. This highlights the need for external and accurate data to guide choices of α and W. Also, in this work, the probability of testing 'recently' infected, conditional on being alive, is understood to be approximately constant (and equal to α) for all times after seroconversion greater than W, based on the analysis constructs of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24]. The behaviours of LS-Vironostika and LS-Vitros may violate this assumption, potentially leading to misalignment between observed and expected proportions.

3.2.3 Parametric Versus Non-Parametric Estimation

The need for parametric assumptions about the probability of testing 'recently' infected as a function of time since seroconversion is circumvented by using only data with sufficiently large ID intervals (see Section 3.2.1). While this protects against bias arising from poor parametric assumptions, the sample size is reduced.

The test characteristics of LS-Vironostika, in the South African repeat donor population, were therefore also estimated using all data and a number of parametric assumptions, captured by specifying various forms for $S_{R'}(t) = S_{R'}(t|\underline{\phi})$, where $\underline{\phi}$ is a vector of parameters to be estimated from the data. The six assumed forms for $S_{R'}(t|\underline{\phi})$ are plotted in Figure 3.12. For simplicity, by design, $\underline{\phi} = \omega$ for each form. Estimates of the MDRI, ω, using the various parametric assumptions, are provided in Table 3.1. Results of a chi-squared goodness of fit test [158], used to assess agreement between data and assumptions, are also provided. Widely varying estimates of ω were obtained, even after discarding those estimates for which data and assumptions did not agree.

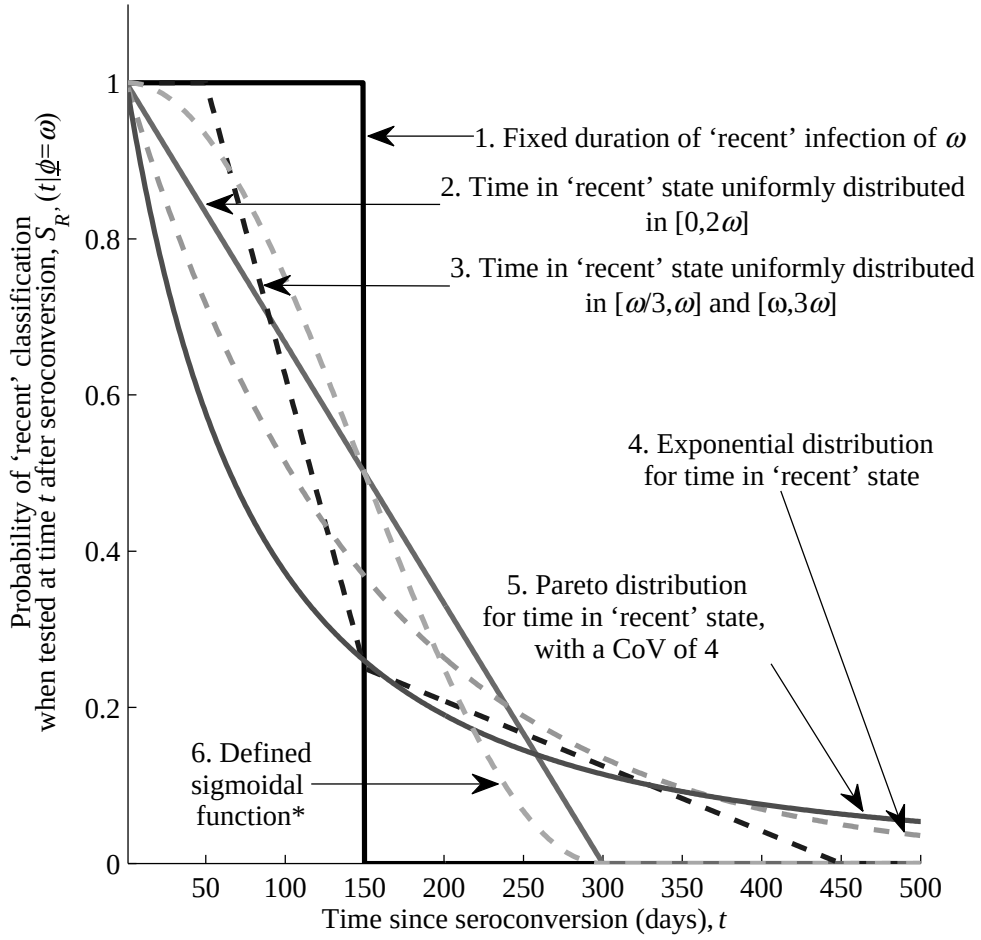


Figure 3.12: Parametric forms for $S_{R'}(t|\underline{\phi})$ used in MDRI estimation and data generation

Each of the six forms for $S_{R'}(t|\underline{\phi})$ used in the analysis is plotted as a function of time since seroconversion (days), where $S_{R'}(t|\underline{\phi})$ is the probability of being 'recently' infected when tested at time t after seroconversion (for those individuals who do transition out of the 'recent' state). The function $S_{R'}(t|\underline{\phi})$ contains the parameter $\underline{\phi}$, which is estimated to maximise the likelihood of observed data for test characterisation purposes, or for which a value is specified for data generation purposes. By design, $\underline{\phi} = \omega$ for each form. The functions are plotted using $\underline{\phi} = \omega = 150$ days. *Form 6 uses a (suitably scaled and shifted) sine function to obtain an s-shaped curve that reaches zero at time 2ω after seroconversion.

	Input α (%)			
	0	**5**	**10**	**15**
Estimated ω, in days (95% CI)				
Parametric form 1	*84 (82-84)	*83 (81-84)	*83 (81-84)	*83 (80-84)
Parametric form 2	316 (266-374)	278 (233-333)	251 (209-301)	229 (188-276)
Parametric form 3	*237 (217-249)	228 (204-246)	219 (192-242)	208 (180-235)
Parametric form 4	*429 (355-520)	379 (309-464)	338 (273-418)	303 (242-378)
Parametric form 5	650 (528-802)	*579 (464-721)	516 (409-649)	*461 (361-585)
Parametric form 6	268 (232-309)	241 (207-281)	221 (188-259)	205 (173-242)
Non-parametric	274 (234-313)	245 (199-289)	216 (165-266)	186 (132-241)
Goodness of fit p-value[a]				
Parametric form 1	0.00	0.00	0.00	0.00
Parametric form 2	0.29	0.10	0.56	0.73
Parametric form 3	0.03	0.57	0.65	0.80
Parametric form 4	0.02	0.12	0.16	0.09
Parametric form 5	0.09	0.02	0.08	0.03
Parametric form 6	0.05	0.56	0.69	0.90

[a] Null hypothesis: The data is consistent with the assumed form for $S_{R'}(t|\underline{\phi})$

Table 3.1: Estimated mean duration of recent infection for LS-Vironostika, South Africa, using various parametric assumptions

Estimates of the MDRI, ω (days), for LS-Vironostika in the South African repeat donor population are shown, using both parametric and non-parametric estimation approaches and an input value of α (%). 95% CI limits are also provided. For the parametric estimation, each of the six forms for $S_{R'}(t|\underline{\phi})$ shown in Figure 3.12 was assumed in turn (Parametric form 1 to 6) and all data were included. P-values from a chi-squared goodness of fit test used to assess agreement between data and parametric assumptions are also shown, and an estimate of ω that corresponds to a p-value below 0.05 is indicated by an asterisk (*). For non-parametric estimation, only seroconverters with ID intervals larger than $W = 1$ year were included in the analysis.

Since the underlying dynamics of the data are unknown, the extent of any bias in the results is unclear. Simulated data was therefore used to investigate the trade-off between the increased precision from larger samples and increased potential for bias from poor parametric assumptions, when moving from a non-parametric to parametric approach. A number of datasets were generated, each consisting of ID intervals and test classifications for 500 seroconverters. For each seroconverter, the ID interval was drawn from a (kernel density) non-parametric distribution that was fitted to the ID intervals contained in the real-world dataset (for LS-Vironostika, South Africa); the infection time was drawn from a uniform distribution spanning the subject's ID interval; and the test classification was generated from a chosen specification of $S_{R'}(t|\underline{\phi})$ and $\alpha = 0\%$. Each of the six forms for $S_{R'}(t|\underline{\phi})$ in Figure 3.12, using $\underline{\phi} = \omega = 150$ days, was used to generate 100 datasets. For each simulated dataset, seven estimates of ω were obtained: six by assuming each of the parametric forms for $S_{R'}(t|\underline{\phi})$ in turn and using all data in the estimation procedure, and one by non-parametric estimation using only ID intervals greater than $W = 1$ year and assuming $\alpha = 0\%$. For each parametric estimation, a chi-squared goodness of fit test was performed to assess agreement between data and parametric assumptions. When data were generated from a form of $S_{R'}(t|\underline{\phi})$ that is non-zero at times greater than $W = 1$ year, underestimation of ω is expected. To assess the performance of MDRI estimation, estimates of ω were compared to its true value, which was 150 days throughout this investigation, and the ability to distinguish between correct and incorrect parametric assumptions was considered.

The results of the investigation, summarised in Table 3.2, indicate that, although moving to a parametric approach allows all data to be exploited, there is the potential for introducing large bias in estimates from poor parametric assumptions. The results of the goodness of fit tests suggest that it is challenging to distinguish between appropriate and poor parametric assumptions, using a given dataset. The average 95% confidence interval (CI) widths, when using the correct parametric assumption or the non-parametric approach, are also provided (Table 3.2). CIs were obtained using large sample maximum likelihood theory (in particular, properties of the deviance statistic) [157]. The increased CI width when moving to the non-parametric approach illustrates the loss of precision incurred when discarding data with insufficiently large ID intervals.

	Parametric form used for data generation					
	1	**2**	**3**	**4**	**5**	**6**
Average estimated ω, in days (95% CI width)[a]						
Parametric form 1	153 (27)	-	102	-	-	108
Parametric form 2	187	153 (51)	154	122	90	164
Parametric form 3	197	129	151 (42)	105	78	154
Parametric form 4	-	191	197	151 (59)	110	207
Parametric form 5	-	273	286	212	151 (68)	295
Parametric form 6	174	138	140	106	84	151 (44)
Non-parametric	150 (79)	153 (79)	148 (79)	144 (78)	128 (74)	149 (79)
Percentage of estimates rejected						
Parametric form 1	2	100	94	100	100	96
Parametric form 2	87	2	10	19	68	5
Parametric form 3	23	67	3	80	92	16
Parametric form 4	100	7	37	5	13	35
Parametric form 5	100	22	68	9	3	79
Parametric form 6	28	32	7	77	97	2

[a] Includes only estimates not rejected (p-value≥0.05) by goodness of fit test

Table 3.2: Summary of results from investigation of parametric versus non-parametric estimation using simulated data

Average estimates of the MDRI, ω (days), and percentages of estimates rejected by goodness of fit tests (p-values below 0.05) are shown for groups of 100 simulated datasets. The parametric form for $S_{R'}(t|\underline{\phi})$ used to generate the datasets is captured by the table columns (Parametric form 1 to 6, see Figure 3.12). The MDRI was estimated parametrically (assumed form for $S_{R'}(t|\underline{\phi})$ captured by table rows) and non-parametrically (using W=1 year). For each group of datasets and estimation method, the average estimate of ω was calculated after excluding those estimates rejected by the goodness of fit test, and the average width of the 95% CI is provided when the assumed parametric form is correct or non-parametric estimation was applied. In both data generation and MDRI estimation, $\alpha = 0\%$.

3.3 Further Applications

Two further applications of the methodology presented are described below. While this approach is not intended to provide highly accurate and precise estimates of the MDRI, it serves an important role in obtaining initial estimates using previously neglected sources of specimens when that is all that is available.

3.3.1 A Biomarker for Recent Infection Using SMARTube™

The sensitivity of an HIV diagnostic test (probability of correctly detecting virus) is expected to increase from zero to, ideally, 100% over some short period after HIV transmission. An antibody-based diagnostic test is unable to detect HIV infection in the period between acquiring the infection and seroconversion, where the seroconversion time captures when the antibody response reaches a measurable level. Therefore, an HIV-positive subject tested in this 'seronegative window period' will produce a false-negative HIV diagnosis. To reduce this period, *SMART Biotech Ltd* [169] developed the *Stimulating Maximal Antibody Response Tube* (SMARTube^{TM}). The technology stimulates in-vivo primed immune cells to produce antibodies in-vitro. By incubating a specimen in a SMARTube^{TM} before applying the HIV diagnostic test, antibodies reach detectable levels sooner after infection [170, 171].

As illustrated in Figure 3.13 (Part A), stimulation of antibody in an incubated specimen is expected to fade over time after infection. This suggests that a measure of the increase in signal could provide a novel biomarker for recent infection. In this analysis, one such measure is defined and its potential explored: the '*Stimulation Index*' (SI) is the ratio of stimulated to unstimulated antibody levels (Part B of Figure 3.13). A test for recent infection is constructed by introducing a threshold, where an SI measurement *above* the threshold is interpreted as indicating 'recent' infection. As the threshold is varied, there is the familiar trade-off between test characteristics: while a lower threshold improves (increases) the MDRI, a higher threshold typically improves (decreases) the FRR.

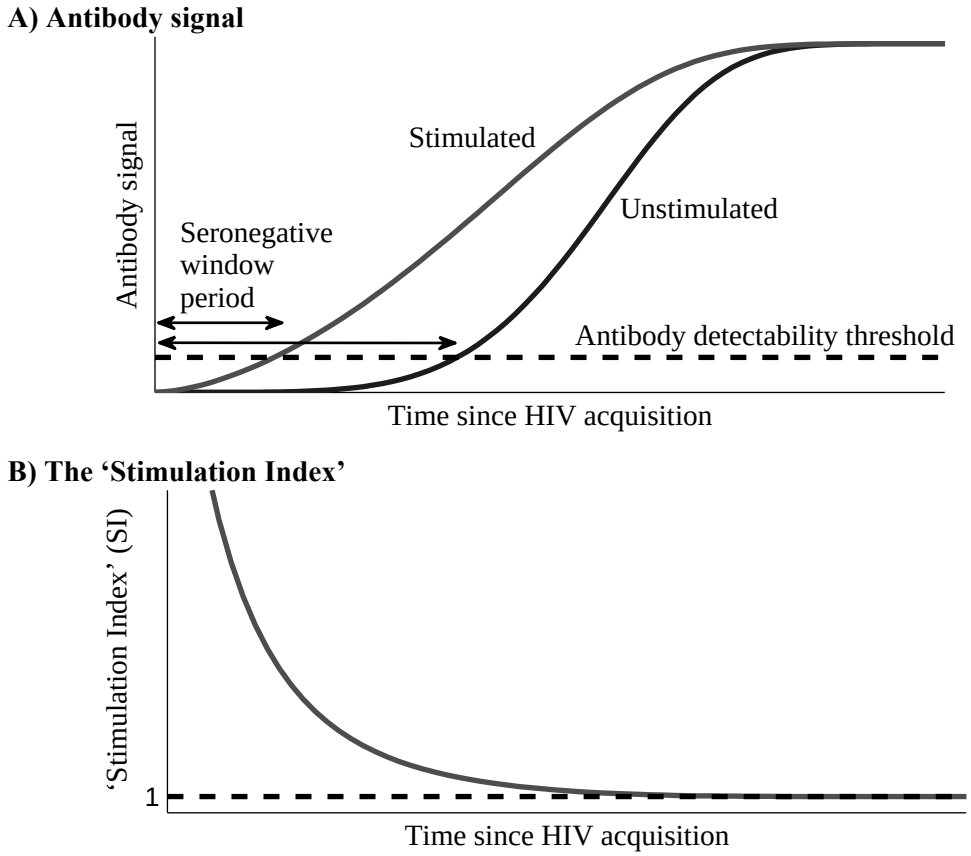


Figure 3.13: The impact of SMARTube™ on antibody signal, and the 'Stimulation Index'

The schematic diagram illustrates the increase in antibody signal resulting from incubating a specimen in a SMARTube™, and the proposed measure of this increase, each as a function of time since HIV acquisition. A) Antibody signal, as measured by a standard semi-quantitative HIV diagnostic test, is shown, without ('unstimulated') and with ('stimulated') the use of a SMARTube™, and the difference in the 'seronegative window period' reflected. B) The 'Stimulation Index' (SI) is shown, and is the ratio of the stimulated to unstimulated antibody levels.

As is typically the case for new biomarkers, very little data to characterise the test were available, and opportunities to generate further, tailored, data would be contingent on demonstrating sufficient promise of the candidate biomarker. An initial, preliminary characterisation of the test was therefore performed, using existing available specimens. The specimens were obtained from subjects in various regions of China by the *Centers for Disease Control and Prevention* (CDC) and the *National Institute for the Control of Pharmaceutical and Biological Products in Beijing* (NICPBP). Intervals between HIV tests were large (the exact sizes were unknown) and there was little background information on the subjects, and therefore such specimen sets have typically been overlooked. Data on the SI biomarker were generated by testing (subsets of) specimens

using two semi-quantitative HIV diagnostic tests, developed by *Abbott Diagnostics* and *Beijing Wantai Biological Pharmacy*. Each specimen was tested both using the standard diagnostic procedure (to obtain an 'unstimulated' antibody measurement) and after incubation in a SMARTube^{TM} (to obtain a 'stimulated' antibody measurement), and SI values were calculated.

The FRR, ε, was estimated using specimens drawn from non-recently infected subjects attending *CDC* clinics ($n = 59$ for Abbott and $n = 73$ for Wantai). The decreasing FRR, as function of increasing threshold, is illustrated in Figure 3.14. Results suggest that a suitably low FRR may be achievable by a choice of threshold of around 1.2 or larger.

The MDRI, ω, was estimated using specimens collected during surveys of a high-risk, injecting-drug-using population ($n = 57$ for Wantai). There was no follow-up of subjects, and specimens included in the analysis were drawn at the times of first HIV-positive tests. Since the times between last HIV-negative and first HIV-positive tests were unknown, these were crudely all assumed to be Δ^* in the analysis. The maximum likelihood approach presented in Sections 3.1 and 3.2 above was used to estimate the MDRI, for a range of thresholds beginning at 1.2 and assuming a zero FRR (based on the results above), and for a range of values of Δ^* beginning at 1 year (assuming $\Delta^* > W$, where W is the maximum time after infection that an individual may remain in the 'recent' state). The point estimate for ω, as a function of threshold and inter-test interval Δ^*, is shown in Figure 3.15, and increases with decreasing threshold or increasing Δ^*. At a high threshold of 1.5 (SI units) and small Δ^* of 1 year, ω is estimated to be 0.2 years (70 days), and ω increases to 1.1 years (404 days) when the threshold decreases to 1.2 and Δ^* increases to 3 years.

Hypothesis tests for superiority or non-inferior of the test characteristics (compared to those of existing tests or reference values) would have little statistical power due to the small sample sizes. Instead, for the FRR, data were used to assess the null hypothesis $\varepsilon = 5\%$ against the alternative hypothesis $\varepsilon > 5\%$. Even at a relatively low threshold of 1.1 (the estimate for ε increases as the threshold decreases), the null hypothesis was not rejected (p-values of 0.80 and 0.71 for Abbott and Wantai kits, respectively). Similarly, for the MDRI, data were used to evaluate the null hypothesis $\omega = 155$ days against the alternative hypothesis $\omega < 155$ days, where 155 days is the MDRI of the (then) widely used BED assay as per package insert [166]. Even at a high threshold of 1.4 and conservatively assuming an FRR of 5% (the estimate for ω decreases as the threshold

or input FRR increases), the null hypothesis was not rejected (p-value of at least 0.46 for an assumed Δ^* of at least 2 years for the Wantai kit). These results suggest that there is a lack of evidence that the FRR is particularly large or MDRI is small. This analysis suggests that efforts should be made to capture a larger set of specimens, appropriate for test characterisation and including relevant background information, to further investigate the potential of using SMARTube^{TM} to construct a biomarker for recent infection. More generally, the investigation strengthens the case for broadening the spectrum of biomarkers conventionally considered.

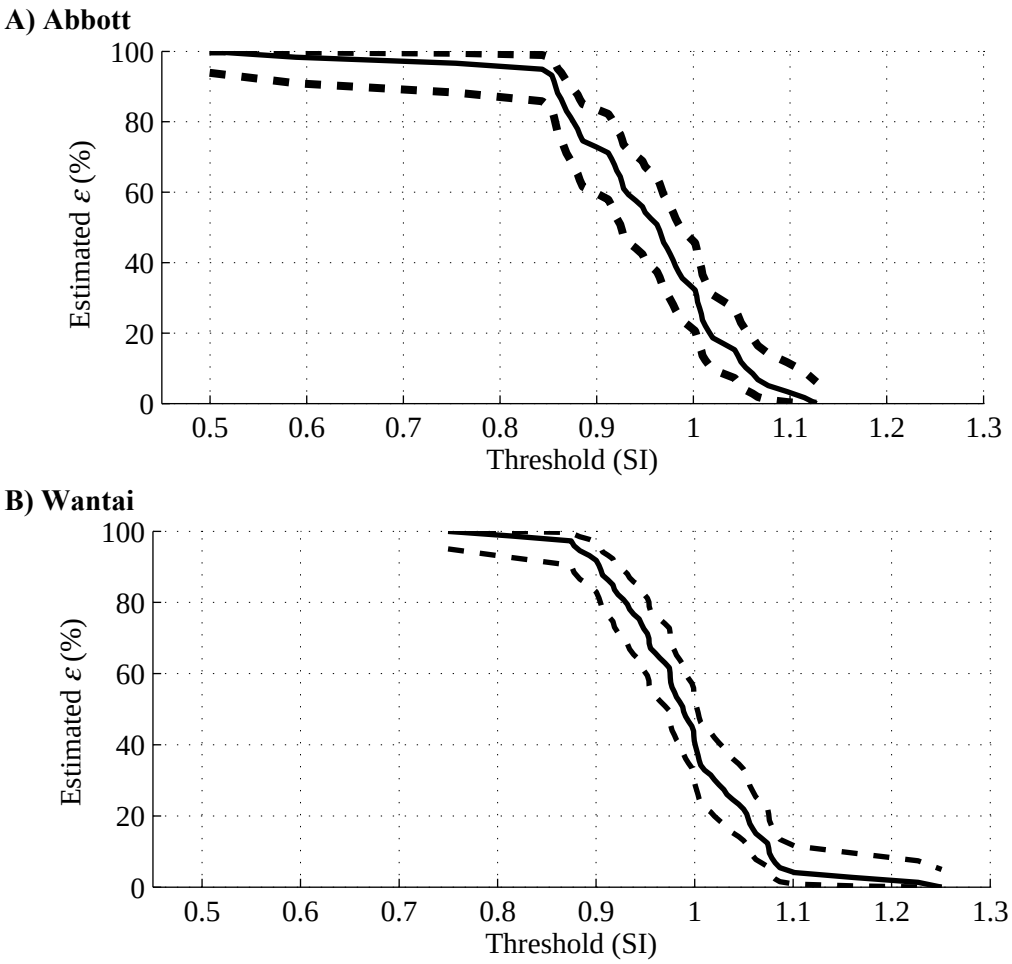


Figure 3.14: Estimated false-recent rate of a test using the 'Stimulation Index' biomarker, as a function of the 'recent'/'non-recent' threshold and stratified by HIV diagnostic test

The estimate of the FRR, ε (%), and its 95% (Clopper-Pearson) confidence interval are shown for a recent infection test that is based on the 'Stimulation Index' (SI), as a function of test threshold (above which an SI value indicates 'recent' infection). Antibody levels were measured using HIV diagnostic tests by A) *Abbott Diagnostics* and B) *Beijing Wantai Biological Pharmacy*.

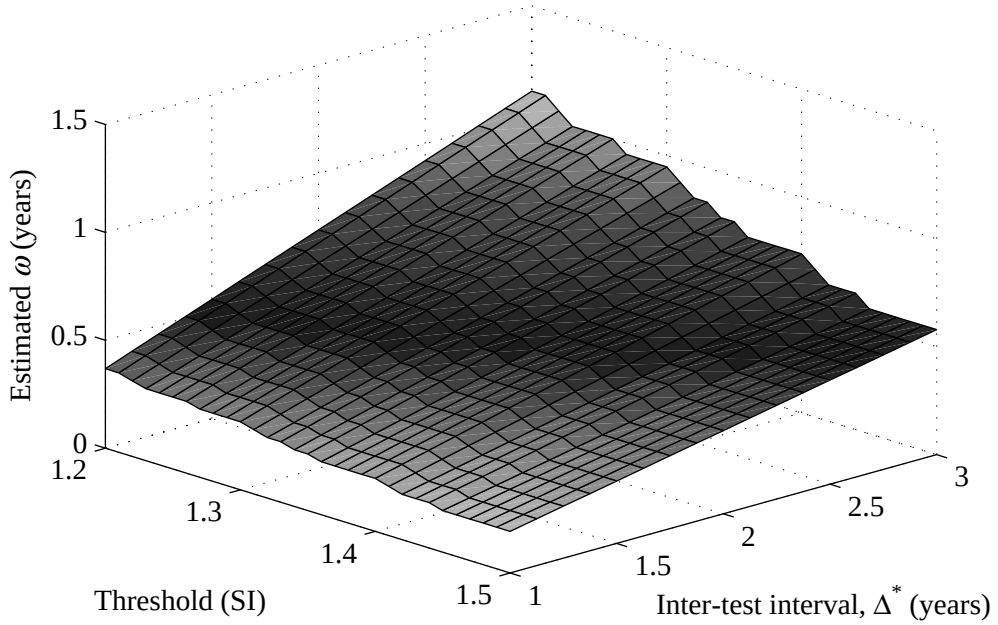


Figure 3.15: Estimated mean duration of recent infection of a test using the 'Stimulation Index' biomarker, as a function of the 'recent'/'non-recent' threshold and assumed inter-test interval

The point estimate for the MDRI, ω (days), is shown for a recent infection test that is based on the 'Stimulation Index' (SI), as a function of both test threshold (above which an SI value indicates 'recent' infection) and the assumed inter-test interval (time between last HIV-negative and first HIV-positive tests, in years). An input FRR of zero was used, and antibody levels were measured using an HIV diagnostic test by Beijing Wantai Biological Pharmacy.

This analysis suggests that efforts should be made to capture a larger set of specimens, appropriate for test characterisation and including relevant background information, to further investigate the potential of using SMARTube^{TM} to construct a biomarker for recent infection. More generally, the investigation strengthens the case for broadening the spectrum of biomarkers conventionally considered.

3.3.2 Local Characterisation of the BED Assay

Potential regional variation in the behaviour of tests for recent infection brings into question the ability to recycle estimates of the MDRI across different incidence studies. In this analysis, existing specimens, captured as part of an ongoing population-based demographic surveillance study in rural South Africa, were used to estimate the MDRI of the BED assay [92, 107], developed in the USA, for a South African context.

Ongoing HIV surveillance is conducted by the *Africa Centre for Health and Population Studies*, *University of KwaZulu-Natal* [172]. The surveillance area is located near the rural market town of Mtubatuba in KwaZulu-Natal. HIV incidence in the area has remained at a high rate of approximately 3.4 infections per 100 person years since 2003 [173], and individuals in the area are eligible for HIV testing as part of the routine study surveillance. For this analysis, the BED assay was applied to specimens (stored as dried blood spots) drawn at times of first HIV-positive tests, for women aged 15-49 and men aged 15-54 who were tested between June 2003 and June 2006. Intervals between last HIV-negative and first HIV-positive tests were large, ranging from 0.5 years to 3 years, with a median interval of 1.3 years. The utility of such data had been previously overlooked, due to the infrequent observation of subjects.

The MDRI was estimated using the approach described in Sections 3.1 and 3.2 above. A local FRR for the BED assay had previously been measured in the same population, and was estimated to be 1.69% (95% CI: 1.00%-2.66%) for a maximum time in the 'recent' state of $W = 306$ days for test 'progressors' [131]. This estimated FRR and the corresponding W were used as inputs in the MDRI estimation. The maximum likelihood estimates of the MDRI, by gender and age, are provided in Table 3.3.

	Number of subjects	**Estimated ω in days (95% CI)**
All data	274	115 (90-139)
Male	72	98 (50-146)
Female	202	120 (91-149)
Ages 15 - 24	138	103 (70-136)
Ages 25 +	136	128 (91-165)

Table 3.3: Estimated mean duration of recent infection for the BED assay in a South African surveillance population, by gender and age
Estimates of the MDRI, ω (days), for the BED assay in a surveillance population in KwaZulu Natal, South Africa, are tabulated. Estimation used an input FRR of 1.96% and $W = 306$ days. 95% confidence intervals (CIs) are also provided (based on large sample maximum likelihood estimator properties). The MDRI was calculated using all data, stratifying data by gender, and stratifying data by age, in turn.

The likelihood functions for the MDRI are shown in Figures 3.16 to 3.18 (using $W = 306$ days throughout). The impact of varying assumptions for the FRR is considered in Figure 3.16, where each of a number of FRR inputs were used in turn: the local estimate of 1.69% described above; an external estimate of 5.6%, which was measured in populations in North America and the Netherlands [16]; and zero. The estimated MDRIs varied by 21 days. In Figure 3.17 and Figure 3.18, the likelihood functions for the MDRI are shown when stratifying data by gender and age respectively (using an input FRR of 1.69%). Age and gender differences were not significant (using a likelihood ratio test and significance level of 5%).

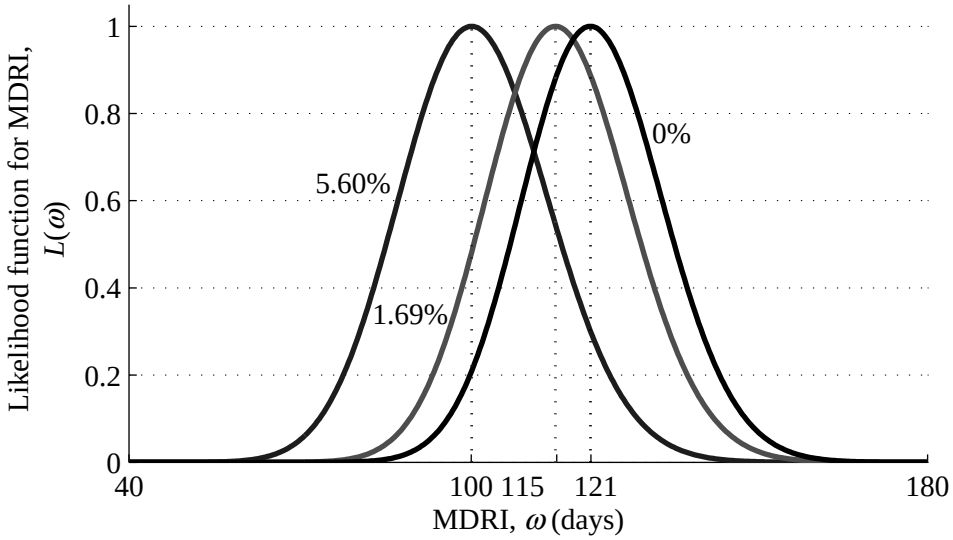


Figure 3.16: Likelihood function for the mean duration of recent infection of the BED assay by input false-recent rate

The likelihood of observing the data (scaled to have a maximum of 1) is shown as a function of the MDRI, ω (days), for the BED assay in the studied South African surveillance population. Data with inter-test intervals larger than $W = 306$ days were included in the analysis, and various FRR inputs were used in turn: the local measurement of 1.69%, an external measurement of 5.60%, and zero.

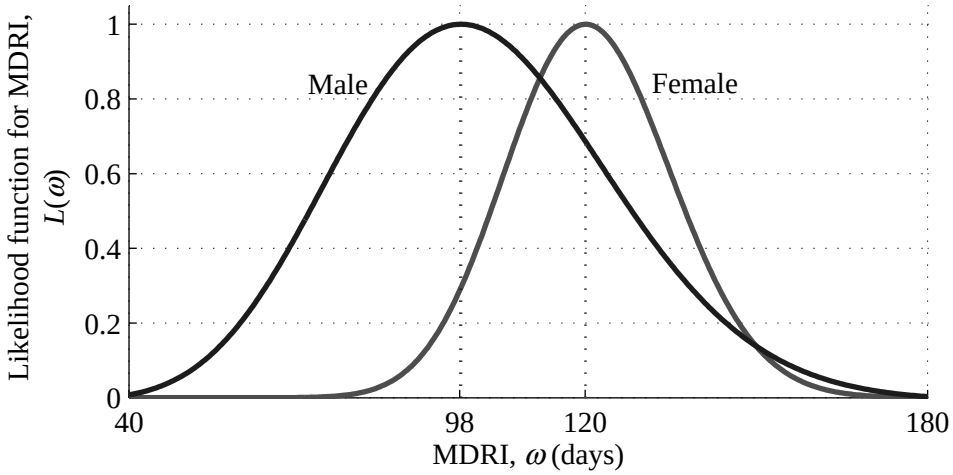


Figure 3.17: Likelihood function for the mean duration of recent infection of the BED assay by gender

The likelihood of observing the data (scaled to have a maximum of 1) is shown as a function of the MDRI, ω (days), for the BED assay in the studied South African surveillance population. Data with inter-test intervals larger than $W = 306$ days were included in the analysis, and an input FRR of 1.69% was used. Data were stratified by gender (Male and Female).

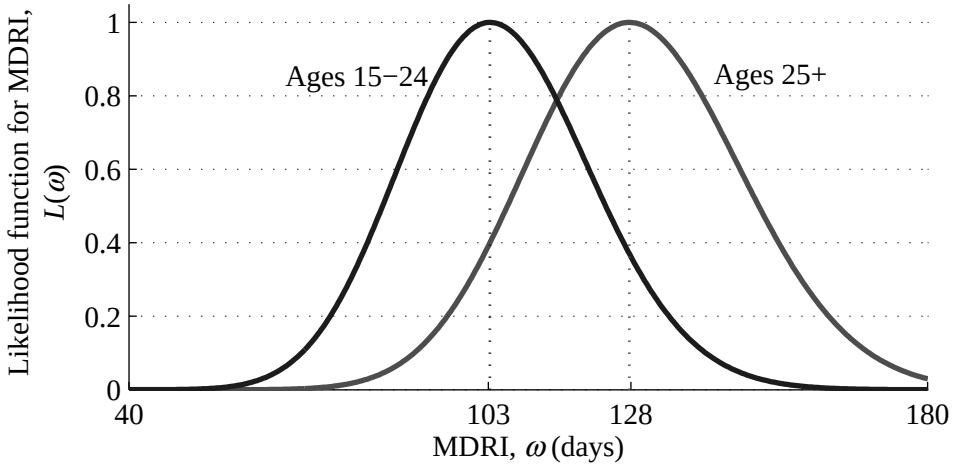


Figure 3.18: Likelihood function for the mean duration of recent infection of the BED assay by age

The likelihood of observing the data (scaled to have a maximum of 1) is shown as a function of the MDRI, ω (days), for the BED assay in the studied South African surveillance population. Data with inter-test intervals larger than $W = 306$ days were included in the analysis, and an input FRR of 1.69% was used. Data were stratified by age (Ages 15-24 and Ages 25+).

This analysis provides the first MDRI estimate for the BED assay using specimens collected as part of population-based HIV surveillance in Southern Africa. These preliminary results suggest that, on average, infected individuals remain in the BED assay-defined state of 'recent' infection for 115 days (95% CI: 90-139 days) in the South African surveillance population. Previously published estimates range from 133 or 153 days in a North American cohort [16] to 187 days in a cohort of Zimbabwean women [17]. Use of these externally obtained MDRI estimates would decrease incidence estimates by between 14% and 39% respectively. Just as other work has cautioned against the generalisation of FRR estimates [131], this work cautions against the use of an MDRI that is not validated in the survey population.

3.4 Estimation within the General Incidence Inference Framework

Throughout this chapter thus far, test characteristics have been defined based on the analyses of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24]. The definitions of the MDRI and FRR emerging from their work rely on specific assumptions about the dynamics of tests for recent infection, known to be violated in reality. These assumptions have since been relaxed and a general incidence inference framework has been developed [29].[21] Estimation of the MDRI that is defined within this general framework is explored below.

The probability that a seroconverter, who tests seropositive at time Δ after testing seronegative, is 'recently' infected at the time of the first seropositive test is

$$p(\Delta) = p\left(\Delta|\underline{\phi}\right) = \frac{1}{\Delta}\int_0^{\Delta} S_R\left(t|\underline{\phi}\right)\,\mathrm{d}t, \tag{3.25}$$

where it is assumed that seroconversion is equally likely to have occurred at any time between the two tests, and $S_R(t|\underline{\phi})$, which is the probability of testing 'recent' at time t after seroconversion, depends on the parameter $\underline{\phi}$. The likelihood of observing the test classifications in a sample of n seroconverters is then

$$L(\underline{\varphi}) = \prod_{i=1}^{n}\left(p(\Delta_i|\,\underline{\phi})\right)^{x_i}\left(1 - p(\Delta_i|\,\underline{\phi})\right)^{1-x_i}, \tag{3.26}$$

[21]The derivation of the general framework for incidence inference is presented in Chapter 2.

where Δ_i and x_i capture the observed data ($i = 1,2,\dots,n$), and $\underline{\phi}$ is the (unknown) model parameter. More specifically, for the i^{th} seroconverter, Δ_i is the inter-test interval and x_i is the observed classification at the first seropositive test, where x_i equals 1 if the subject is 'recently' infected and equals 0 if the subject is 'non-recently' infected.

Any definition of a mean duration in the 'recent' state would involve some averaging of $S_R(t|\underline{\phi})$ over time t, and therefore the MDRI is generally a function of $\underline{\phi}$. Estimation of the MDRI then naturally proceeds by specifying a functional form for $S_R(t|\underline{\phi})$, estimating $\underline{\phi}$ by maximising the likelihood function, and calculating the implied MDRI estimate. However, poor assumptions about $S_R(t|\underline{\phi})$ could lead to substantial bias in results, and therefore a large part of this chapter has been dedicated to finding special cases of this estimation procedure which require less extensive or no parametric assumptions. More specifically, cases have been sought where $p(\Delta)$, and therefore the likelihood function, depend only on the test characteristics of interest, and, if required, a small number of 'nuisance' parameters (that restrict the form of $S_R(t|\underline{\phi})$ less than a full specification of the function in terms of $\underline{\phi}$). In all cases considered, large sample maximum likelihood theory can be used to obtain properties of the MDRI estimator.

Under the assumptions of McDougal et al [16], McWalter and Welte [25] and Wang and Lagakos [24], $S_R(t|\underline{\phi})$ can be expressed as

$$S_R(t|\underline{\phi}) = \varepsilon + (1-\varepsilon)S_{R'}(t|\underline{\phi}^*), \tag{3.27}$$

where $S_{R'}(t|\underline{\phi}^*) = 0$ for sufficiently large t ($t > W$), and the parameter ε is the FRR. The MDRI is defined as $\omega = \int_0^\infty S_{R'}(t|\underline{\phi}^*)\,\mathrm{d}t$ (assuming guaranteed survival until W after infection). For a subject with a sufficiently large inter-test interval Δ ($\Delta > W$), the probability of testing 'recently' infected becomes

$$p(\Delta) = p(\Delta|\varepsilon,\omega) = \varepsilon + (1-\varepsilon)\frac{\omega}{\Delta}. \tag{3.28}$$

Therefore, by including only subjects with large inter-test intervals in the analysis dataset ($\Delta_i > W, i = 1,2,\dots,n$), the likelihood becomes a function of only the test characteristics ε and ω, which can be estimated directly by maximising the likelihood function. While the test characteristics can be jointly estimated in principle, results from the sections above show that estimation of the MDRI, ω, is challenging without an input value of ε for a carefully chosen value of W.

The more general analysis of Kassanjee et al [29] relaxes all assumptions about test dynamics by introducing a post-infection time cut-off T in the construction of the incidence estimator. The MDRI, Ω_T, is now the average time that an individual is 'recently' infected and alive, while infected for less than T. The FRR, β_T, is the probability that an individual who is infected for longer than T will return a 'recent' result. The probability of testing 'recently' infected, $p(\Delta)$, can be expressed in a form that contains Ω_T directly (assuming guaranteed survival until T after infection):

$$p(\Delta) = \frac{1}{\Delta}\left(\int_0^T S_R\left(t|\underline{\phi}\right) \mathrm{d}t + \int_T^\Delta S_R\left(t|\underline{\phi}\right) \mathrm{d}t\right)$$

$$= \frac{\Omega_T}{\Delta} + \frac{1}{\Delta}\int_T^\Delta S_R\left(t|\underline{\phi}\right) \mathrm{d}t. \tag{3.29}$$

When $\Delta = T$, this probability reduces to

$$p(\Delta = T) = p(\Delta = T|\Omega_T) = \frac{\Omega_T}{T}. \tag{3.30}$$

Therefore, by designing a study (or constructing a dataset) where all inter-test intervals are equal to T ($\Delta_i = T,\ i = 1,2,\ldots,n$), the likelihood function is expressed in terms of only the MDRI, Ω_T, which can then be estimated directly by maximising the likelihood of the data. In particular, the likelihood function, which now captures a binomial process, will be maximised when the observed probability of a 'recent' classification at the first seropositive test is used as an estimate of the true probability that is given in Equation (3.30). That is, the estimated MDRI is

$$\widehat{\Omega}_T = T\frac{1}{n}\sum_{\mathrm{i}=1}^{n} x_i, \tag{3.31}$$

where $\sum_{\mathrm{i}=1}^{n} x_i$ is the total number of 'recent' results among the n available results. Importantly, no external estimates of any parameters are required to guide data inclusion rules or assess the likelihood function.

The HIV surveillance data on the BED assay, described in Section 3.3.2, was re-analysed to demonstrate estimation of the MDRI, Ω_T. A time cut-off of $T = 548$ days (1.5 years) was chosen for the exploratory analysis, which aimed to provide preliminary estimates of the MDRI using two simplifications of Equation (3.29).

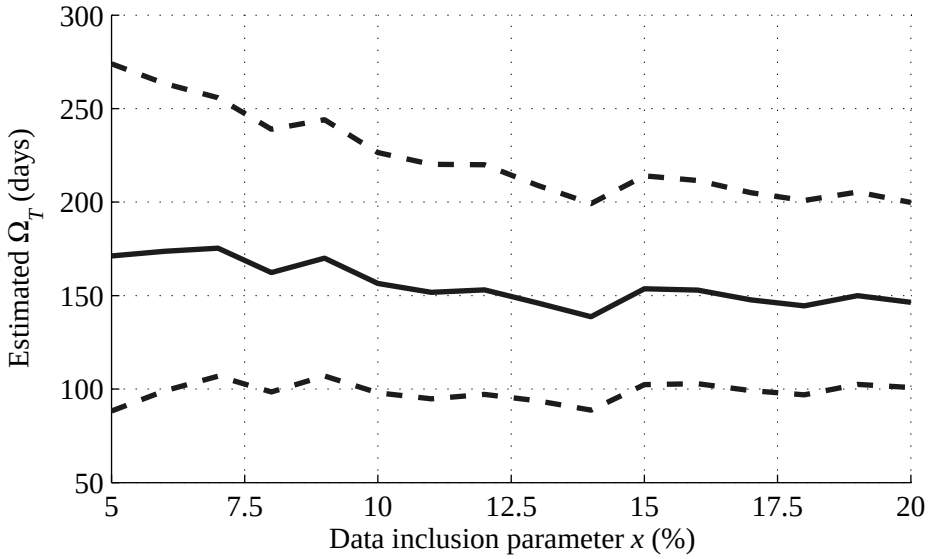


Figure 3.19: Estimated mean duration of recent infection for the BED assay using subjects with inter-test intervals close to T
The estimate of the MDRI, Ω_T (days), for the BED assay in a South African surveillance population is shown as a function of x (%), where subjects with inter-test intervals within $[T \cdot (1-x), T \cdot (1+x)]$ are included in the analysis and $T = 548$ days. In the likelihood function approximation, all inter-test intervals are treated as being equal to T. 95% confidence interval limits are also indicated (dashed lines).

Firstly, all subjects with inter-test intervals *close* to T were included in the analysis, and the simplification given by Equation (3.30) – which requires inter-test intervals to be exactly equal to T – was nevertheless used. A subject was included in the analysis if his inter-test interval, Δ, was within $[T \cdot (1-x), T \cdot (1+x)]$ for a chosen value of x. The estimated MDRI, as a function of x, is shown in Figure 3.19. Point estimates appear relatively stable, and range from 138 days to 175 days for x between 5% and 20%**.** As the value of x increases from zero, bias is introduced as the likelihood function is no longer exact. However, the number of subjects included in the analysis increases and therefore the confidence interval width decreases ($n = 32$, for $x = 5\%$, increases to $n = 101$, for $x = 20\%$).

Secondly, all subjects with inter-test intervals in some range $[L, U]$, which contains T, were included in the analysis, and it was assumed that $S_R(t|\underline{\phi})$ remains constant at some

value θ for all t such that $t \in [L, U]$.[22] The probability of a seroconverter, with inter-test interval Δ, being 'recently' infected is then

$$p(\Delta) = p(\Delta|\Omega_T, \theta) = \frac{\Omega_T + \theta \cdot (\Delta - T)}{\Delta}. \quad (3.32)$$

The likelihood function now contains Ω_T and θ, where the value of θ was specified as an input in the estimation. Using $L = T$ and $U = 3$ years ($n = 115$), the estimated MDRI, as a function of θ, is shown in Figure 3.20. The point estimate drops from 150 days for $\theta = 0\%$ to 130 days for $\theta = 15\%$. The sensitivity of results to the choice of L, U and θ was explored by repeating the estimation for all combinations of values of L between 300 days and 550 days inclusive (in 25 day increments), U between 600 days and 1100 days (25 day increments), and θ between 0% and 15% (1% increments). A histogram of all resulting MDRI point estimates is provided in Figure 3.21, where the minimum, 25th percentile, median, 75th percentile and maximum point estimates are 113, 131, 137, 146 and 170 days, respectively (the sample size ranged from $n = 37$ to $n = 275$).

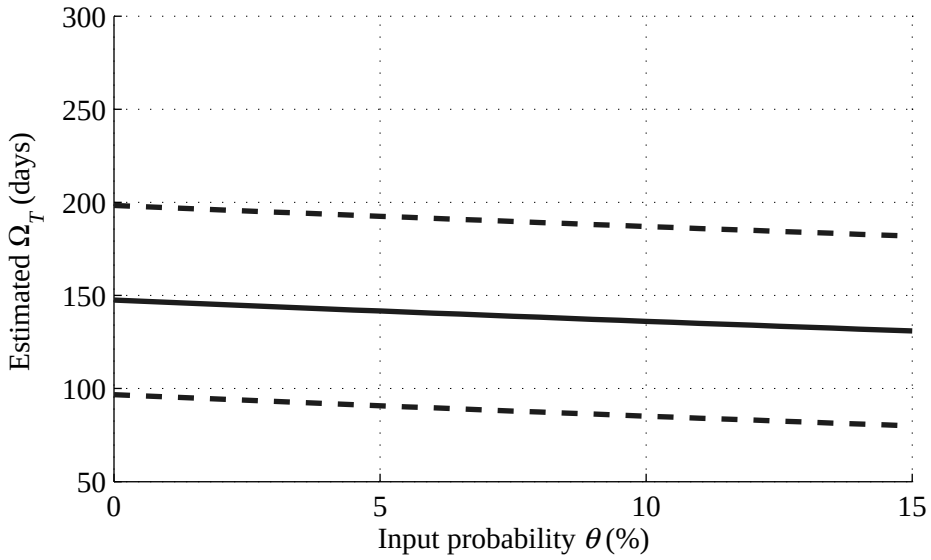


Figure 3.20: Estimated mean duration of recent infection for the BED assay using subjects with inter-test intervals larger than *T*

The estimate of the MDRI, Ω_T (days), for the BED assay in a South African surveillance population is shown as a function of θ (%), where the probability of testing 'recent' is assumed to be constant at θ for times since infection in $[L, U]$. Subjects with inter-test intervals within $[L, U]$ are included in the analysis, and $L = T = 548$ days and $U = 3$ years. 95% confidence intervals are also indicated (dashed lines).

[22]This simplification was also used in the analysis of simulated data in Section 2.1 and 2.2.

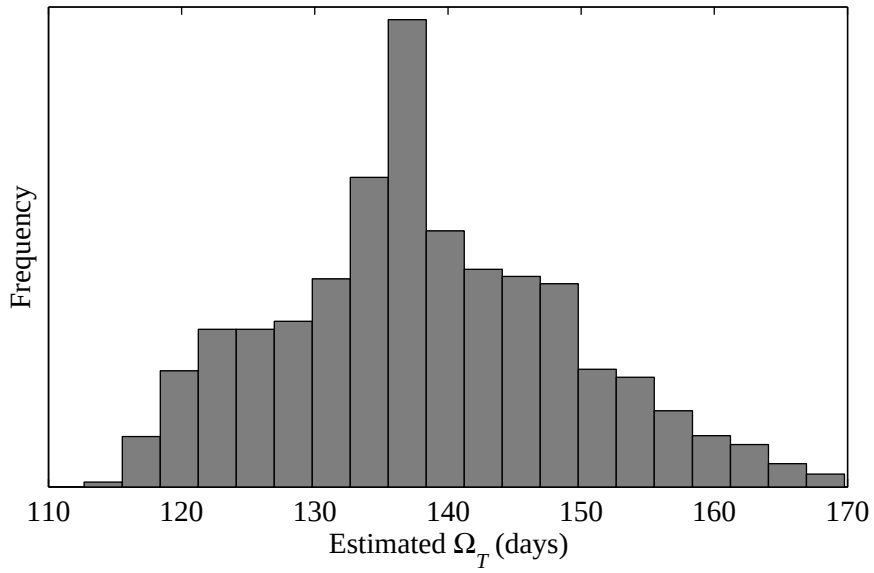


Figure 3.21: Distribution of mean duration of recent infection estimates for the BED assay when varying estimation input parameters
A histogram of estimates for the MDRI, Ω_T (days), for the BED assay in a South African surveillance population is shown. Subjects with inter-test intervals within $[L, U]$ are included in the analysis, and the probability of testing 'recent' is assumed to be constant at θ for times since infection in $[L, U]$. The MDRI was estimated for combinations of L in [300,550] days, U in [600,1100] days and θ in [0%,15%].

While the focus of this chapter is MDRI estimation using specimens from subjects who are observed only once after seroconversion, the approach presented could be generalised to contexts where subjects contribute multiple specimens after seroconversion.[23] Estimation of the MDRI without making parametric assumptions about $S_R\left(t|\underline{\phi}\right)$ is of particular interest. A generalisation of the non-parametric estimation approach described above (captured by Equations (3.30) and (3.31)), to contexts where inter-test intervals can be smaller than T and subjects are followed up over time, is therefore briefly outlined below.

[23]The estimation of the MDRI from longitudinal data, using alternative estimation methods, is discussed in Chapter 4. The generalisation of the approach for MDRI estimation that is presented in this chapter (using single specimens from infected subjects), to the context where there are follow-up specimens, is presented here for completeness.

The inter-test interval Δ_d should be a divisor of T – that is, there must exist some natural number n_d such that $T/\Delta_d = n_d$. For example, for $T = 1$ year, some possible values for Δ_d are 2 months ($n_d = 6$), 4 months ($n_d = 3$) and 1 year ($n_d = 1$). Subjects should have n_d specimens drawn, Δ_d apart, from (and inclusive of) the time of the first seropositive test (although subjects can miss draws and some subjects can be lost to follow-up). Assuming seroconversion events are uniformly distributed between last seronegative and first seropositive tests, the probability of a 'recent' classification at the first seropositive test is

$$p(\Delta_d) = p\left(\Delta_d|\underline{\phi}\right) = \frac{1}{\Delta_d}\int_0^{\Delta_d} S_R\left(t|\underline{\phi}\right)\,\mathrm{d}t, \tag{3.33}$$

as before (see Equation (3.25)). More generally, the probability of a 'recent' classification at draw k ($k = 1,2,\ldots,n_d$), which occurs at time $\Delta_d \cdot (k\text{-}1)$ when using the first seropositive visit as reference time 0, is

$$p^*(k,\Delta_d) = p^*\left(k,\Delta_d|\underline{\phi}\right) = \frac{1}{\Delta_d}\int_{\Delta_d\cdot(k-1)}^{\Delta_d\cdot k} S_R\left(t|\underline{\phi}\right)\,\mathrm{d}t. \tag{3.34}$$

The MDRI can be written in terms of these probabilities:

$$\begin{aligned}\Omega_T &= \int_0^T S_R\left(t|\underline{\phi}\right)\,\mathrm{d}t \\ &= \sum_{k=1}^{n_d}\int_{\Delta_d\cdot(k-1)}^{\Delta_d\cdot k} S_R(t)\,\mathrm{d}t \\ &= \Delta_d\sum_{k=1}^{n_d} p^*(k,\Delta_d),\end{aligned} \tag{3.35}$$

and estimated using the observed probabilities of 'recent' results by

$$\widehat{\Omega}_T = \Delta_d\sum_{k=1}^{n_d}\frac{y_k}{m_k}, \tag{3.36}$$

where m_k results are observed for visit k, and y_k of those are 'recent'. Note that a subject's classifications (for various visits) would not be independent, and the estimated uncertainty for the MDRI estimator should consistently account for this – for example, by using the observed correlations between results as proxies for true correlations in an analytical expression for $var(\widehat{\Omega}_T)$, or through bootstrap resampling at a subject level [174].

In summary, the approach for obtaining preliminary MDRI estimates, using specimens from subjects observed only once after infection, is still as applicable and valuable when defining test characteristics using the general incidence inference framework. By assuming a parametric form for the probability of testing 'recent' as a function of time since infection, all inter-test intervals can be accommodated. However, there are special cases where, by restricting the inter-test intervals included in the analysis, the extent of any parametric assumptions (and potential bias arising from these) can be reduced. When all inter-test intervals are exactly equal to the time cut-off T (or some divisor of T when infected subjects are in fact followed-up), the MDRI can be estimated non-parametrically and directly, without any 'nuisance' parameters or the need for any external knowledge of test dynamics.

Chapter 4

Estimating the Mean Duration of Recent Infection II: Longitudinal Follow-Up of Infected Subjects

Traditionally, estimation of the mean duration of recent infection (MDRI) has relied on longitudinal data, which provides results for the test for recent infection at multiple time points after infection for each subject. Various approaches have been used in the literature to analyse such data, leading to questions about best practices and the implications of methodological differences. This chapter therefore systematically investigates various MDRI estimation methods, and presents some ideas for potentially reducing artefacts in analyses.

A detailed benchmarking of estimation methods was performed using a data simulation platform, and is presented in Section 4.1. The investigation forms part of a project involving an international group of researchers and analysts brought together by the *HIV Modelling Consortium* [53]. The analysis presented here forms a substantial portion of the full body of work by the group, on which a manuscript is currently being prepared.[24]

[24]The analysis presented in Section 4.1 contributes to work being published by the following authors, listed alphabetically and publishing on behalf of an *HIV Modelling Consortium* working group: Daniela De Angelis (Medical Research Council, and Cambridge University – United Kingdom), Marian Farah (Medical Research Council, and Cambridge University – United Kingdom), Debra Hanson (Centers for Disease Control and Prevention – USA), Reshma Kassanjee (SACEMA – South Africa), Phillip Labuschagne (SACEMA – South Africa), Oliver Laeyendecker (National Institute of Allergy and Infectious Diseases – USA), Stephán Le Vu (French Public Health Institute – France), Brian Tom (Medical Research Council, and Cambridge University – United Kingdom), Rui Wang (Harvard University – USA), Alex Welte (SACEMA – South Africa) and Ping Yan (Public Health Agency of Canada – Canada).

The unknown infection times of subjects in the sample pose a particularly subtle obstacle to accurate MDRI estimation, and potential approaches for limiting bias are explored in the remainder of the chapter. Firstly, by redefining the effective 'HIV-positive' state through the introduction of an artificially high 'diagnostic detectability' threshold on a semi-quantitative assay, the time of entry into this state could potentially be more accurately estimated by reducing extrapolation required in analyses given typically available data. The concept is demonstrated in Section 4.2 using simulated data from the benchmarking exercise, and was first presented in a conference poster [34] using actual data.[25] Secondly, when estimating a subject's infection time, diagnostic testing history data should be carefully analysed, and the context-dependent definition of (perforce, *detectable*) infection considered. These topics, inspired by the CEPHIA data preparation process, are briefly explored in Section 4.3. A summary of the ideas presented has been included in a manuscript by CEPHIA.[26]

[25]The concept presented in Section 4.2 was first published in the following conference poster: 'Kassanjee R, Hargrove J, Marinda E, Humphrey J, McWalter TA and Welte A. New criteria for defining biomarker-derived 'recent HIV infection' for the purposes of incidence estimation. E-poster CDC0474 at the *XVIII International AIDS Conference*, 18 – 23 July 2010, Austria'.

[26]Some of the ideas presented in Section 4.3, which were developed while preparing and analysing CEPHIA data, are summarised in a (currently unpublished) manuscript by CEPHIA, authored by: Christopher D Pilcher (University of San Francisco, California – USA), Sheila M Keating (Blood Systems Research Institute – USA), Reshma Kassanjee (SACEMA – South Africa), Elaine McKinney (Public Health England – United Kingdom), Shelley N Facente (University of San Francisco, California – USA), Kara Marson (University of San Francisco, California – USA), Alex Welte (SACEMA – South Africa), Michael P Busch (Blood Systems Research Institute – USA) and Gary Murphy (Public Health England – United Kingdom).

4.1 Benchmarking Estimation Approaches Using Simulated Data

4.1.1 Introduction

Conventionally, estimation of the MDRI has relied on longitudinal data, which captures the dynamics of the test for recent infection as a function of time after infection. Constructing such datasets requires specimens from (initially HIV-negative) subjects to be collected regularly over time before and after infection. Such specimens are costly and logistically difficult to obtain,[27] and, when available, a number of approaches for analysing the resulting longitudinal data could be considered. While different methodologies have been adopted by various research groups [16, 17, 91-103], robust (and widely-accepted) methods for estimating the MDRI are essential for the success of a cross-sectional incidence surveillance approach as unbiased incidence estimation requires unbiased test characterisation. It is important to identify factors that significantly influence test dynamics, and disentangling the variation in MDRI estimates caused by study population differences (such as the subtype of HIV infections) from that caused by analytical differences requires careful consideration of the estimation methods employed. Also, to optimally design studies for characterising tests for recent infection, the relationships between features of the data, such as sample sizes and visit gaps, and the performance of MDRI estimation methods need to be understood.

Consequently, in 2012, the *HIV Modelling Consortium* tasked SACEMA with coordinating a collaborative project to investigate the performance of MDRI estimation approaches using simulated data [53]. The project team consists of researchers from SACEMA, the *Centers for Disease Control and Prevention* (CDC), the *National Institute of Allergy and Infectious Diseases* (NIAID), the *British Medical Research Council* (MRC), *Cambridge University*, the *French Public Health Institute*, *Harvard School of Public Health* and the *Public Health Agency of Canada*. The working group gathered at

[27] In an attempt to address the bottleneck created by the reliance on longitudinal data, the preliminary estimation of the MDRI, using specimens from subjects observed only once after infection, is discussed in Chapter 3.

Harvard University in Boston in July 2012 to discuss the theoretical framework for estimating incidence (and thus defining the MDRI), outline the project scope and methodology, and interpret preliminary outputs. Thereafter, a teleconference was held every second week, so that each step of the process could be critically reviewed and tasks allocated to team members. Online structures were developed to efficiently share relevant documents, datasets, MDRI estimation outputs and drafts of the manuscript (to be published).

In principle, the 'recent' or 'non-recent' classification produced by a test for recent infection may rely on multiple measured biomarkers and clinical indicators [10]. This investigation was restricted to considering a test for recent infection based on a single biomarker. The biomarker need not represent the measurement of a single quantity, but may be a single summary metric of multiple measurements. In line with currently used 'incidence assays', a measurement below a chosen threshold, Y, was interpreted as indicating 'recent' infection. The MDRI summarises the average time that biomarker measurements are below this threshold – more specifically, the MDRI, Ω_T, of relevance for incidence estimation, is the average time 'recently' infected and alive while infected for less than some chosen time cut-off T [29].

A defining feature of this project is the use of simulated data: by simulating data, not only is the true underlying MDRI computable (against which MDRI estimates can be compared), but experiments can be replicated thousands of times and therefore the behaviour of estimation methods fully understood. The SACEMA team developed a comprehensive simulation platform that automated the generation and storage of datasets, application of estimation methods, and storage of outputs in a database.

A large number of MDRI estimation methods were implemented, including both approaches previously used and some new approaches proposed by the project team members. The accuracy and precision of methods were assessed in a number of modelled *scenarios* that capture essential features of what could be encountered in reality, such as different underlying biomarker dynamics (for example, forms of the biomarker signals post infection and levels of measurement noise) and different study designs or subject behaviours (for example, sample sizes, intended visit schedules and tendency of subjects to miss visits).

4.1.2 Design

In practice, an analyst would typically be provided with a single dataset for estimation of the MDRI. However, to understand the performance of any given MDRI estimation method, this 'experiment' needs to be repeated a large number of times and the collection of MDRI estimates considered. The simulation platform[28] that was developed as a part of this project allowed for efficient replication of experiments through the automated generation of datasets and application of estimation methods. Detailed discussions follow about the (i) generation of data, (ii) MDRI estimation methods, and (iii) metrics of performance used.

Broadly outlining the approach taken, the performance of MDRI estimation methods was assessed in a *'base case' scenario* and then in each of a number of *comparison scenarios*. The base case scenario captures a somewhat optimistic study design and high adherence by subjects, and a biomarker dynamic inspired by experiences of researchers in the group. Comparison scenarios were generated by systematically varying aspects of this base case scenario, and capture features of processes that could be encountered in practice. For each scenario, 1 000 datasets were generated and each method of MDRI estimation was applied to each of the datasets (other than for the non-linear mixed models, which were each applied to a common subset of 250 datasets due to the computational expense of the estimation approach). Summary measures of accuracy and precision were used to assess the performance of the MDRI estimation methods, relative to one another and by context.

[28] The simulation platform was implemented and administered by Phillip Labuschagne of SACEMA. The platform automated the generation of data, the application of MDRI estimation methods (which were provided by others in the group, including myself, in the form of Matlab, R or SAS scripts) and the storage of all results. The system was developed using Python, R and MySQL. Due to the large run times involved, computing resources provided by *Amazon Web Services* were utilised.

Data generation

The underlying processes that produce a real-world dataset for MDRI estimation can be considered in two parts. Firstly, the study design, adherence to the protocol and subject behaviour would drive the visit times and infection times of subjects in the sample. Secondly, particulars of the biological signal (as a function of time since infection) and the noise around the signal (from fluctuations within the host or simply imperfect measurement in the laboratory) would drive the observed biomarker readings for the visits (according to when the subjects got infected). Based on imitating these real-world processes, stochastic models were constructed for generating the visit times, infection times and biomarker readings of subjects. The models, inspired by experiences of researchers in the project team, were designed to be simple and rely on relatively few input parameters, yet provide sufficient flexibility to explore the features to be investigated. The base case scenario was defined by choosing particular values for the data generation parameters, based on an idealistic adherence to a somewhat optimistic study design and insights into existing biomarkers for recent infection. Data generation parameters were then systematically varied to create the comparison scenarios.

The model used to generate visit times was intended to portray a study where initially HIV-negative subjects are visited regularly over time, and the *intended* visit gaps and total follow-up times are specified, but the *realised* visits that are generated also account for variability, missed visits and loss to follow-up. More specifically, the starting point is to specify, for all subjects in the study, the intended time between visits while HIV-negative, the intended visit gap once observed to be HIV-positive, and the intended total follow-up time from a subject's first HIV-positive test. To then account for loss to follow-up (for example, due to participants moving out of the study area or termination of the study period), a proportion of subjects are lost earlier than intended, and allocated reduced total follow-up times. To account for variability in visit gaps, the potential visit gaps (in the absence of missed visits) fluctuate around the intended or mean visit gaps. To account for missed visits, there is a chance that a subject misses any potential visit. Since one may expect some subjects to miss visits more frequently than others, subject-specific probabilities of missing visits deviate from some population average probability. A total sample size (or number of subjects who become HIV-positive) is selected, which, in reality, results from factors such as the scale of study, observed HIV incidence, and recruitment and retention rates.

The timing of infection within the 'infection interval', used to refer to the time interval from the last HIV-negative visit to the first HIV-positive visit, will depend on both the study design and subject behaviour. For example, if visit times are specified by the study administrators and strictly maintained, it is valid to assume that a subject is equally likely to have been infected (as defined by the HIV diagnostic test used) at any time in the infection interval. However, subjects may exhibit test-seeking behaviours (advancing visits after exposures specifically to obtain HIV tests) or test-deferring behaviours (delaying visits after exposures, possibly out of fear or from illness with acute HIV) [155, 156, 175, 176]. This complexity was summarised by specifying a distribution for generating infection times within infection intervals.

The model for generating a biomarker reading for each visit aimed to mimic a process whereby realised biomarker readings consist of both signal and noise, and the biomarker evolves differently in different subjects. For example, the signal may grow quickly and saturate at a high value in one subject, and grow slowly and settle at a low level in another. Based on plausible behaviour of a viral or host-response biomarker in a subject, a flexible sigmoid curve was chosen to capture the biomarker signal as a function of time after infection, thus allowing for a period of little growth, clear evolution, and then levelling-off of the signal. To allow for subject-specific evolutions, the three parameters (asymptote, scale and shape) defining the signal were independently generated for each subject, with each subject's parameters drawn from some common multivariate population-level distribution. A flexible noise structure allowed the magnitude of the noise to depend on the value of the signal. While this biomarker model was used to generate readings for almost all scenarios considered (including the base case), an alternative model was implemented for exploring the sensitivities of MDRI estimation results to the true form of the underlying biomarker. In this alternative biomarker model, the biomarker signal was based on a power of time, with a time lag between infection and when the signal was observed

The described models for the simulation of data are formally outlined below, in terms of input parameters and the specific functional forms and statistical distributions used to generate the data. Values for the input parameters for the base case scenario are also specified. The models produce sets of visit times $\{t_{i,j}\}$, infection times $\{t_{inf,i}\}$, and biomarker readings $\{x_{i,j}\}$, where $t_{i,j}$ is the time of the j^{th} HIV-positive visit $(j = 1,2, \dots n_i)$ for subject i $(i = 1,2, \dots, n)$ and $x_{i,j}$ is the corresponding biomarker

reading, and $t_{inf,i}$ is the time of infection for subject i. For each subject, the time of the last HIV-negative visit provides the reference time of 0.

The generation of visit times, $\{t_{i,j}\}$, depends on eleven parameters which are introduced in order below, namely n, F_{max}, p_{ltfu}, μ_{vgn}, σ_{vgn}, α_{vmpn}, β_{vmpn}, μ_{vgp}, σ_{vgp}, α_{vmpp}, and β_{vmpp}.

For any given subject i ($i = 1,2, \dots, n$, where n is specified as an input), the total follow-up time from the first HIV-positive visit, u_i, is drawn from the following mixed distribution:

$$f(x) = \begin{cases} 1 - p_{ltfu} & \text{if } x = F_{max} \\ \dfrac{p_{ltfu}}{F_{max}} & \text{if } x \in (0, F_{max}) \\ 0 & \text{elsewhere} \end{cases}, \tag{4.1}$$

where F_{max} is the intended and maximum follow-up time and p_{ltfu} is the proportion of subjects lost to follow-up. The distribution treats follow-up times as uniformly distributed. in $(0,F_{max})$ for those subjects who are lost to follow-up.

The time of the first HIV-positive visit for subject i, relative to the last HIV-negative visit, depends on the study protocol and behaviour of the subject while HIV-negative. The first HIV-positive visit time, $t_{i,1}$, is drawn from a normal distribution with mean $\mu_{vgn} \cdot (n_{vmn,i} + 1)$ and variance $\sigma_{vgn}^2 \cdot (n_{vmn,i} + 1)$ (truncated to exclude values below zero), where μ_{vgn} is the mean visit gap while HIV-negative and σ_{vgn} captures the variability in the gap, and $n_{vmn,i}$ is the number of visits that are missed (before the first HIV-positive visit). The count $n_{vmn,i}$ is drawn from a geometric($p_{vmn,i}$) distribution ($n_{vmn,i} = 0,1,2, \dots$), where $p_{vmn,i}$ is the subject's probability of independently missing any visit. This probability, $p_{vmn,i}$, is drawn from a population-level beta($\alpha_{vmpn}, \beta_{vmpn}$) distribution.

Denoting the realised visit gap after the k^{th} HIV-positive visit by $\delta_{i,k}$ ($k = 1,2, \dots$), the time of the subsequent HIV-positive visit, $t_{i,k+1}$, is generated using $t_{i,k+1} = t_{i,k} + \delta_{i,k}$, where $\delta_{i,k}$ depends on the study protocol and subject behaviour while HIV-positive. The visit gap, $\delta_{i,k}$, is drawn from a normal distribution with mean $\mu_{vgp} \cdot (n_{vmp,i,k} + 1)$ and variance $\sigma_{vgp}^2 \cdot (n_{vmp,i,k} + 1)$ (truncated to exclude values below zero), where μ_{vgp} is the mean visit gap while HIV-positive and σ_{vgp} captures variability in the gap, and $n_{vmp,i,k}$ is the number of visits that are missed (after the k^{th} HIV-positive visit). The number of

visits missed, $n_{vmp,i,k}$, is drawn from a geometric($p_{vmp,i}$) distribution ($n_{vmp,i,k} = 0,1,2, \ldots$), where $p_{vmp,i}$ is the subject's probability of independently missing any visit. This probability, $p_{vmp,i}$, is drawn from a population-level beta($\alpha_{vmpp}, \beta_{vmpp}$) distribution.

Only visits occurring within the subject's total follow-up time are retained – that is, n_i is the largest value such that $t_{i,n_i} - t_{i,1} \leq u_i$.

The generation of infection times, $\{t_{inf,i}\}$, depends on two parameters, namely α_{inf} and β_{inf}. The infection time for the i^{th} subject, $t_{inf,i}$, is $t_{i,1} \cdot u_i$ where u_i is drawn from a beta($\alpha_{inf}, \beta_{inf}$) distribution. Setting $\alpha_{inf} = \beta_{inf} = 1$ recovers uniformly distributed infection times, and $\alpha_{inf} = \beta_{inf}$, $\alpha_{inf} > \beta_{inf}$ and $\alpha_{inf} < \beta_{inf}$ produce *average* test-neutral, test-seeking and test-deferring behaviours respectively.

The generation of biomarker readings, $\{x_{i,j}\}$, depends on the thirteen parameters contained in the inputs $\underline{\mu}_b$, Σ_b, and $\underline{e}$. For the i^{th} subject, the signal at time t post infection is given by:

$$y_i(t) = y(t|\underline{b}_i) = b_{i,1} \cdot \left(1 + \left(\frac{t}{b_{i,2}}\right)^{-b_{i,3}}\right)^{-1}, \tag{4.2}$$

where $b_{i,1}$ captures the maximum height or asymptote, $b_{i,2}$ captures a horizontal scaling, and $b_{i,3}$ describes the shape, of the signal which is anchored at zero at infection. The subject's signal parameters, contained in $\underline{b}_i = [b_{i,1}, b_{i,2}, b_{i,3}]$, are drawn from a population-level multivariate $N(\underline{\mu}_b, \Sigma_b)$ distribution (truncated to have only a non-negative support for each parameter). Allowing for noise, a biomarker measurement, $x_{i,j}$, for the visit at time $t_{i,j}$, is drawn from a

$$N\left(y_i(t^*_{i,j}), \left(e_0 + e_1 \cdot y_i(t^*_{i,j}) + e_2 \cdot \left(y_i(t^*_{i,j})\right)^{e_3}\right)^2\right)$$

distribution (and non-positive values are censored to be 0), where the noise parameters are contained in $\underline{e} = [e_0, e_1, e_2, e_3]$ and $t^*_{i,j} = t_{i,j} - t_{inf,i}$ is the time since infection at the visit.

For the alternative biomarker model, the data depends on the five inputs $b_{1,l}$, $b_{1,u}$, $b_{2,l}$, $b_{2,u}$ and e. The signal for subject i at time t post infection is:

$$y_i(t) = y(t|\underline{b}_i) = 8 \cdot \max(0, \mathrm{t} - \mathrm{b_{i,1}})^{b_{i,2}}, \tag{4.3}$$

where $b_{i,1} > 0$ captures the time between infection and the appearance of signal, and $b_{i,2} > 0$ describes the shape of signal growth. The subject's lag parameter, $b_{i,1}$, is drawn from a uniform$(b_{1,l}, b_{1,u})$ distribution, and the shape parameter is drawn from a uniform$(b_{2,l}, b_{2,u})$ distribution. Including noise, a biomarker measurement, $x_{i,j}$, for the visit at time $t_{i,j}$, is drawn from a $N\left(y_i\left(t_{i,j}^*\right), e^2\right)$ distribution (and non-positive values are censored to be 0), where $t_{i,j}^* = t_{i,j} - t_{inf,i}$ is the time since infection at the visit.

For each dataset generated, the above mechanisms produce $\{t_{i,j}\}$, $\{t_{inf,i}\}$ and $\{x_{i,j}\}$ for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots n_i$. The infection times, $\{t_{inf,i}\}$, would be unknown in practice, and only the observable visit times and biomarker readings, $\{(t_{i,j}, x_{i,j})\}$, are passed to the MDRI estimation methods.

The base case scenario is described below, and values of the data generation input parameters are provided. For each of the 50 subjects captured in the sample ($n = 50$), the intended visit gaps were 3 months and 1 month while HIV-negative and HIV-positive respectively ($\mu_{vgn} = 3$ months and $\mu_{vgp} = 1$ month), and intended follow-up time was 2 years ($F_{max} = 2$ years). There was a 10% coefficient of variation for visit gaps ($\sigma_{vgn} = 3$ days and $\sigma_{vgp} = 9$ days), and no missed visits and no loss to follow-up ($\alpha_{vmpn} \to 0$, $\beta_{vmpn} = 1$, $\alpha_{vmpp} \to 0$, $\beta_{vmpp} = 1$, $p_{ltfu} = 0$). Visit times were considered to be fixed by study design and independent of infection, and therefore infection times were uniformly distributed in infection intervals ($\alpha_{inf} = 1$ and $\beta_{inf} = 1$). In terms of the biomarker dynamics, for visit times measured in days, the mean signal parameter values were given by $\underline{\mu_b} = [85,190,5]$ (capturing height, scale and shape, in order). The covariance matrix for subject-specific deviations, Σ_b, was such that the standard deviations of the height, scale and shape parameters were 7.5, 50 and 1.4 respectively; and there was a correlation of -0.4 between the shape and scale parameters, 0.3 between the shape and height parameters, and -0.12 between the scale and height parameters (arising from the relationship of each parameter with shape, without any additional or partial correlation). The noise inputs were $\underline{e} = [2,0,0.3,0.5]$, producing a standard deviation that began at 2 at infection and levelled off at 4.7 (based on the average asymptote of 85). One or more parameters were varied to produce each of the comparison scenarios, and details are specified in the tables of results.

Approaches for estimating the mean duration of recent infection

Based on the general incidence inference framework [29], the MDRI can be expressed mathematically as $\Omega_T = \int_0^T P_R(t)\,\mathrm{d}t$, where $P_R(t)$ is the probability of being 'recently' infected and alive at time t after infection.[29] Estimation of the MDRI therefore inevitably entails using the longitudinal data to make inferences about the function $P_R(t)$. This can be achieved either directly by fitting a chosen model for $P_R(t)$ to the 'recent' and 'non-recent' classifications of subjects, or indirectly by modelling the biomarker measurements and then subsequently computing the implied probability of obtaining a measurement below the threshold Y (that is, obtaining a 'recent' result). Throughout this work, $T = 1$ year and $Y = 40$, and negligible mortality within T after infection was assumed in the estimation of the MDRI.

Various approaches for estimating the MDRI appear in the literature [16, 17, 91-103]. For this benchmarking exercise, a number of methods were implemented, intended to be representative of those previously published and to provide some extensions. The methods are represented in the mind map in Figure 4.1, and were broadly divided into three categories, capturing (i) interpolation methods, (ii) survival analysis, and (iii) parametric regression. Those methods that were implemented by others in the project team appear in grey text (Methods 18-24) and are not discussed in this thesis (additional interpretation of all results will be provided in the manuscript to be published by the group). Specific implementations within each category of methods are described in more detail below.

[29] The derivation of the general incidence inference framework and the definitions of test characteristics appearing in the incidence estimator are presented in Chapter 2.

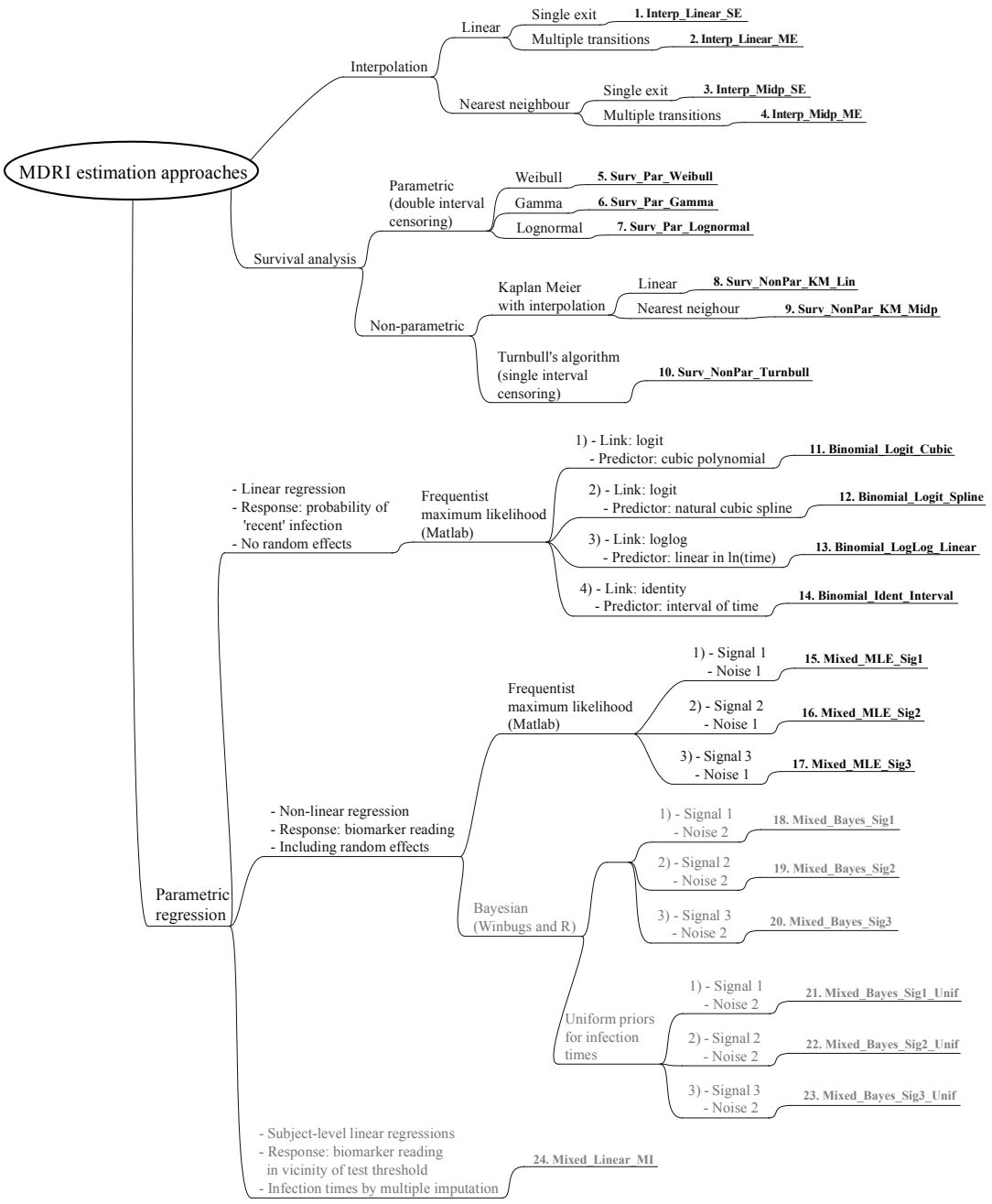


Figure 4.1. Mind map of the mean duration of recent infection estimation approaches included in the benchmarking exercise

The 24 MDRI estimation methods that were implemented and evaluated are captured in the mind map (each numbered and labelled in bold). Methods shown in grey text (Methods 18-24) were implemented by other members of the team, and are therefore not discussed in this thesis. When parametrically modelling biomarker measurements, the three forms for the biomarker signal that were used (Signal 1-3) are given by 1) $a(1 + (t \cdot \exp(b))^c)^{-1}$, 2) $a(1 - \exp(-(t \cdot \exp(b))^c))$, and 3) $a\big(1 - \exp\big(-(t \cdot \exp(b))\big)\big)$ where t is time since infection and a, b and c are model parameters; and the two noise structures that were used (Noise 1-2) specify that the standard deviation of noise is 1) a linear function of the signal, or 2) constant.

Generally, a challenge faced when estimating the MDRI from longitudinal data is the uncertainty of subjects' infection times. It is only known that infection (as defined by the HIV diagnostic test being used) occurs somewhere between the last HIV-negative and first HIV-positive visits, and explicitly allowing for this uncertainty substantially increases the complexity of estimation methods. One simple approach is to approximate each subject's infection time by some expected infection time, but even this relies on assumptions about subject behaviour. In studies where visits times are fixed by design, it is valid to assume that the infection time is uniformly distributed in the infection interval. In this case, an infection time could be estimated by the midpoint of the infection interval. A number of the MDRI methods outlined below use this 'midpoint infection time' approach.

Interpolation methods (Methods 1-4) use mathematical interpolation for each subject to obtain a biomarker reading at every time point after infection (using a 'midpoint infection time', assuming a biomarker reading of zero at infection, and without any extrapolation beyond the subject's last data point). At every time t post infection, the proportion of those subjects with available biomarker readings (namely those that have not been lost to follow-up) that have measurements below the test threshold Y provides an estimate of $P_R(t)$. Biomarker readings were interpolated either linearly or by a nearest neighbour approach (implying that transitions between states occur at the midpoints of periods between visits). Furthermore, the approach was implemented either using all data as is or assuming a subject's readings remain above the threshold, Y, after a first measurement above Y. The first implementation allows for *multiple transitions* between the 'recent' and 'non-recent' states, while the second assumes and enforces *single exits* from the state of 'recent' infection (disregarding information contained in later data points), as done in some of the literature [91, 93-96, 98-100, 102]. This class of methods provides a basic, informal approach for analysing the longitudinal data. The Matlab function that was developed to implement Methods 1 and 2 is provided in Appendix B.3.

Conventional statistical **survival analysis** techniques are used to model the time from entering to exiting a state of interest. For this application, even if the biomarker signal is monotonically increasing over time after infection (at least until the time cut-off T), measurement noise implies that subjects may fluctuate in and out of the state of 'recent' infection many times, and therefore this single sojourn view of 'recent' infection is too restrictive. These statistical methods have nevertheless been employed in this area [93, 94, 96, 98-100, 102] and can be easily implemented using standard software, and were

therefore included in this benchmarking exercise. To utilise the survival analysis framework, all data points beyond a subject's first 'non-recent' data point were effectively discarded.

Survival analysis methods are also more amenable to accommodating (at least some cases of) data censoring. In this application, data is double interval censored as both entry and exit times are interval censored. More specifically, a subject's entry time lies within the infection interval, and exit time lies in what is referred to here as the 'exit interval'. The exit interval is taken to be the time interval either between the subject's first 'non-recent' visit and the preceding visit, or between the latest visit and infinity (or some very large time) if there is no 'non-recent' result (typically referred to as a right censored exit time).

The first set of survival analysis methods (Methods 5-7) are based on fitting a parametric distribution (Weibull, Gamma or Lognormal distribution) to the time in the 'recent' state, treating the time in the state as double interval censored (that is, the infection time is uniformly distributed in the infection interval, and then the exit time is uniformly distributed in the exit interval). A maximum likelihood approach was used to estimate the distribution's parameters, and the integration required to assess the likelihood function was performed numerically (using the composite trapezoidal rule).

Parametric assumptions about the distribution of times in the 'recent' state were avoided by using a Kaplan-Meier or Product-Limit estimator [177, 178] of the survival function (Methods 8 and 9), where the estimated survival function is a step function that maximises the likelihood of the observed data. The standard Kaplan-Meier approach accommodates only right censored data, and therefore 'midpoint infection times' were used and exit times were estimated by interpolation between biomarker readings on either side of the exit interval (assuming a zero biomarker measurement at infection), either linearly or by a nearest neighbour approach. If the last visit of the subject was 'non-recent', the exit time was treated as right censored. The MDRI estimates obtained from these methods will be equal to those produced by the corresponding single-exit interpolation methods (Methods 1 and 3) when there are no right censored exit times within T after infection.

Founded on the principle of an Expectation-Maximisation algorithm, Turnbull's algorithm [178, 179] extends the Kaplan-Meier estimator of the survival function to allow for interval censored data (Method 10). The double interval censoring in the data implies that a trapezoidal distribution should be used to describe the uncertainty in the time in the

'recent' state, while the single interval censoring accommodated by Turnbull's algorithm utilises a uniform distribution. Therefore, while not formally exact, to facilitate application of this approach and reproduce previous implementations [98], the data was interpreted as capturing single interval censored times in the 'recent' state, with bounds given by the minimum and maximum possible times in the 'recent' state implied by the data. In this implementation, the survival function was taken to evolve piecewise linearly (rather than being piecewise constant, as is conventionally assumed).

In **parametric regression** (Methods 11-17), a particular form for the expected response, as a function of predictors, is fitted to the data. In this application, time since infection (or transformations thereof) provide the predictor(s), and either the biomarker reading or test classification ('recent' versus 'non-recent') provides the response. Different classes of parametric regression can be considered, depending on the following, for example: the method of model fitting (such as whether a Bayesian or Frequentist approach is used, and the specific software or algorithms utilised); whether the response is the biomarker reading or test classification; whether the defined model forms are linear or non-linear in the parameters; whether subject-level clustering of data is accounted for by random effects; and how the unknown infection times are accommodated. The choices made greatly impact the statistical complexity, and computational stability and expense, of the approach. Two classes of parametric regression models were employed and are discussed below, namely linear binomial regression models and non-linear (normal-response) mixed models, both using 'midpoint infection times'.[30]

The linear binomial regression models (Methods 11-14) assumed functional forms for $P_R(t)$, with parameters estimated to maximise the likelihood of classifications in the data (using Matlab's 'glmfit' tool). The models were of the form $g\left(P_R(t)\right) = \underline{\beta}^T \underline{x}(t)$, where $g(.)$ is the link function, and $\eta = \underline{\beta}^T \underline{x}(t)$ is the linear predictor containing both the model parameters in $\underline{\beta}$ and the predictors, which are functions of time since infection, in $\underline{x}(t)$. This class of models neglects the subject-level clustering of data points. Four parametric forms of the model were implemented (Forms 1-4), intended to provide varying degrees of flexibility: 1) $g(.)$ is a logit link and η is a cubic polynomial of time (a four-parameter model); 2) $g(.)$ is a logit link and η is a linear combination of the basis

[30]Additional classes of parametric regression models were implemented by other members of the project team, and are to be presented in the group's manuscript.

functions of a natural cubic spline that has six equally spaced knots within $[0, T]$ (a six-parameter model); 3) $g(.)$ is a loglog link and η is a linear function of the natural logarithm of time (a two-parameter model, with $P_R(t)$ resembling a Weibull survival function); and 4) $g(.)$ and η are such that $P_R(t)$ is constant within each of six equally-sized subintervals segmenting the post-infection time interval $[0, T]$ (six parameters). In some scenarios considered, the sparseness of data did not support meaningful use of Forms 2 and 4. For Form 4, when fewer than five subjects contributed data points to any subinterval, the number of subintervals was reduced in decrements of one (to a minimum of one).

Non-linear mixed models (Methods 15-17) for the biomarker measurements are substantially more complex and difficult to implement than any of the approaches discussed above, but are able to capture various features of the data. In general, a parametric form for the biomarker signal over time since infection is chosen, as is a measurement noise structure, and subject-level clustering or correlation of data is accounted for through subject-specific deviations (random effects) of signal parameters from the population-level average signal parameters (fixed effects).

More specifically, the chosen model structure specified that the biomarker measurement for subject i at time t after infection was given by

$$f_i(t) = f(t|\underline{a}_i) = y(t|\underline{a}_i) + \left(k_1 + k_2 y(t|\underline{a}_i)\right) \cdot \varepsilon, \tag{4.4}$$

where the signal was described by $y(t|\underline{a}_i)$ and the noise was captured by ε, which is a standard normal random variable (independently drawn for every biomarker measurement). The subject's n_a signal parameters contained in $\underline{a}_i$ followed a multivariate normal distribution, with mean $\underline{\mu}_a$ (the fixed effects) and covariance matrix Σ_a (capturing the variability of the random effects or subjects' deviations). The model parameters were contained in $\underline{\mu}_a$, Σ_a and $\underline{k} = [k_1, k_2]$.

Three forms for the signal were implemented (Signals 1-3). Signal 1 is

$$y(t|\underline{a}_i) = a_{i,1} \cdot \left(1 + \left(t \times exp\left(a_{i,2}\right)\right)^{a_{i,3}}\right)^{-1}, \tag{4.5}$$

and Signal 2 is

$$y(t|\underline{a}_i) = a_{i,1} \cdot \left(1 - \exp\left(-\left(t \times exp\left(a_{i,2}\right)\right)^{a_{i,3}}\right)\right), \tag{4.6}$$

and both result in eleven-parameter models ($n_a = 3$). Signal 3, a special case of Signal 2, is

$$y(t|\underline{a}_i) = a_{i,1} \cdot \left(1 - \exp\left(-\left(t \times exp\left(a_{i,2}\right)\right)\right)\right) \tag{4.7}$$

and produces a seven-parameter model ($n_a = 2$). Signal 1 matches the biomarker signal used to generate base case scenario data, up to transformations of the parameters.

The mixed model parameters were estimated to maximise the likelihood of the data. The large search space and complex likelihood function can cause instabilities when using optimisation algorithms to search for the parameters' maximum likelihood estimates, and therefore Markov Chain Monte Carlo (MCMC) approaches are often used. For this exercise, stability of processes was important as estimation was performed for several thousands of datasets through an automated system, and therefore the MCMC approach provided by Matlab's 'nlmefitsa' tool was used. Also, while (in practice) convergence criteria would be carefully assessed when analysing any given dataset (as would the choice of parametric forms), here the number and lengths of the chains (which aim to converge to the maximum likelihood estimates) used in a single estimation were increased until a balance was found between feasible run times and suitably small variation in MDRI estimates if analysing the same dataset multiple times (three chains of 1 000 steps were used, each with its own starting value, and the parameter estimates providing the largest likelihood value were considered to be the maximum likelihood estimates).

For the parametric approaches above (Methods 5-7, 11-17), data points occurring at large times after infection, relative to the time cut-off T, were discarded before model fitting, in an attempt to achieve the best model fit specifically to data over post-infection times in $[0, T]$. More specifically, data points were excluded which were beyond T plus some margin (ranging from 0 to $0.5 \cdot T$ across methods). Since the maximum follow-up time for subjects was 2 years and biomarkers were well behaved over that time by design, this data exclusion may have limited impact here. However, in practice, biomarker dynamics may become less predictable years after infection, and therefore care should be taken in choosing data to train models that are intended to describe early test dynamics.

The output of each method of estimation described above can be expressed as an estimated $P_R(t)$ – possibly after some transformation of output model parameters. The MDRI was then estimated by the area under this inferred curve, from $t = 0$ to $t = T$, either analytically or numerically (using the composite trapezoidal rule).

In practice, when any parameter is inferred, the reported estimate has little meaning without an estimate of uncertainty, typically expressed as a confidence interval. However, for this exercise, confidence intervals were not reported. By operating within a simulation environment where experiments could be replicated, the bias and variance of point estimates were directly measured. These metrics inform the coverage and widths of confidence intervals that may be produced. Furthermore, for any given estimation method, it is possible that a number of methods could be used to obtain confidence intervals. The optimisation of the confidence interval approach, and finding a balance between its complexity or computational expense and its performance, would extend this benchmarking exercise well beyond its already broad scope.

Metrics of performance for estimation methods

For each scenario and each MDRI estimation method, a large number of MDRI estimates were obtained by analysing each of the 1 000 datasets that were generated. The distribution of point estimates provides information about the behaviour and performance of the estimation method. Many metrics could be considered for summarising this distribution, and potentially how it relates to the true MDRI – for example, percentiles, variance, the expected value, bias, or the mean squared error could be computed. For each scenario and each estimation method, two such performance statistics are reported here: (i) accuracy is summarised by relative bias (difference between average MDRI estimate and true MDRI, relative to the true MDRI), and (ii) precision is captured by the relative standard deviation or coefficient of variation (standard deviation of estimates relative to the mean estimate).

To quantify the bias in estimation, the average MDRI estimate needs to be compared to the true MDRI, which was therefore computed for each scenario considered. The MDRI, Ω_T, depends on the dynamics of the underlying biomarker as a function of time since infection.

For scenarios where data was generated using the primary, base case biomarker model, the MDRI was calculated using the following expression (substituting in the scenario's data generation parameter values):

$$\Omega_T = \int_{t=0}^{t=T} \int_{\underline{b}\in R^3} f_{\underline{b}}(\underline{b})\, \Phi\left(\frac{Y - y(t|\underline{b})}{e_0 + e_1 \cdot y(t|\underline{b}) + e_2 \cdot \left(y(t|\underline{b})\right)^{e_3}}\right) \mathrm{d}\underline{b}\, \mathrm{d}t, \qquad (4.8)$$

where $Y > 0$ is the test threshold used to distinguish between 'recent' and 'non-recent' results, T is the time cut-off appearing in the definition of the MDRI, $y(t|\underline{b})$ is the signal given by Equation (4.2), $\Phi(.)$ is the standard normal cumulative distribution function, and $f_{\underline{b}}(\underline{b})$ is a multivariate normal probability density function with mean $\underline{\mu}_b$ and covariance matrix Σ_b, truncated to have only a positive support.

For the alternative data generation biomarker form discussed, the MDRI was computed by:

$$\Omega_T = \int_{t=0}^{t=T} \int_{b_1=b_{1,u}}^{b_1=b_{1,l}} \int_{b_2=b_{2,l}}^{b_2=b_{2,u}} f_{b_1}(b_1)\, f_{b_2}(b_2) \Phi\left(\frac{Y - y(t|\underline{b})}{e}\right) \mathrm{d}b_2\, \mathrm{d}b_1\, \mathrm{d}t, \qquad (4.9)$$

where $\underline{b} = [b_1, b_2]$, $y(t|\underline{b})$ is the signal given by Equation (4.3), $f_{b_1}(b_1) = \left(b_{1,u} - b_{1,l}\right)^{-1}$ if $b_{1,l} \le b_1 < b_{l,u}$ (0 elsewhere) and $f_{b_2}(b_2) = \left(b_{2,u} - b_{2,l}\right)^{-1}$ if $b_{2,l} \le b_2 < b_{2,u}$ (0 elsewhere).

4.1.3 Results

The results from each method of MDRI estimation are plotted for the base case scenario in Figure 4.2, and summary performance statistics for all scenarios are provided in Tables 4.1 to 4.4 – namely the relative bias of the estimation procedure (accuracy) and the relative standard deviation or coefficient of variation of point estimates (precision). The contents of each of the tables are outlined below, and then some qualitative insights gained from interpreting the results are highlighted.

The base scenario is somewhat idealistic, capturing monthly visits for 50 HIV-positive subjects over two years after infection and infection times that are uniformly distributed in intervals of 3 months. For this scenario, all categories of estimation methods appear to provide useful estimates (Figure 4.2), but there is a clear sensitivity to parametric assumptions when considering the mixed model results (Methods 15-17). Therefore, an initial investigation into parametric assumptions for the mixed models was performed, and is presented in Table 4.1. This investigation was used to down select and fine tune the (computer-intensive) mixed models that were to be subsequently assessed alongside the remaining estimation methods (and resulted in Signal 3 being abandoned). For each of the three fitted signal forms (Methods 15-17), four variations of the mixed model were implemented for this auxiliary investigation, based on assumptions about the noise structure and correlation between random effects (Table 4.1). The scenarios that were used to assess performance captured alternative forms of the underlying biomarker, and variations of the noise structure and levels of correlation between random effects (compared to the base case biomarker dynamic).

Features of the data related to study protocol and subject behaviour (that is, the generation of visit and infection times) are considered in Table 4.2 and Table 4.3. The performance of estimation methods when varying the number of subjects and mean visit gaps, while HIV-negative and HIV-positive in turn, is summarised in Table 4.2; while the impact of (increasing levels) of missed visits, loss to follow-up and non-uniformity of infection times is explored in Table 4.3. The impact of features of the underlying biomarker dynamic is investigated in Table 4.4, where the levels of measurement noise and inter-subject variability are varied (for the base case biomarker form) and the alternative biomarker form is considered.

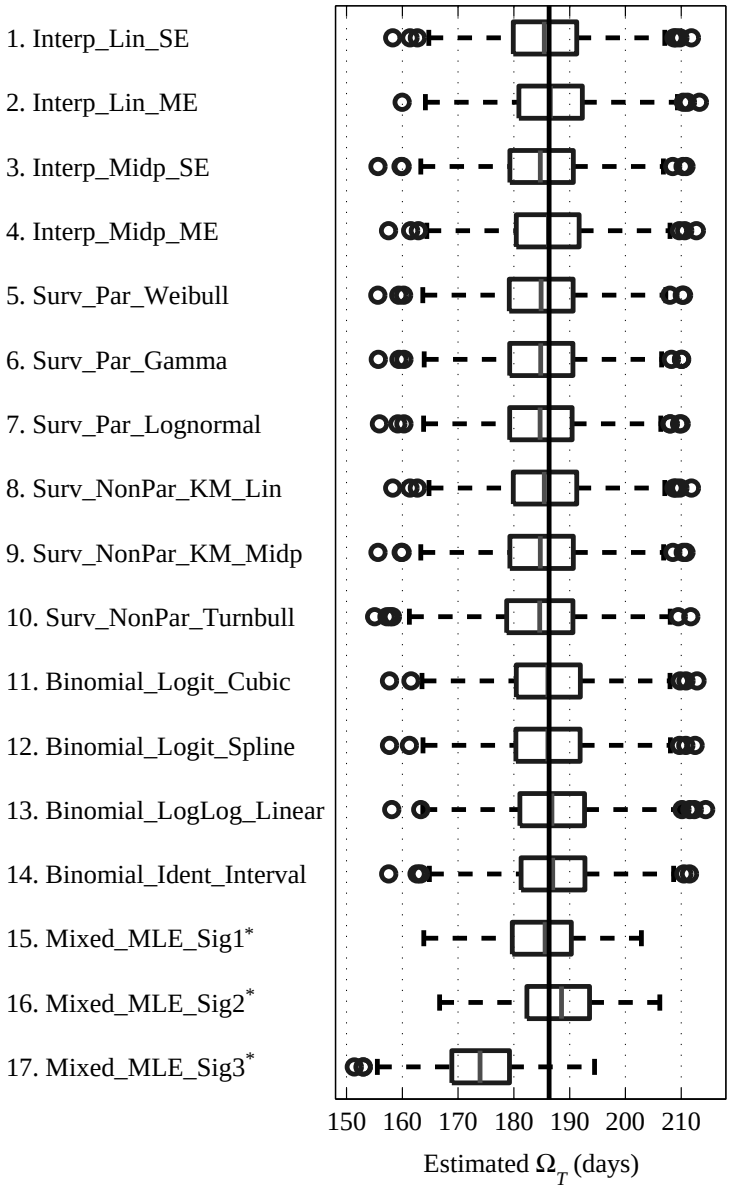


Figure 4.2. **Box-and-whisker plots of the mean duration of recent infection point estimates for the base case scenario by method of estimation**

Box-and-whisker plots of the point estimates for the MDRI, Ω_T (days), are provided for the base case scenario, for each of the MDRI estimation methods ($T = 1$ year). The box and dividing line indicate the central 50% and median of estimates respectively, and whiskers and circles capture remaining data points and outliers respectively (outliers are more than 1.5 times the interquartile range or box length away from the central box). The vertical black line indicates the true MDRI.

* Methods 15-17 appear to have fewer outliers because fewer experiments were replicated (250 instead of 1 000).

	15. Mixed_MLE_Sig1	*16. Mixed_MLE_Sig2*	*17. Mixed_MLE_Sig3*	*15. Mixed_MLE_Sig1*	*16. Mixed_MLE_Sig2*	*17. Mixed_MLE_Sig3*	*15. Mixed_MLE_Sig1*	*16. Mixed_MLE_Sig2*	*17. Mixed_MLE_Sig3*	*15. Mixed_MLE_Sig1*	*16. Mixed_MLE_Sig2*	*17. Mixed_MLE_Sig3*
FITTED MODEL SPECIFICATIONS												
Noise standard deviation	Constant			Linear in signal			Constant			Linear in signal		
Correlation between 'random effects'	None			None			Included			Included		
RELATIVE BIAS (%)												
Underlying biomarker form												
Base case*	-1 0	-0 2	-6 1	-1 1	-0 4	-6 0	-0 6	0 9	-6 7	-0 7	0 9	-6 7
Power function, no time lag[1]	-0 8	-0 1	-2 8	-0 7	-0 1	-2 8	-0 4	0 2	-3 3	-0 3	0 1	-3 4
Power function, with time lag[2]	1 7	2 2	-1 1	1 7	2 2	-1 1	-1 1	0 0	-1 8	-1 1	-0 0	-1 8
Noise structure												
Base case*	-1 0	-0 2	-6 1	-1 1	-0 4	-6 0	-0 6	0 9	-6 7	-0 7	0 9	-6 7
Alternative 1[3]	-0 6	0 3	-7 0	-0 6	0 3	-7 0	-0 4	0 9	-6 8	-0 4	1 0	-6 8
Alternative 2[4]	-1 7	-1 4	-8 7	-1 4	-0 4	-8 7	-1 6	-1 0	-8 1	-1 5	-0 0	-8 2
Correlation between signal 'random effects'												
No correlation	-0 3	0 6	-4 6	-0 4	0 5	-4 7	-0 3	1 3	-5 7	-0 3	1 3	-5 6
Base case*	-1 0	-0 2	-6 1	-1 1	-0 4	-6 0	-0 6	0 9	-6 7	-0 7	0 9	-6 7
Double correlation coefficients	-1 4	-0 6	-6 8	-1 4	-0 8	-6 8	-0 6	1 0	-7 2	-0 7	0 9	-7 2
RELATIVE STANDARD DEVIATION (%)												
Underlying biomarker form												
Base case*	4 0	4 0	4 1	4 1	4 1	4 1	4 1	4 1	4 7	4 8	4 8	4 7
Power function, no time lag[1]	4 3	4 3	4 4	4 3	4 4	4 4	4 4	4 4	3 5	3 5	3 6	3 6
Power function, with time lag[2]	3 8	3 8	3 4	3 3	3 7	3 6	3 3	3 3	3 3	3 3	3 4	3 4
Noise structure												
Base case*	4 0	4 0	4 1	4 1	4 1	4 1	4 1	4 1	4 7	4 8	4 8	4 7
Alternative 1[3]	4 4	4 4	4 5	4 5	4 3	4 4	4 5	4 5	5 4	5 4	5 2	5 1
Alternative 2[4]	4 5	5 1	4 5	5 3	4 4	4 5	4 4	4 5	5 1	5 3	5 0	5 1
Correlation between signal 'random effects'												
No correlation	4 6	4 6	4 5	4 6	4 6	4 6	4 5	4 6	5 1	5 2	5 1	5 0
Base case*	4 0	4 0	4 1	4 1	4 1	4 1	4 1	4 1	4 7	4 8	4 8	4 7
Double correlation coefficients	4 0	3 9	4 0	4 0	3 9	3 8	4 0	4 0	4 9	4 9	4 8	4 7

[1] Alternative biomarker form specified in text, $b_{1,l} = b_{1,u} = 0$, $b_{2,l} = 0.3$, $b_{2,u} = 0.35$ and $e = 4$, where time is measured in days

[2] Alternative biomarker form specified in text, $b_{1,l} = 15$, $b_{1,u} = 25$, $b_{2,l} = 0.3$, $b_{2,u} = 0.35$ and $e = 4$, where time is measured in days

[3] Noise standard deviation increases and then decreases with growing signal: $\underline{e} = [4, -0.26, 2.8, 0.5]$

[4] Noise standard deviation increases and then decreases with growing signal: $\underline{e} = [4, -0.22, 10, 0.25]$

Table 4.1: Performance of variations of mixed models in scenarios constructed for exploring sensitivities to parametric assumptions

The estimated relative bias (%) and relative standard deviation (%) for each MDRI estimation procedure are shown, for variations of the non-linear mixed models for the biomarker readings. Scenarios capture different underlying biomarker forms, and changes in the noise structure and the correlations between signal 'random effects' (relative to the base case). * Scenario corresponds to base case.

	1. Interp_Lin_SE	2. Interp_Lin_ME	3. Interp_Midp_SE	4. Interp_Midp_ME	5. Surv_Par_Weibull	6. Surv_Par_Gamma	7. Surv_Par_Lognormal	8. Surv_NonPar_KM_Lin	9. Surv_NonPar_KM_Midp	10. Surv_NonPar_Turnbull	11. Binomial_Logit_Cubic	12. Binomial_Logit_Spline	13. Binomial_LogLog_Linear	14. Binomial_Ident_Interval	15. Mixed_MLE_Sig1	16. Mixed_MLE_Sig2
RELATIVE BIAS (%)																
Number of subjects																
5[1]	-0 0	0 5	-0 1	0 5	-0 3	0 1	-0 2	0 0	-0 1	-0 0	0 4	-0 9	0 7	0 9	0 3	1 8
10	-0 6	-0 0	-0 8	-0 1	-0 7	-0 7	-0 7	-0 6	-0 8	-0 8	-0 2	-0 2	0 1	0 3	-0 5	1 1
20	-0 3	0 1	-0 5	0 0	-0 4	-0 5	-0 5	-0 3	-0 5	-0 5	0 0	-0 0	0 3	0 5	-0 6	0 9
50*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
100	-0 4	0 1	-0 7	-0 0	-0 6	-0 7	-0 7	-0 4	-0 6	-0 9	-0 0	-0 0	0 3	0 4	-0 3	1 2
150	-0 3	0 2	-0 5	0 1	-0 5	-0 5	-0 6	-0 3	-0 5	-0 8	0 1	0 1	0 4	0 6	-0 2	1 3
Mean HIV-negative visit gap[2]																
1 week	-0 4	0 1	-0 6	0 0	-0 7	-0 6	-0 5	-0 4	-0 6	-0 6	0 1	0 0	0 5	0 5	-0 2	1 5
2 weeks	-0 5	-0 0	-0 7	-0 1	-0 8	-0 7	-0 6	-0 5	-0 7	-0 8	-0 1	-0 1	0 3	1 3	-0 2	1 6
1 month	-0 5	-0 0	-0 6	-0 1	-0 7	-0 6	-0 5	-0 5	-0 6	-0 6	-0 0	-0 0	0 3	0 6	0 1	1 8
3 months*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
6 months	-0 4	0 1	-1 5	-0 9	-0 8	-0 6	-0 6	-0 4	-1 5	1 6	-0 5	-1 0	0 2	-2 4	-0 5	-0 2
Mean HIV-positive visit gap[2]																
1 week	-2 4	0 1	-2 7	0 2	-2 6	-2 7	-2 7	-2 4	-2 6	-2 6	0 2	0 2	0 5	0 0	-0 1	1 2
2 weeks	-1 3	0 1	-1 5	0 1	-1 5	-1 5	-1 5	-1 3	-1 5	-1 5	0 1	0 1	0 4	-0 0	-0 8	0 6
1 month*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
3 months	1 0	1 1	0 1	0 2	0 1	0 0	-0 0	1 0	0 1	0 4	-0 2	1 8	0 4	-1 8	-1 5	0 9
6 months[3]	1 1	1 1	-3 5	-3 5	0 1	0 4	0 2	1 1	-3 5	14 1	-3 2	-24 5	-0 1	-10 3	-8 8	-19 4
RELATIVE STANDARD DEVIATION (%)																
Number of subjects																
5[1]	14 1	14 1	14 4	14 4	14 8	14 1	14 6	14 1	14 4	14 9	14 5	13 9	14 6	14 6	13 8	13 9
10	9 7	9 8	9 8	9 9	9 8	9 8	9 8	9 7	9 8	10 2	10 0	9 9	10 0	9 9	9 0	8 9
20	7 0	7 0	7 1	7 1	7 1	7 1	7 1	7 0	7 1	7 5	7 1	7 2	7 2	7 2	7 3	7 2
50*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
100	2 9	3 0	3 0	3 0	3 0	3 0	3 0	2 9	3 0	3 1	3 0	3 0	3 1	3 0	2 9	2 9
150	2 5	2 5	2 6	2 6	2 6	2 6	2 6	2 5	2 6	2 7	2 6	2 6	2 6	2 6	2 4	2 4
Mean HIV-negative visit gap[2]																
1 week	4 0	4 0	4 1	4 1	4 1	4 1	4 0	4 0	4 1	4 1	4 1	4 1	4 2	4 1	3 9	3 8
2 weeks	4 1	4 1	4 1	4 2	4 1	4 1	4 1	4 1	4 1	4 2	4 2	4 2	4 3	4 1	4 2	4 1
1 month	4 0	4 1	4 1	4 2	4 1	4 1	4 1	4 0	4 1	4 2	4 2	4 2	4 3	4 1	3 9	3 9
3 months*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
6 months	5 3	5 3	5 6	5 6	5 6	5 5	5 4	5 3	5 6	4 9	5 6	6 3	5 4	6 4	5 0	5 1
Mean HIV-positive visit gap[2]																
1 week	4 2	4 3	4 2	4 3	4 2	4 2	4 2	4 2	4 2	4 2	4 3	4 3	4 3	4 3	4 2	4 2
2 weeks	4 2	4 3	4 3	4 3	4 3	4 3	4 3	4 2	4 3	4 4	4 4	4 4	4 4	4 4	4 1	4 1
1 month*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
3 months	4 4	4 4	4 9	4 8	4 9	5 0	4 9	4 4	4 9	6 1	4 9	7 1	5 0	6 1	5 1	4 4
6 months[3]	4 4	4 4	6 1	6 1	6 2	6 9	7 4	4 4	6 1	4 8	22 2	37 7	6 5	8 8	13 6	16 1

[1] Convergence issues arose for Methods 5-7, which were run on only 250 datasets and 80% of runs produced outputs

[2] The standard deviation for visit gaps (σ_{vgn} and σ_{vgp}) was 10% of the mean

[3] Convergence issues were encountered for Methods 15 and 16; and Methods 12 and 14 utilised unrealistic model forms given the sparseness of data

Table 4.2: Performance of the mean duration of recent infection estimation methods in scenarios capturing various study designs

The estimated relative bias (%) and relative standard deviation (%) for each of the MDRI estimation procedures are shown. Scenarios capture varying numbers of (HIV-positive) subjects and average visit gaps, while HIV-negative and HIV-positive.

* Scenario corresponds to base case.

	1. Interp_Lin_SE	2. Interp_Lin_ME	3. Interp_Midp_SE	4. Interp_Midp_ME	5. Surv_Par_Weibull	6. Surv_Par_Gamma	7. Surv_Par_Lognormal	8. Surv_NonPar_KM_Lin	9. Surv_NonPar_KM_Midp	10. Surv_NonPar_Turnbull	11. Binomial_Logit_Cubic	12. Binomial_Logit_Spline	13. Binomial_LogLog_Linear	14. Binomial_Ident_Interval	15. Mixed_MLE_Sig1	16. Mixed_MLE_Sig2
RELATIVE BIAS (%)																
Missed visit probability[1]																
0% and 0%*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
50% and 0%	1 5	1 9	-0 2	0 3	-0 5	-0 5	-0 5	1 5	-0 2	1 1	2 2	2 2	2 5	1 9	2 1	2 6
75% and 0%	11 0	11 3	7 9	8 3	-0 7	-0 6	-0 6	11 0	7 9	3 5	4 8	4 8	4 7	4 2	7 7	7 2
0% and 50%	1 1	1 3	0 6	0 9	-0 2	-0 4	-0 5	1 1	0 6	-0 6	-0 1	-0 1	0 2	0 3	-0 5	1 4
0% and 83%	13 3	13 7	14 2	14 8	0 6	-0 2	-0 6	13 5	14 5	-0 2	0 0	-0 2	0 4	0 6	-1 3	-0 5
50% and 50%	2 9	3 1	1 3	1 5	-0 1	-0 2	-0 3	2 9	1 3	0 7	1 9	1 8	2 4	2 0	1 5	2 6
50% and 83%	12 4	12 8	13 2	13 7	0 3	-0 3	-0 6	12 5	13 4	-1 6	1 4	1 2	2 2	1 8	0 5	2 7
Loss to follow-up[2]																
0% within 2 years*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
100% within 2 years	-2 0	-0 0	-2 2	-0 2	-0 7	-0 7	-0 7	-0 6	-0 8	-1 0	-0 1	-0 2	0 2	0 8	-0 0	1 1
100% within 1 5 years	-2 6	0 0	-2 9	-0 1	-0 7	-0 7	-0 6	-0 5	-0 7	-1 9	-0 1	-0 1	0 3	1 1	-0 0	0 8
100% within 1 year[3]	-4 1	0 1	-4 5	-0 1	-0 6	-0 2	0 0	-0 7	-0 9	-2 9	-0 0	-0 0	0 1	1 8	1 0	1 3
Infection times																
Uniformly distributed*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
Test-seeking behaviour[4]	7 8	8 2	7 6	8 1	7 6	7 5	7 5	7 8	7 6	7 4	8 2	8 1	8 6	8 7	7 9	9 6
RELATIVE STANDARD DEVIATION (%)																
Missed visit probability[1]																
0% and 0%*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
50% and 0%	5 2	5 2	5 5	5 5	5 0	5 1	5 0	5 2	5 5	5 1	5 1	5 1	5 2	5 2	5 0	5 0
75% and 0%	6 0	6 0	6 4	6 4	6 5	6 5	6 5	6 0	6 4	6 6	6 6	6 6	6 7	6 8	6 7	6 6
0% and 50%	4 4	4 4	4 8	4 8	4 9	5 0	4 9	4 4	4 8	5 4	5 1	5 1	5 1	5 0	4 2	4 1
0% and 83%	4 6	4 6	5 2	5 2	6 8	7 0	7 0	4 6	5 2	8 4	7 2	7 3	7 2	7 2	6 6	7 1
50% and 50%	5 4	5 4	5 8	5 8	5 8	5 8	5 8	5 4	5 8	6 3	6 0	6 0	6 0	6 0	5 5	5 4
50% and 83%	5 6	5 6	6 3	6 3	7 6	7 6	7 6	5 6	6 3	9 7	8 0	8 0	7 9	8 0	6 3	5 9
Loss to follow-up[2]																
0% within 2 years*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
100% within 2 years	5 0	5 0	5 1	5 2	5 0	5 0	5 0	4 9	5 0	5 2	5 1	5 1	5 2	5 1	4 9	4 8
100% within 1 5 years	4 9	5 1	5 0	5 2	5 0	4 9	4 9	4 9	4 9	5 1	5 1	5 1	5 2	5 1	4 9	4 8
100% within 1 year[3]	5 4	5 8	5 5	5 9	5 6	5 6	5 7	5 5	5 6	5 7	5 8	5 8	5 8	5 7	5 4	5 5
Infection times																
Uniformly distributed*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
Test-seeking behaviour[4]	4 0	4 0	4 0	4 1	4 1	4 1	4 1	4 0	4 0	4 2	4 1	4 1	4 2	4 1	4 0	3 9

[1] Reported as the mean of the population-level distribution for the probability of missing a visit, which has an absolute standard deviation of 5%, for HIV-negative subjects and HIV-positive subjects in turn

[2] Reported as the percentage of subjects lost within the specified maximum follow-up time

[3] Methods 2 and 4 require that at least one subject is followed-up until $T = 1$ year after infection, and therefore 2% of estimations could not be completed

[4] Infection times occur a third of the infection interval away from the first HIV-positive visit on average, and 80% of infection times occur closer to the first HIV-positive visit than the last HIV-negative visit

Table 4.3: Performance of the mean duration of recent infection estimation methods in scenarios capturing various non-ideal features of subject visits

The estimated relative bias (%) and relative standard deviation (%) for each of the MDRI estimation procedures are shown. Scenarios capture varying levels of missed visits and loss to follow-up, and non-uniformity of infection times.

* Scenario corresponds to base case.

	1. Interp_Lin_SE	2. Interp_Lin_ME	3. Interp_Midp_SE	4. Interp_Midp_ME	5. Surv_Par_Weibull	6. Surv_Par_Gamma	7. Surv_Par_Lognormal	8. Surv_NonPar_KM_Lin	9. Surv_NonPar_KM_Midp	10. Surv_NonPar_Turnbull	11. Binomial_Logit_Cubic	12. Binomial_Logit_Spline	13. Binomial_LogLog_Linear	14. Binomial_Ident_Interval	15. Mixed_MLE_Sig1	16. Mixed_MLE_Sig2
RELATIVE BIAS (%)																
Magnitude of noise[1]																
4	-0 4	0 2	-0 6	0 0	-0 6	-0 6	-0 6	-0 4	-0 6	-0 8	0 1	0 0	0 4	0 5	-0 5	1 1
10	-3 9	-0 2	-4 5	-0 2	-4 5	-4 5	-4 5	-3 9	-4 5	-4 7	-0 1	-0 2	0 6	0 3	-0 3	1 0
20	-15 9	-0 2	-17 3	-0 1	-17 2	-17 3	-17 1	-15 9	-17 2	-17 4	-0 2	-0 3	1 2	-0 1	-3 0	-0 1
Inter-subject variability[2]																
Base case*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
Doubled	-0 8	0 5	-1 2	0 2	-1 1	-1 8	-2 1	-0 8	-1 2	-0 6	-0 1	-0 0	0 4	-0 0	-3 4	-2 5
Tripled	-1 4	-0 0	-1 8	-0 3	-2 5	-3 3	-4 6	-1 4	-1 8	-0 5	-0 8	-0 8	-0 3	-1 0	-6 3	-5 3
Underlying biomarker form																
Base case*	-0 4	0 1	-0 7	-0 1	-0 7	-0 7	-0 7	-0 4	-0 7	-0 9	-0 1	-0 1	0 3	0 4	-0 7	0 9
Power function, no time lag[3]	-10 1	-0 9	-11 6	-0 4	-11 5	-11 4	-11 4	-10 1	-11 5	-11 8	-0 4	-0 4	0 1	-0 0	-0 3	0 1
Power function, with time lag[4]	-8 7	-0 6	-9 9	-0 1	-9 9	-9 9	-9 9	-8 7	-9 9	-10 0	-0 0	-0 1	0 5	0 5	-1 1	-0 0
RELATIVE STANDARD DEVIATION (%)																
Magnitude of noise[1]																
4	4 5	4 5	4 5	4 6	4 5	4 5	4 5	4 5	4 5	4 7	4 6	4 6	4 7	4 6	4 7	4 7
10	4 4	4 5	4 6	4 5	4 6	4 6	4 5	4 4	4 6	4 8	4 5	4 5	4 6	4 6	4 3	4 3
20	5 5	4 5	5 9	4 6	5 9	5 8	5 7	5 5	5 9	6 0	4 6	4 6	4 6	4 7	6 9	4 7
Inter-subject variability[2]																
Base case*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
Doubled	6 9	6 8	7 0	6 9	7 1	7 0	6 9	6 9	7 0	7 0	7 0	7 0	6 9	7 1	6 1	6 1
Tripled	7 6	7 5	7 7	7 6	7 6	7 5	7 2	7 6	7 7	7 6	7 8	7 8	7 6	7 9	7 6	7 6
Underlying biomarker form																
Base case*	4 4	4 4	4 5	4 5	4 5	4 5	4 4	4 4	4 5	4 6	4 5	4 5	4 6	4 5	4 1	4 1
Power function, no time lag[3]	4 8	4 6	5 1	4 8	5 1	5 1	5 1	4 8	5 1	5 3	4 7	4 8	4 8	4 9	4 3	4 4
Power function, with time lag[4]	4 3	4 1	4 6	4 2	4 6	4 6	4 6	4 3	4 6	4 8	4 2	4 2	4 3	4 2	3 3	3 3

[1] Reported as the (constant) standard deviation of noise, with values representing 10%, 25% and 50% of the test threshold (or 5%, 12% and 24% of the average range of the signal)

[2] Described in terms of the standard deviations of signal 'random effects', maintaining their correlations

[3] Alternative biomarker form specified in text, $b_{1,l} = b_{1,u} = 0$, $b_{2,l} = 0.3$, $b_{2,u} = 0.35$ and $e = 4$, where time is measured in days

[4] Alternative biomarker form specified in text, $b_{1,l} = 15$, $b_{1,u} = 25$, $b_{2,l} = 0.3$, $b_{2,u} = 0.35$ and $e = 4$, where time is measured in days

Table 4.4: Performance of the mean duration of recent infection estimation methods in scenarios capturing various features of the underling biomarker dynamics

The estimated relative bias (%) and relative standard deviation (%) for each of the MDRI estimation procedures are shown. Scenarios capture varying magnitudes of measurement noise, levels of inter-subject variability, and underlying biomarker forms.

* Scenario corresponds to base case.

Single sojourn assumptions. A primary cause of bias in MDRI estimation is the assumption of a single sojourn in the 'recent' infection state, as contained in all single-exit and survival analysis methods (Methods 1 and 3, 5-10). The underestimation of the MDRI became more pronounced as measurement noise was increased and HIV-positive visit gaps were reduced – both of these features made it more likely that an (increasingly) 'early' upward fluctuation above the test threshold would be observed, artificially clipping a subject's sojourn in the 'recent' state. In principle, the level of bias (for a given visit schedule) would depend on both the magnitude (and structure) of noise and the growth in the signal (such as captured by its slope), *in a vicinity* of the threshold (namely, over the time when noise may cause biomarker readings to fluctuate across the threshold). For example, when data was generated using the alternative biomarker form, the relatively small gradient of the signal at the test threshold resulted in greater underestimation of the MDRI by the single sojourn approaches, even though the magnitude of measurement noise around the threshold remained similar to that for the base case scenario. Given the limitations of single sojourn approaches, the interpretations below mainly focus on the results from the remaining estimation approaches.

Variability of estimates. The various methods of estimation exhibited similar variability (relative standard deviations) in a given context (neglecting the most extreme contexts). Results indicate that the standard deviation of estimates varies approximately inversely proportionally to the square root of the number of subjects, and proportionally to the standard deviation of the distribution of individual durations in the 'recent' state (see investigation into inter-subject variability). Larger visit gaps, more frequently missed visits, increased loss to follow-up, and greater measurement noise all added to the variability in more nuanced ways.

Loss to follow-up. Increased loss to follow-up would be expected to increase the variability of estimates, as the biomarker behaviour at later times after infection is inferred from fewer subjects. In the extreme case that no subjects are followed until T after (an estimated) infection, only approaches that extrapolate beyond the latest data points can be used. When there is drop-out, those subjects that are observed to transition out of the 'recent' state would over-represent the shorter sojourns, and therefore biases arise when using methods that naïvely average over data (see Methods 1 and 3).

Parametric assumptions for biomarker mixed models. When considering the parametric assumptions used in the mixed models, Signal 3 produced large biases and

was removed early on in the benchmarking exercise (Table 4.1). This bias was anticipated as Signal 3 is concave downwards and does not allow for a period of little growth in the biomarker signal immediately after infection. Models using Signals 1 and 2 (Methods 15 and 16), allowing for varying noise magnitude and non-zero correlations between random effects, were retained for the remainder of the investigation. Signal 1 exactly matched the data generation process (up to transformations of parameters), although this ideal alignment of assumptions with reality would not occur in practice. When analysing any given dataset, knowledge of the underlying process, plots of the data, and statistical model fitting diagnostic tools would be used to select a reasonable parametric form. A comparison of the results for Signal 1 and 2 suggests that, while the true underlying form would never be known, similar inferences can be made even when fitting the 'wrong' parametric model by choosing a form is that reasonably aligned with the data and sufficiently flexible. The model fit (and therefore biases from incorrect parametric assumptions) depends on factors such as the frequency of visits and magnitude of noise.

Noise structure. Many statistical models assume noise has either a constant standard deviation (additive noise) or a constant coefficient of variation (proportional noise). When incorrectly assuming the latter noise structure (or even when correctly assuming it, while having an incorrect assumption for the biomarker signal), performance of the estimation approach would be poor – models are forced to describe the data points with near-zero signal values very well, at the expense of fitting the rest of the data very poorly. A constant coefficient of variation noise structure was therefore not implemented, after some preliminary explorations (not shown).

Binomial regression models. The binomial regression models appear to be particularly stable across scenarios, with performance typically at least on par with that of the mixed models. While the binomial models were computationally stable and easy to implement, they do not account for all the data features that are captured by the mixed models – notably, the subject-level clustering of data points – and also assume that data is missing completely at random. Results appear to be less sensitive to parametric assumptions, with all four models (ranging from two-parameter to six-parameter models) providing similar results (excluding scenarios where data were insufficient to fit certain parametric forms). Intuitively, many different biomarker dynamics could potentially be summarised into fewer (suitably flexible) forms for $P_R(t)$ – a topic that could now be further investigated using this simulation platform.

Large visit gaps. All methods of estimation were inaccurate when infection intervals were very large (compared to the duration of 'recent' infection). The number of (HIV-positive) data points captured in any dataset could be reduced by either a higher probability of missed visits, or larger intended visit gaps. Methods employing subject-specific biomarker interpolation were vulnerable to high missed visit probabilities (some subjects then had very large realised visit gaps and their times in the 'recent' state were particularly poorly estimated). On the other hand, population-level model fitting became challenging and inaccurate with large scheduled visit gaps for all subjects (as there was then no data to describe the dynamic over certain intervals of post-infection time after anchoring infection times).

Infection times. Unknown infection times pose a fundamental obstacle to MDRI estimation as substantial bias can be introduced when the assumptions about infection times are violated. All methods of estimation introduced a bias (in absolute terms) equal to the difference between the mean infection time in reality and the mean infection time under model assumptions (namely uniformly distributed infection times within infection intervals).

Correlation between estimates. Lastly, the MDRI estimates produced by any two estimation procedures in a chosen scenario were highly correlated (Pearson correlation coefficients averaged around 0.95 after excluding single-exit estimation methods and contexts with extremely sparse data). This suggests that (the unbiased) estimation procedures should give similar results when applied to any given dataset (keeping in mind the similarities in precision), and limited benefit can be gained from averaging the results of multiple methods.

4.1.4 Discussion

For tests for recent infection to be of utility for cross-sectional incidence surveillance, tests must first be characterised – that is, test properties of relevance for incidence estimation must be measured. Through the development of a theoretical framework for incidence inference [29], two generally-defined test properties have emerged: the mean duration of recent infection (MDRI), which is the average time 'recently' infected and alive while infected for less than some time cut-off T, and the false-recent rate (FRR),

which is the probability that a person who is infected for a time larger than T will produce a 'recent' result.[31] While the FRR in principle captures a mixture of test dynamics and epidemiological and demographic history (and is therefore expected to be context-dependent), the MDRI captures (primarily) the early biological dynamics of the test for recent infection (and should therefore be stable across contexts, which is critical to the overarching concept that a once-calibrated test should be useful if transferred to other contexts).

Estimation of the MDRI has traditionally relied on longitudinal data, consisting of test results observed over time after some (estimable) infection times for a sample of subjects. Through an extensive benchmarking exercise using simulated data, the accuracy and precision of various methods for estimating the MDRI from such longitudinal data were assessed in a number of modelled scenarios capturing what may be encountered in practice. In this exercise, incidence assays, or tests for recent infection based on single biomarkers, were considered, where measurements below a chosen test threshold indicate 'recent' infection. The methods of estimation therefore model either the biomarker measurements or the 'recent' and 'non-recent' classifications, as functions of time since infection.

Results highlight the danger of using estimation procedures that assume single continuous sojourns in the state of 'recent' infection, such as conventional survival analysis approaches. Simplistic approaches, such as the interpolation of biomarker readings (while allowing for multiple transitions between the 'recent' and 'non-recent' states) are useful for obtaining 'quick and dirty' estimates provided the times between visits are sufficiently small.

Formal regression approaches were generally the strongest. While non-linear mixed models for the biomarker readings most comprehensively captured the expected real-world features of the data (such as subject-specific evolutions of the biomarker), they were computationally demanding and required a higher level of analytical sophistication. Furthermore, when analysing any given dataset, knowledge of the biomarker, plots of the data, and model fit diagnostics should be carefully used to inform parametric assumptions, at the risk of large bias occurring otherwise. While not accounting for the

[31] The general framework for inferring incidence and defining test characteristics of relevance for incidence estimation is presented in Chapter 2.

subject-level clustering of the data, the linear binomial regression models proved particularly useful – they provided accurate results, were potentially less sensitive to parametric assumptions, were computationally stable, and could be fitted using standard statistical software. The benefits of including random effects or changing to non-linear model forms could be considered, although results suggest current implementations of binomial regression perform well. However, these extensions may become important when operating in scenarios outside of those considered. For example, if study drop-out depends on the observed biomarker readings (that is, data is missing at random) over the relevant post-infection timescale set by T, then random effects could be included to control for bias that may otherwise arise. Other potential sensitivities, for example, to non-normal distributions of random effects, could also be explored in more detail.

Uncertainty in infection times, or rather the violation of assumptions about infection times, poses a particular challenge to obtaining unbiased MDRI estimates. While computationally demanding, some research groups have formally accounted for this uncertainty in infection times, by using a marginal likelihood function (integrating out the infection times) [97, 103], sampling infections times [101], or incorporating prior distributions for infection times in a Bayesian model fitting [94]. However, assumptions about the timing of infection are still required (to inform the distributions used). Other groups have attempted to use the biomarker readings themselves to estimate infection times [17, 92, 96, 100, 102]. In this exercise, a single, standard diagnostic test, such as an enzyme immunoassay (EIA), was assumed to be used at all visits, and therefore the analyst would only know that a subject's infection (as defined by the diagnostic test) occurred between the last HIV-negative and first HIV-positive visit. In settings where various HIV diagnostic tests are used and the testing histories of subjects are documented, this external data could be used to inform the distributions of infection times.[32] For example, a subject with detectable p24 antigens and undetectable antibodies at a visit would have been infected within the preceding few weeks [101, 180, 181].

As guidance for MDRI estimation is further developed, the nature of tests for recent infection that are expected to be used in the future should be kept in mind. For example, as the field moves towards tests that rely on multiple biomarkers, some methods for estimating the MDRI (such as those that utilise the dichotomous test classifications)

[32] The inference of infection times using diagnostic testing histories is briefly explored in Section 4.3.

would be more amenable to this extension than others (such as those that model biomarker measurements, where many parameters would be needed to describe the multiple biomarkers and the relationship among them). Also, biomarkers may represent complex summary metrics of multiple responses, or quantify biological processes that are not well-understood, and thus approaches for choosing and testing parametric assumptions may require particular attention. Another nuance worth noting is the assumption of guaranteed survival until T after infection implicit in most MDRI estimations. This assumption may often lead to little bias in the MDRI for small values of T, but, in settings where early mortality is high or the value of T is chosen to be relatively large (to capture an enduring 'recent' state), analyses to estimate the MDRI should incorporate data on survival.

It is hoped that the results presented here will help inform the design of studies, which are costly and challenging to conduct. For example, particularly in low HIV incidence settings, hundreds of HIV-negative subjects need to be followed to obtain just a handful of subjects who become HIV-positive. While limited resources will always restrict study design options, useful MDRI estimates can be obtained in a range of realistic scenarios. It was only in the extreme scenarios of sparse data that all methods of estimation became problematic. The results of this benchmarking exercise can therefore provide guidance on how best to direct efforts – for example, by drawing attention to the varying benefits of increasing the sample size versus increasing the number of visits per subject, or of having stringent visits times that are the same for all subjects versus allowing for variability among subjects.

The simulation platform that has been developed can now be used to extend this benchmarking exercise, for example, to include other hypothetical scenarios, assess future proposed estimation procedures, or investigate confidence interval coverage. The simulation approach also holds the promise of extending the use of this environment to explore other topics related to cross-sectional incidence surveillance, particularly those that are intractable to being explored analytically – such as the detection of incidence trends in populations.

4.2 Redefining Entry into 'Recent' Infection for More Accurate Test Characterisation

An obstacle to accurate estimation of the MDRI is the unobservable infection times of subjects in the sample, where infection typically refers to *detectable* infection as defined by the HIV diagnostic test being used. As highlighted by the results in Section 4.1, substantial biases can arise when incorrect assumptions are made about subjects' testing behaviours. Whether simply using expected infection times in analyses [95, 96, 100, 102] or formally accounting for their uncertainty [91, 94, 97, 101, 103], assumptions about the relationship between infection and visit times are inevitably required. Alternatively, infection times could be estimated from the observed recent infection test biomarker measurements [17, 92, 96, 100, 102], rather than relying on external assumptions about the distributions of infection times between visits. However, the extrapolation of biomarker readings to times earlier than the first HIV-positive visit requires assumptions or knowledge about the very early dynamics of the biomarker.

An approach for redefining the 'HIV-negative', 'HIV-positive and recently infected' and 'HIV-positive and non-recently infected' states is therefore proposed, and is intended to reduce the reliance on assumptions about testing behaviour and on extrapolation of biomarker readings back in time when estimating the MDRI. The concept is demonstrated using the data simulation platform and estimation methods that have been presented in Section 4.1. While the approach can be generalised, a test for recent infection based a single biomarker is considered below.

The approach entails redefining the empirically observed 'HIV-positive' state by introducing a lower 'diagnostic' threshold on the dynamic of the biomarker for recent infection. In a cross-sectional survey, a subject who is classified as infected by the HIV diagnostic test (which may be, for example, a standard enzyme immunoassay or Western blot) must also return a recent infection biomarker measurement above this diagnostic threshold to be classified as 'HIV-positive' for purposes of estimating incidence. Effectively, an artificially less-sensitive HIV diagnostic algorithm is created. When estimating the MDRI from longitudinal data, benefits are gained if 'infection' times, defined consistently with the diagnostic algorithm described, can be more accurately

modelled. As illustrated in Figure 4.3, a subject is now 'HIV-positive and recently infected' when the biomarker measurement is between some lower threshold Y_L (acting as a diagnostic threshold) and upper threshold Y_U (distinguishing 'recent' from 'non-recent' infection as before), and the MDRI summarises the average time that this occurs (within time T of entering the 'HIV-positive' state).

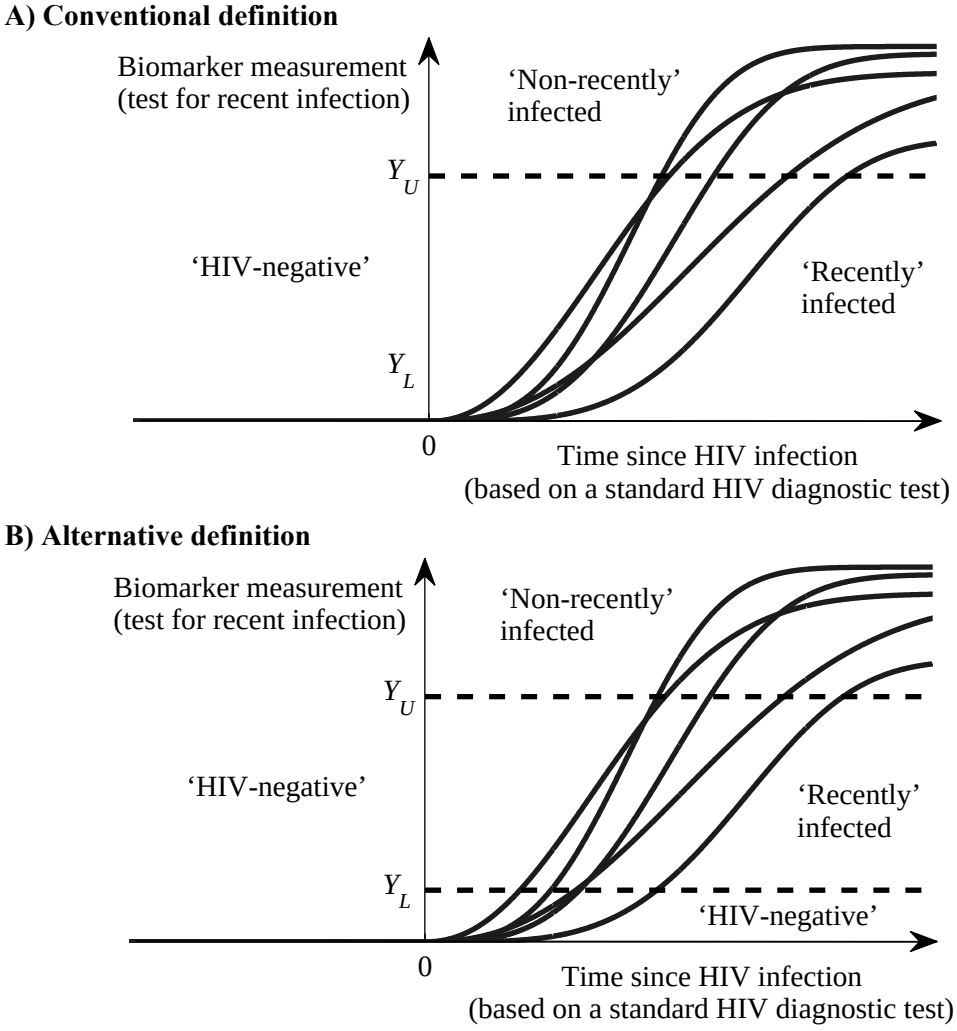


Figure 4.3: Conventional and alternative definitions of the 'HIV-negative', 'HIV-positive and recently infected' and 'HIV-positive and non-recently infected' states
The three states of relevance for incidence estimation, namely the 'HIV-negative', 'HIV-positive and recently infected' and 'HIV-positive and non-recently infected' states, as defined by the HIV diagnostic test and the biomarker for recent infection, are shown. In A), which illustrates the conventional definition used, a subject is 'HIV-positive' if diagnosed as infected by the HIV diagnostic test, and then 'recently' infected if the biomarker for recent infection is below a threshold Y_U. In B), which illustrates the alternative definition of states proposed, a subject is 'HIV-positive' if diagnosed as infected by the HIV diagnostic test *and* the biomarker for recent infection is above a threshold Y_L, and is then 'recently' infected if the biomarker is between Y_L and Y_U.

The simulation platform described in Section 4.1 was used to generate datasets, for estimation of the MDRI using both the conventional and alternative definitions of 'HIV-positive and recently infected'. Visits times for subjects were generated as in the base case scenario, and the distribution of infection times between HIV-negative and HIV-positive visits (all defined according to some standard HIV diagnostic test) was tuned to create three scenarios which capture: (i) test-neutral behaviour where infection times were uniformly distributed in the infection interval ($\alpha_{inf} = \beta_{inf} = 1$), (ii) test-seeking behaviour where infection times were closer to the first HIV-positive visits on average ($\alpha_{inf} = 3$, $\beta_{inf} = 1.5$), and (iii) test-deferring behaviour where infection times were closer to the last HIV-negative visits on average ($\alpha_{inf} = 1.5$, $\beta_{inf} = 3$). In all three scenarios, biomarker measurements were generated from the base case biomarker model, using parameters that produced a slowly evolving signal and little inter-subject variability ($\underline{\mu}_b$=[100,250,2.7], the standard deviations of the height, scale and shape random effects were 5, 5 and 0.5 respectively, there was no correlation between random effects, and measurement noise had a constant standard deviation of 2 biomarker units).

The MDRI was estimated by linear interpolation of biomarker readings, as described in Section 4.1.2 (Method 2). The average time a biomarker is between two thresholds Y_L and Y_U can be expressed as the difference between (i) the average time the biomarker is below Y_U, and (ii) the average time the biomarker is below Y_L. Therefore, conveniently, no modification of the estimation method was required, although it outputs an estimate of the average time that a biomarker is below a specified threshold (rather than between two thresholds). To estimate the MDRI under the conventional definition of states, the estimation method was applied using a threshold of $Y_U = 60$. To estimate the MDRI under the alternative definition of states, the estimation method was then also applied using a threshold of $Y_L = 10$ and the difference between the results for Y_U and Y_L taken. By design, the 'recently' infected state persists for less than a couple of years, and therefore T was chosen to be large enough for the MDRI to capture all 'recent' results.

Results from estimating the MDRI, from the 1 000 datasets generated for each of the three scenarios, are shown in Figure 4.4. Bias in estimation of the MDRI was reduced by moving from the conventionally-defined MDRI (biomarker readings below Y_U indicate 'recent' infection) to the alternatively-defined MDRI (readings between Y_L and Y_U indicate 'recent' infection). For the alternatively-defined MDRI, there was little bias because the vast majority of subjects (82%) had readings below Y_L at their first HIV-

positive visits (where, for purposes of referring to the data, last HIV-negative and first HIV-positive visits still relate the standard HIV diagnostic test). For these subjects, the 'infection' times now of interest (that is, when biomarkers exceeded Y_L) were estimated by interpolating between observed biomarker measurements (which were one month apart). In other words, for the vast majority of subjects in the sample, no assumptions were made about when infections became detectable by the HIV diagnostic test (in the three month intervals between the last HIV-negative and first HIV-positive visits) nor was there any extrapolation of biomarker readings to times earlier than first HIV-positive visits. These assumption and extrapolations were unavoidable in estimation of the conventionally-defined MDRI.

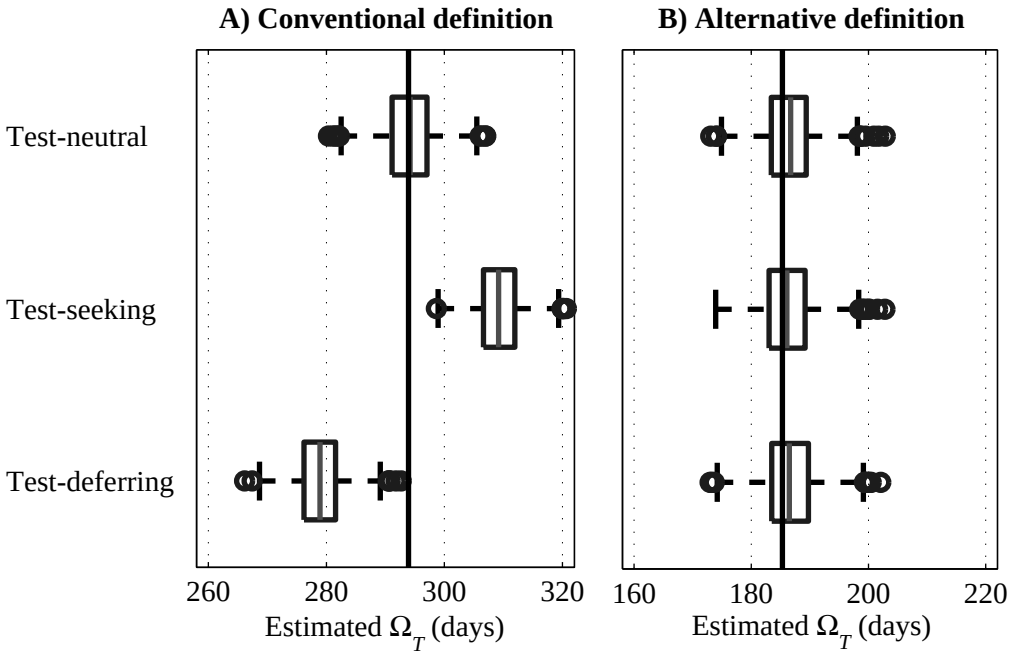


Figure 4.4: Box-and-whisker plots of the mean duration of recent infection point estimates, using conventional and alternative definitions of the 'HIV-positive and recently infected' state, in scenarios capturing different testing behaviours

Box-and-whisker plots summarise the 1 000 point estimates for the MDRI, Ω_T (days), in test-neutral, test-seeking and test-deferring scenarios, where the MDRI was estimated using each of two definitions for the 'HIV-positive and recently infected' state. The MDRI was estimated using linear interpolation ($T = 2$ years). In A), a biomarker reading below $Y_U = 60$ indicates 'recent' infection (conventional definition), while in B), a biomarker reading between $Y_L = 10$ and $Y_U = 60$ indicates 'recent' infection (alternative definition). For each scenario, the box and dividing line indicate the central 50% and median of estimates respectively, and whiskers and circles capture remaining data points and outliers respectively (outliers are more than 1.5 times the interquartile range away from the central box). The vertical black line indicates the true MDRI.

The definitions of the MDRI and FRR are now anchored by an 'infection' time that captures when a subject's biomarker for recent infection crosses above the threshold Y_L. This is familiar in design – previously, the infection time captured when some viral or host response (that grows after HIV transmission) crossed above some 'detectability' threshold of an HIV diagnostic test (for example, when seroconversion occurs if considering an antibody-based test). The differences are that a biomarker for recent infection produces a signal that will more slowly evolve and has a more useful dynamic range, compared to that of a diagnostic test, and the 'diagnostic' or 'detectability' threshold is set higher than necessary to detect the virus or an immune response.

By creating a less-sensitive HIV diagnostic algorithm, a time lag is introduced to incidence estimation as subjects who are infected in the weeks preceding a surveillance survey (and therefore return biomarker measurements below Y_L) would not contribute to the measured incidence. However, the weighting of incidence that is measured is already stretched over several months prior to a study, and therefore any blurring or shifting of this weighting over a few weeks (for suitable choices of Y_L) would not be meaningful. Also, while it is a fine detail that relates to any definition of 'infection', it is possible that a subject can fluctuate in and out of the 'HIV-positive' state for some short period (although this is unlikely to be observed in practice given reasonable visit gaps), leading to some negligible blurring of the weighting function.[33]

By introducing the lower threshold Y_L, there is a trade-off between increased accuracy and decreased precision of incidence estimation. As the discussion above highlights, MDRI estimates become less prone to bias, and will therefore bias incidence estimates less. However, the state of 'recent' infection becomes more transient, and therefore the variance of incidence estimates increases (as it becomes more difficult to observe subjects in this state). This gain in accuracy and loss in precision would grow as the threshold Y_L is increased, and therefore a suitable balance between these would need to be found.

[33]Section 2.2.1 provides some discussion about the relationship between the weighting function for past incidence and diagnostic sensitivity. Subtleties around the detection of HIV are noted in passing here merely for completeness, but the impact of these is considered to be negligible for the analyses presented.

The potential of this approach would also depend on the particular dynamics of the biomarker for recent infection and subjects' visit gaps. Frequent testing is not feasible for the large numbers of HIV-negative subjects enrolled in a prospective study, and therefore frequent follow-up is typically initiated only after a subject presents as HIV-positive. To restrict bias in MDRI estimation, namely by reducing the reliance on assumptions about (standard diagnostic) infection times and on extrapolation of biomarker readings, the threshold Y_L should be large enough for subjects' biomarker measurements to have not evolved far beyond it, preferably still be well below it, at their first HIV-positive visits. It should also be large enough to lie above any early noisy biomarker dynamics, so that the evolution of the biomarker in its vicinity can be confidently modelled from the data. All of this needs to be achieved while keeping the value of Y_L small enough to obtain a suitably enduring state of 'recent' infection (for an appropriate choice of the upper threshold Y_U for distinguishing between 'recent' and 'non-recent' infections).

In principle, the framework for cross-sectional incidence estimation allows for arbitrary definitions of the 'HIV-negative', 'HIV-positive and recently infected' and 'HIV-positive and non-recently infected' states (although not all definitions will provide useful weightings of past incidence). Accurate incidence estimation then requires only that tests for recent infection are *consistently* characterised and applied in cross-sectional incidence studies. This illustrative analysis shows how alternative definitions of HIV diagnostic states and classifications of infections as 'recent' or 'non-recent' could be considered, in an attempt to reduce artefacts in test characterisations arising from modelling unknown infection times. As with estimating any parameter, there is unavoidable uncertainty in both MDRI and incidence estimates, and therefore any gains in accuracy should be viewed in the light of this sampling variability. While the dynamics of biomarkers developed in the future may make this approach more appealing, current experiences with data suggest there may be limited benefit to this strategy.

4.3 Estimation of Infection Times from Diagnostic Testing Histories

Throughout this thesis, 'infection' has consistently referred to 'detectable infection', which is largely sufficient for the needs at hand. In this section, some finer points about estimating infection times are considered, and this warrants a review of the applicable terminology. The time of exposure refers to the time of HIV acquisition or transmission (for example, during a sexual contact), while the time of infection or test conversion refers to when HIV becomes detectable by some HIV diagnostic test (which is typically only weeks after exposure). While subjects can only be tested for detectable infection in a study, results are dependent on the particular diagnostic test used, and therefore exposure provides a useful general reference event for analytical purposes.

In a prospective study intended to generate longitudinal data for estimation of the MDRI, a subject's exact time of infection is unknown, but constrained to lie between the subject's last HIV-negative visit and first HIV-positive visit. When HIV exposures do not influence the visit times of subjects, it is reasonable to assume that infection is equally likely to have occurred at any time in this interval. This leads to uniform distributions or flat priors for the infection times [91, 94, 97, 101, 103], often summarised into expected infection times at the midpoints of intervals between visits [95, 96, 100, 102]. Implicit in this approach is the assumption that the same HIV diagnostic test is used at all visits, and that no staging information is produced by the diagnostic test (that would further influence the analyst's view on when a subject was infected). Furthermore, by using the estimated MDRI for incidence estimation, it is also assumed that the same diagnostic test (or one with the same sensitivity) is used in the incidence survey. This section aims to consider briefly these important analytical subtleties, which are overlooked in the literature. The estimation of exposure or infection times from subjects' diagnostic testing histories is briefly explored, and the need for consistent use of diagnostic rules for both test characterisation and surveillance application is highlighted.

A large number of HIV diagnostic tests have been developed, each based on the detection of specific components of the virus itself or specific host antibody responses (or both) [182]. Tests may also differ by type of specimen analysed (such as plasma, dried blood

spots or saliva) and whether they need to be performed in a laboratory or can be completed on site ('rapid tests'). Classes of diagnostic tests that detect antibodies include enzyme immunoassays (EIAs) – sometimes called enzyme-linked immunosorbent assays (ELISAs), Western blot and immunofluorescence assays (IFAs); while those that measure the virus itself include assays to detect p24 antigens and nucleic-based assays to detect HIV ribonucleic acid (RNA). Also, to reduce diagnostic misclassification, algorithms of HIV diagnostic tests are often used – for example, two EIAs may be performed, and if at least one detects HIV infection, a confirmatory Western blot test is conducted.

Despite the large variety of HIV diagnostics that are available and in use, in principle, the behaviour of any given HIV diagnostic test (or component of a diagnostic algorithm) that is applied to a subject who has acquired HIV can be summarised into the test sensitivity as a function of time since exposure. The sensitivity of a test is the probability that it correctly detects HIV, and is shown for a hypothetical test in Figure 4.5 (part A). It is possible that this function is not monotonically increasing – for example, p24 antigens may decline to undetectable levels after an initial period of detectability. When a subject (who acquires HIV at some time, during the study) is tested for HIV at a study visit, the result of the diagnostic test can be combined with the test's sensitivity to infer likely HIV exposure times, as shown in Figure 4.5 (part B).

Expressing this formally, the likelihood of observing the diagnostic test result d at calendar time t_{test} (choosing an arbitrary reference time 0), as a function of the subject's HIV exposure time u, is

$$L(d|u) = \big(P_+(t_{test} - u)\big)^d \big(1 - P_+(t_{test} - u)\big)^{1-d}, \tag{4.10}$$

where d equals 1 if the diagnostic result is HIV-positive and 0 if it is HIV-negative, and $P_+(s)$ is the sensitivity of the test at time s after HIV exposure. This concept can be extended to account for the subject's full diagnostic testing history. Various HIV diagnostic tests (or algorithms) may be applied at each visit, and all visits times and individual diagnostic test results then collectively used to infer the subject's exposure time, as outlined below.

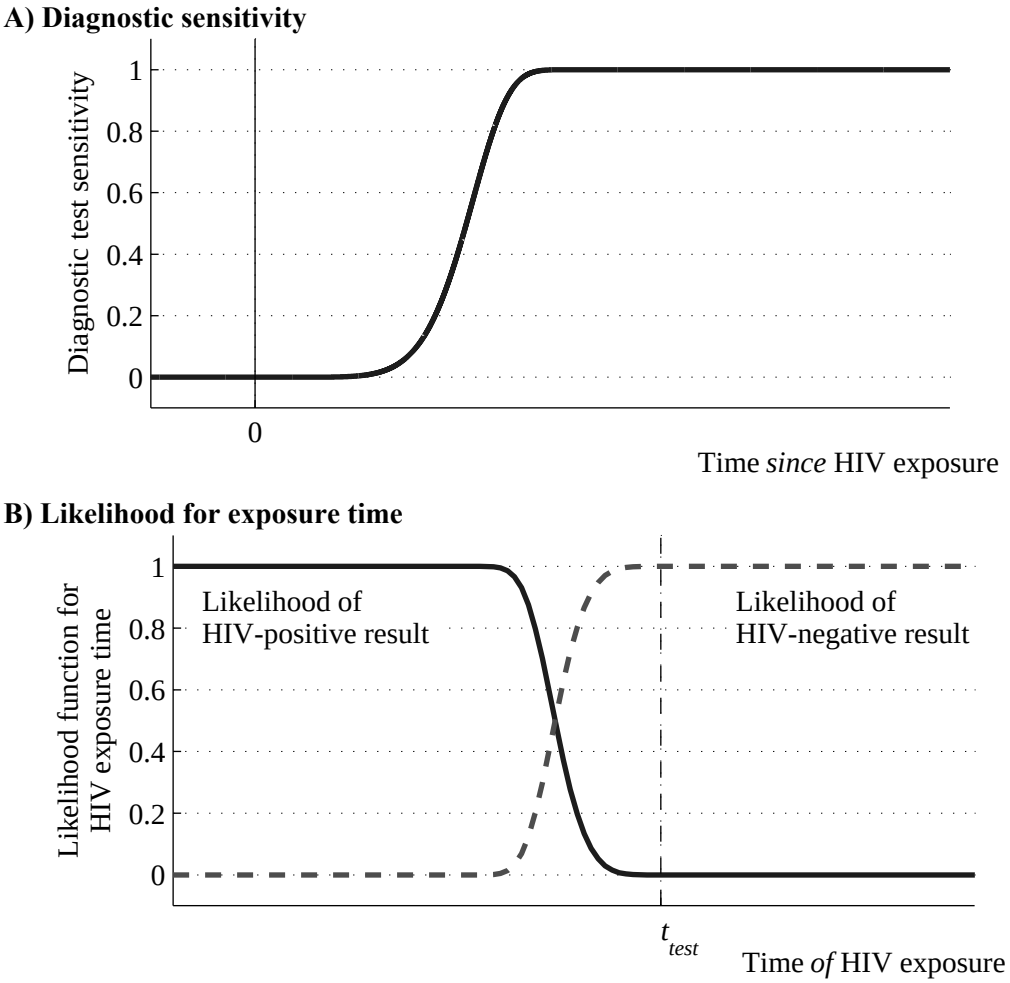


Figure 4.5: Sensitivity of a hypothetical HIV diagnostic test and implied likelihood function for a subject's HIV exposure time
In A), the sensitivity of a hypothetical HIV diagnostic test is shown as function of time since HIV exposure, where the test's sensitivity is its probability of correctly detecting HIV. In B), the likelihood function for a subject's exposure time is shown. This provides the probability of observing a result for the diagnostic test, which is applied at some time t_{test}, as a function of the HIV exposure time, and depends on the test's sensitivity. The solid and dashed curves are the likelihood functions when the observed diagnostic results are HIV-positive and HIV-negative respectively.

To simplify some of the interpretations that follow, it is now assumed that a subject has a single time of entering the infected state (defined by some diagnostic test) and remains in the state thereafter (this view can be relaxed, as captured by the generality of Equation (4.10)). The sensitivity function is now viewed as the cumulative distribution function for the time from exposure to (detection of) infection. The calendar time of the subject's j^{th} visit is denoted by t_j and the result of diagnostic test k at that visit (if available) by $d_{j,k}$, which equals 1 if the result is HIV-positive and 0 if HIV-negative

($j = 1,2, \ldots n_j$ and $k = 1,2, \ldots, n_k$). First considering a particular diagnostic test K, the likelihood of observing its results at all visits, as a function of exposure time u, can be derived:

$$L(\{d_{j,K}\}_j|u) = P_{K+}(t_{pos} - u) - P_{K+}(t_{neg} - u), \quad (4.11)$$

where $P_{K+}(s)$ is the sensitivity of diagnostic test K at time s after exposure, t_{pos} is the earliest visit time where the diagnostic test produced an HIV-positive result, and t_{neg} is the latest visit time where the diagnostic test produced an HIV-negative result. If there are no negative test results, t_{neg} can be set to negative infinity $(P_{K+}(t_{neg} - u) = 0)$ and if there are no positive test results, t_{pos} can be set to positive infinity $(P_{K+}(t_{pos} - u) = 1)$.

Assuming independence of diagnostic tests, the overall likelihood function for the time of HIV exposure u is

$$L(u) = L(\{d_{j,k}\}_{j,k}|u) = \prod_{K=1}^{n_k} L(\{d_{j,K}\}_j|u), \quad (4.12)$$

where $L(\{d_{j,K}\}_j|u)$ is given by Equation (4.11).

Examples of obtaining this net likelihood function from a subject's testing history are provided in Figure 4.6. In general, the lower bound for feasible exposure times is driven by the most-sensitive of the diagnostic tests that returned HIV-negative results, at the latest visit where any HIV-negative result was produced. Conversely, the upper bound for feasible exposure times is driven by the least-sensitive of the diagnostic tests that returned HIV-positive results, at the earliest visit where any HIV-positive result was produced.

A limitation of the framework presented above is the assumption of independence of HIV diagnostic tests. For example, it is conceivable that the virus and host response matures rapidly in some subjects (resulting in relatively early detection of HIV by a number of diagnostic tests) and slowly in other subjects (resulting in relatively late detection by multiple tests). A completely general framework would capture a multivariate distribution for the times from exposure to (detection of) infection by various HIV diagnostic tests, and a similar likelihood approach could be used to infer likely exposure times. However, the practical use of such a framework would rely on (currently unrealistic) inputs to describe the relationship among multiple tests. The analysis presented above may already be challenging to implement without further simplifications (discussed below), as it relies on full specifications of tests' sensitivity functions.

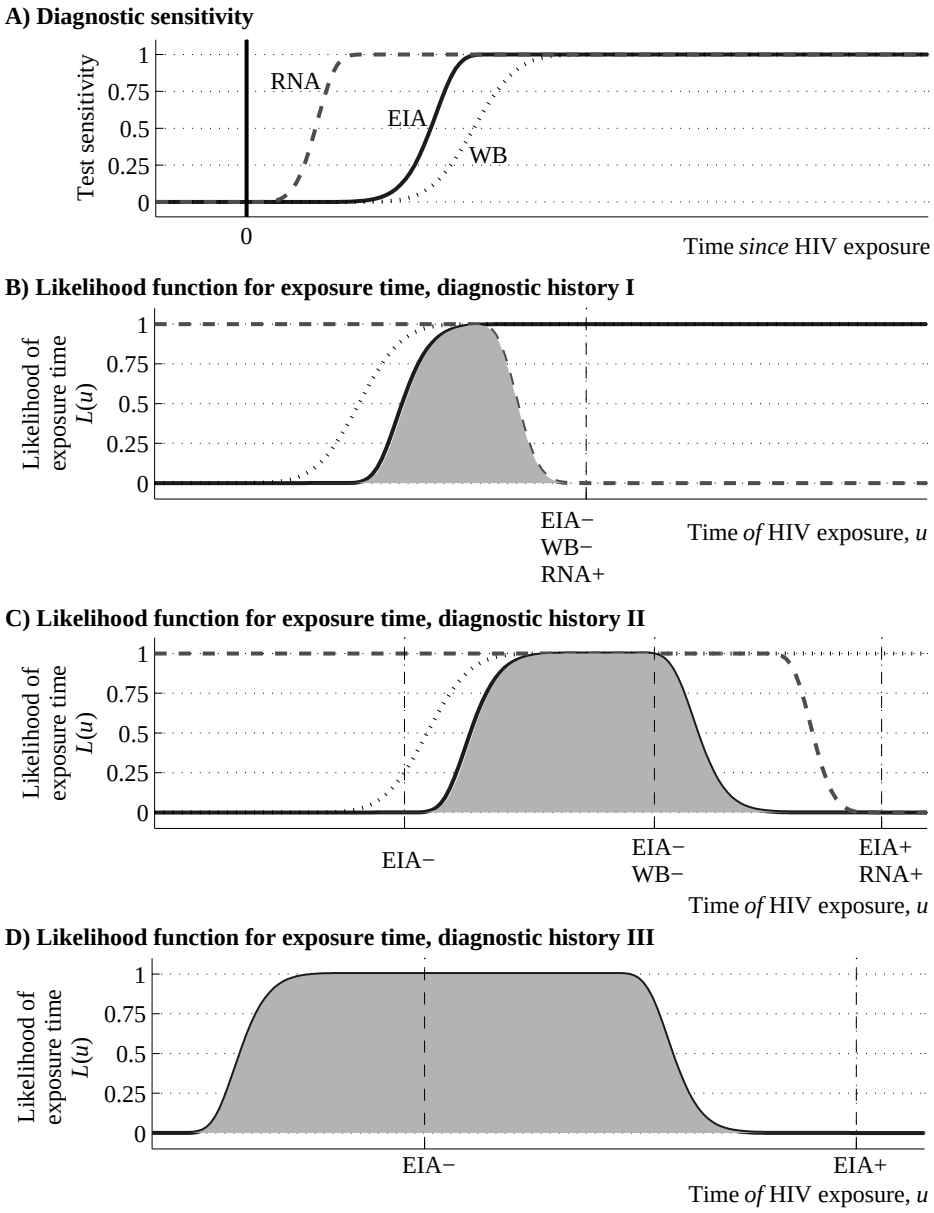


Figure 4.6: Likelihood functions for HIV exposure times for a number of hypothetical diagnostic testing histories

Three hypothetical diagnostic testing histories and resulting likelihood functions for a subject's HIV exposure time are provided. In A), presumed sensitivities of three diagnostic tests are shown as a function of time since exposure. In order of decreasing sensitivity, the three tests are an RNA-detection assay (dashed line), an EIA (solid line) and Western blot (dotted line). Possible testing histories for a subject are captured in B) to D). In each figure, the (dashed, solid and dotted) lines capture the likelihood functions for the exposure time based on the individual diagnostic tests, and these are then used to obtain the overall likelihood function shown by the curve with shaded area (scaled to have a height of one). Diagnostic test results and the times of visits are captured on the x-axis. In B), the subject has a single visit with both HIV-positive and HIV-negative results; in C), the subject has multiple test dates and diagnostic test results; and in D), a single diagnostic test is used and only the last HIV-negative and first HIV-positive visits are shown.

The discussion above has been framed in terms of inferring a subject's HIV exposure time from the testing history. In practice, for unbiased incidence estimation, the MDRI should be estimated using a definition of infection that is consistent with what is detectable in the incidence surveillance study. For example, if a subject's HIV status is to be determined by Western blot in the surveillance survey, then the (estimated) time of becoming Western blot positive should be used as the infection time when analysing data to estimate the characteristics of the test for recent infection. The framework presented above (summarised by Equation (4.12)) can still be used, but now the reference time needs to be the time of test conversion by the diagnostic algorithm used in the incidence study, rather than the exposure time. In other words, $P_{K+}(s)$ appearing in Equation (4.11) should provide the sensitivity of diagnostic test K at time s after test conversion (when using the test used in the incidence study), and a likelihood function $L(w)$ for the appropriately-defined test conversion or infection time w would be obtained.

Given typical visit gaps (of a few months) and the much shorter times from exposure to infection by diagnostic tests, likelihood functions may often plateau (implying non-unique maximum likelihood estimates of test conversion times). One approach for interpreting the information contained in the likelihood function is through a posterior distribution for the test conversion time of the subject. Using a uniform prior distribution for the test conversion time, its posterior distribution is proportional to the likelihood function (by Bayes theorem). Analyses for estimating the MDRI could make use of the posterior expected test conversion time, or formally incorporate the full distribution – for example, as prior knowledge, for purposes of sampling possible test conversion times, or for producing a marginal likelihood function for the MDRI model parameters [91, 94, 97, 101, 103].

In the case that the same diagnostic test is used at all visits in the MDRI study (such as in Part D of Figure 4.6) and in the incidence study, the above analysis recovers a uniform posterior distribution for the test conversion time in the interval between the last HIV-negative and first HIV-positive visits. This is the special case encountered in many of the published analyses [91, 94, 97, 101, 103], although one may expect that more complex diagnostic testing histories often arise, and therefore efforts should be made to document these and incorporate them into analyses.

Given sufficient information about the sensitivities of diagnostic tests, a subject's full testing history could be used to formally make inferences about the 'infection' time. The

empirical category of 'infected' will depend on the diagnostic algorithm, and should be consistent across MDRI estimation and incidence studies. Given that limited knowledge about the sensitivities of diagnostic tests may be available, simplifications of the framework presented here may be required. For example, for a diagnostic test which is expected to have a sensitivity that rapidly climbs from 0 to 1, *after* some period of undetectability of HIV following exposure, the test dynamic could be summarised by the average time from exposure to (detection of) infection, thus neglecting any inter-subject variability. Inputs describing these average durations could be obtained from studies that have assessed diagnostic test performance (for example, see [180, 181, 183-185]). In this case, all posterior distributions for infection times would be uniform distributions (with bounds determined by the data and average durations).[34]

The preceding analysis also highlights that more sensitive diagnostic tests naturally extend the MDRI for any recent infection test. The trade-off between the added expense of a more sensitive diagnostic test and more precise incidence estimation from a (probably only very marginally) larger MDRI would need to be assessed in practice.

[34] This simplification for estimating the (posterior distribution of the) infection time, using a subject's diagnostic testing history, is applied in the analysis presented in Chapter 5.

Chapter 5

Theory to Practice: Characterisation of Candidate Tests for Recent Infection

An important, practical application of the framework and methodology for test characterisation described in the previous chapters is presented here. Over the last four years, the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA) has established a vast repository of specimens and produced a large volume of data to characterise, optimise and compare proposed tests for recent infection. The consortium's first primary analysis outputs, which describe the behaviours of five prominent incidence assays, are presented in Section 5.1. The report is a reproduction of a published journal article [32].[35]

[35] Section 5.1 presents a manuscript that has been published as: 'Kassanjee R, Pilcher CD, Keating SM, Facente SN, McKinney E, Price MA, Martin JN, Little S, Hecht FM, Kallas EG, Welte A, Busch MP, Murphy G, on behalf of the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA). Independent assessment of candidate HIV incidence assays on specimens in the CEPHIA repository. *AIDS*. 2014; 28(16):2439-2449'. The manuscript was published under the terms of the Creative Commons License Attribution-Non-commercial No Derivative 3.0 (CCBY NCND), and therefore no permission was required from the publishers to reproduce the work. The manuscript was written by RK. GM, AW, CDP and MPB conceived the study design and sourced funding. RK and AW led the data analysis. CDP, SNF, SJL, MAP, JNM, EGK and FMH led on specimen acquisition and related data collection. GM, MPB, SMK and EM led on assay performance and quality, and assay results reporting. All authors assisted in the interpretation of findings, provided input and suggestions for analysis, and reviewed the manuscript. Funding for CEPHIA was provided by the *Bill and Melinda Gates Foundation* (grant OPP1017716).

5.1 Independent Assessment of Candidate HIV Incidence Assays on Specimens in the CEPHIA Repository

5.1.1 Introduction

Reliable measurement of HIV incidence (the rate of new infections) is essential for monitoring the epidemic, assessing interventions and planning studies. Traditionally, incidence is measured by counting the number of new infections acquired in a cohort of subjects followed-up over time. However, such longitudinal studies are often costly, time consuming, and unrepresentative. Therefore, the estimation of incidence from cross-sectional surveys, using 'incidence assays' that distinguish 'recent' from 'non-recent' infection, has attracted wide interest [9, 10, 12-14].

Cross-sectional surveillance is founded on the heuristic that a high prevalence of 'recent' infection indicates a high incidence [7, 15]. However, current incidence assays that provide a reasonably enduring state of 'recent' infection also tend to produce substantial 'false-recent' results at large times after infection [9, 10, 12-14, 75-77, 79, 81, 112]. As methodology matured [11, 16-27, 29, 82, 186], a general theoretical framework was developed that supports the consistent analysis of 'false-recent' results [29].[36] However, there have not been independent assessments of candidate assays, or consensus metrics of an assay's utility for incidence estimation.

In 2010, the *Bill and Melinda Gates Foundation* supported the establishment of the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA) [47]. Over the last four years, CEPHIA has entered into collaborations and material transfer agreements to establish a large repository of valuable plasma specimens with sufficient clinical background data. Test developers can apply for access to a small

[36]The derivation of the general incidence inference framework, performed as part of earlier work, is presented in Chapter 2.

'qualification panel' of specimens, and, if the assay is suitably promising, the assay can be independently applied (by a CEPHIA laboratory) to a much larger 'evaluation panel'.

Results are presented below for the first five assays that have successfully passed through the full evaluation, namely Limiting Antigen-Avidity (LAg) [98], BED [92], Less-Sensitive or Detuned Vitros [100], Vitros Avidity [100] and BioRad Avidity [187]. In principle, a test for recent infection can be arbitrarily complex in design [9, 10, 12-14], and can be optimised by tuning numerous parameters. The present evaluation is of tests for recent infection which are based on single biomarkers, termed incidence assays, applied according to developers' test conditions and interpretive guidelines. Test optimisation, by the application of alternative thresholds in the interpretation of results, and using the assays in combination with one another or with supplemental markers (such as viral load), is ongoing.

Translating survey counts (of HIV-negative, 'recently' HIV-positive and 'non-recently' HIV-positive subjects) into incidence estimates [29] requires knowledge of two test properties:

- The mean duration of recent infection (MDRI), which is the average time spent alive and 'recently' infected, while infected for less than some time cut-off denoted by T.
- The false-recent rate (FRR), which is the probability that a randomly chosen subject, who is infected for longer than T, will produce a 'recent' result.

A 'Target Product Profile' for tests for recent infection has been developed and attracted some attention [13, 14, 188], and provides a number of objectives that incidence assays should meet to be of utility for incidence estimation. To achieve usefully precise incidence estimates, in real-world household surveys in high incidence settings, an incidence assay should have a sufficiently enduring MDRI (of around one year) and small FRR (definitely less than 2%, and ideally zero). Furthermore, for feasible widespread use of the assay, results should be highly reproducible, and the training, equipment and sample type requirements should be modest.

In this analysis, each assay's MDRI and FRR were evaluated. As the behaviour of incidence assays may vary across subpopulations [75-77, 79, 189], the characteristics of the incidence assays in various specimen sets were also explored.

5.1.2 Methods

The CEPHIA specimen repository and the evaluation panel

The CEPHIA repository is housed at *Blood System Research Institute* (San Francisco; CA) and currently consists of more than 5000 plasma specimens obtained from over 1200 subjects. The specimens used in this analysis were obtained through collaborations with blood banks, and clinical research studies enrolling and following subjects over time: *American Red Cross* [190]; *Blood Centers of the Pacific* [191]; *South African National Blood Service* [192]; *Hemocentro do São Paulo* [193]; the *University of California, San Francisco*, *Options* study [194]; *San Francisco Men's Health Study* [195]; the *San Diego Primary Infection Cohort* [196]; the multi-centre *AMPLIAR* cohort [197]; the multi-centre *International AIDS Vaccine Initiative* (IAVI) *African Early Infection Cohort* (Protocol C) [198]; and the *University of California, San Francisco*, *SCOPE* study [199].

Two 'panels' of specimens were created for the present purpose: a 250-member 'qualification panel' for preliminary assessments (see [41] for results); and a 2500-member 'evaluation panel' for the full assessments of assays showing suitable promise, which forms the basis of this investigation.

The evaluation panel specimens were drawn from 928 subjects and 60% of subjects contributed multiple specimens over time (these subjects contributed up to 13 specimens and a median of 3 specimens each). Follow-up after infection ranged from 1 week to more than 10 years, and the median follow-up time was 3 years (for subjects with estimable infection dates, as discussed below).

Laboratory procedures and interpretation of assay results

Each of the five assays measures an aspect of an individual's immune response, with measurements below some threshold interpreted as indicative of 'recent' infection.

BED [92, 166] and LAg [98, 113, 200] (Sedia Biosciences Corporation; Portland; OR) were developed specifically as incidence assays by the *Centers for Disease Control and Prevention* (CDC). The immunoglobulin G (IgG) capture BED enzyme immunoassay (EIA) measures the proportion of IgG that is specific to HIV, and a normalised optical density (ODn) below 0.8 indicates 'recent' infection. The single-well Limiting Antigen-Avidity EIA is responsive to the avidity of HIV-1 specific IgG, as it presents marginally

low concentrations of a multi-subtype recombinant HIV-1 antigen, typically affording just a single binding site to the multivalent IgG or IgM antibodies. While a 'recent'/'non-recent' threshold of 1.0 ODn was initially proposed, this was recently revised to 1.5 [200, 201], following a review of the assay in which CEPHIA participated.

Both less-sensitive Vitros (LS-Vitros) and Vitros Avidity [100] are based on the Vitros ECi/ECiQ Immunodiagnostic System, a chemiluminescence assay that gives a quantitative measure of HIV antibodies (Ortho-Clinical Diagnostics, Inc.; Rochester; NY). For LS-Vitros, a reported signal-to-cutoff (S/C) below 20, for a diluted specimen, is interpreted as a 'recent' result. For Vitros Avidity, the ratio of the S/C in an aliquot treated with a chaotropic agent (guanidine) to that in an aliquot not thus treated yields an avidity index (AI), and a 'recent'/'non-recent' threshold of 60% is used to classify the infection.

The BioRad Avidity test [187] is based on a modification of the Genetic Systems HIV-1/HIV-2 plus O EIA (Bio-Rad Laboratories, Inc.; Hercules; CA), which involves the testing of each specimen in the presence and absence of a chaotropic agent (Diethylamine). The ratio of the reactivity of the treated to untreated aliquot produces an avidity index (AI), with values below 40% indicating 'recent' infection.

All assays were applied according to developers' standard operating procedures and package inserts [92, 98, 100, 166, 200], and protocols are available on the CEPHIA project website [47]. Testing was performed independently in CEPHIA laboratories, by technicians trained by the test developers and blinded to specimen background information. Three large volume 'control' specimens (obtained from blood donations, and chosen to represent a range of serological responses) were supplied to laboratory technicians with each panel, for regular confirmation of reproducibility and stability of assays.

Data analysis

All data captured within CEPHIA are stored in a (MySQL) relational database.[37] Database queries linked assay results to the background information on subjects and specimens for data analysis (performed in Matlab R2013b, the MathWorks Inc.).

Test properties were evaluated in specimen sets defined by stratifying on treatment history, viral load, CD4 cell count, time from infection to specimen draw, and HIV subtype (based on country, for the 48% of specimens which lack explicit laboratory subtype confirmation). The performance of assays in 'elite controllers' (ECs), broadly defined as subjects who maintain undetectable or very low HIV viral loads without antiretroviral therapy (ART), is of particular interest. As the SCOPE study purposefully recruited ECs, specimens from these subjects were analysed separately. The subjects were ART-naïve (or without ART for at least 6 months), with all off-treatment viral load measurements (HIV-1 RNA) below 200 copies/ml and at least 50% of these measurements below 75 copies/ml.

The definitions of the MDRI and FRR rely on the previously mentioned construct of a post-infection time cut-off T [29]. If T is chosen to be too short, this limits the possible MDRI and typically raises the FRR. If T is chosen to be too long, it becomes difficult to obtain sufficient data to characterise the test with sufficient precision over this time post infection, and the MDRI will also develop variation by time and place (properties inevitable for the FRR) rather than capture stable biological properties of the test. A cut-off of $T = 2$ years was used throughout this analysis.

In practice, the notion of 'infection' implicit in the test property definitions refers to 'detectable infection' – which depends on the particular HIV diagnostic test used in the incidence study. In this analysis, 'detectable infection' was defined as the time of seroconversion on an HIV viral lysate-based Western blot assay. Based on a methodology described elsewhere,[38] infection dates were estimated for the 56% of subjects who had recorded dates of last HIV-negative and first HIV-positive tests (not more than 120 days

---

[37] The CEPHIA database is currently administered by David Matten of SACEMA.

[38] Members of CEPHIA have prepared a (currently unpublished) manuscript describing the work of CEPHIA and outlining the envisioned development pathway for new biomarkers for recent infection. The framework for estimating infection dates from diagnostic test data, summarised in the manuscript, is presented in Section 4.3.

apart) and descriptions of the diagnostic assays used. Average durations of Fiebig stages [180, 181] were used to estimate times at which subjects seroconverted (corresponding to entering Fiebig stage 5). Subjects with unambiguous acute retroviral syndrome (ARS) symptoms onset dates [202-205] between their last HIV-negative and first HIV-positive test dates, were estimated to seroconvert 17 days after ARS onset (based on the observation that the incubation period of ARS symptoms is about 14 days [206-209], and that the time from exposure to Western blot seroconversion averages 31 days [180, 181]).

A number of methods can reasonably be used to estimate the MDRI, each with its own accuracy, precision and complexity – as explored in a separate, detailed benchmarking exercise [53].[39] In this analysis, binomial regression was applied; this is an approach that was found to be robust across a number of scenarios explored in the benchmarking project and has been previously used for MDRI estimation [101]. The model form is $g(P_R(t)) = f(t)$, where $P_R(t)$ is the probability of testing 'recent' at time t after infection, g is the chosen link function and $f(t)$ contains the model parameters, which are estimated by a maximum likelihood approach. Results from a four-parameter model form are presented, where g is the logit link, and $f(t)$ is a cubic polynomial in t (Model A). Data points more than $1.1 \times T$ post (estimated) infection were discarded before model fitting (Data Exclusion Rule I), with the aim of achieving the best fit of the model over [0,T] post-infection, while avoiding diluting the data around the boundary at T. The sensitivity of results when increasing the data exclusion cut-off to $2 \times T$ (Data Exclusion Rule II) was also considered. Variation in results was explored when fitting two other model forms, namely (i) a more restrictive two-parameter model where g is the log-log link and $f(t)$ is a linear function of $\ln(t)$ (Model B), and (ii) a flexible seven-parameter model where g is the logit link and $f(t)$ is a linear combination of the natural cubic spline basis functions with interior knots occurring every 3 months after infection, between 0 and $T = 2$ years after infection (Model C). In all cases, the MDRI, expressed mathematically as $\int_0^T P_R(t)\,\mathrm{d}t$, was estimated using the fitted $P_R(t) = g^{-1}(f(t))$ (negligible mortality within T post infection was assumed).

To correctly account for the structure of the data, in the absence of explicit subject-level clustering in the fitted models, bootstrapping was performed by sampling subjects (not

[39] The benchmarking of approaches for estimating the MDRI from longitudinal data is presented in Section 4.1.

observations) with replacement. The 2.5th and 97.5th percentiles of 10 000 MDRI estimate replicates provided 95% confidence interval (CI) limits [174].

A population-level FRR is inherently dependent on the epidemiological and demographic history of a study population [29], and so a set of specimens, such as in the CEPHIA repository, can only be used to estimate the FRR in well-defined subpopulations. Therefore, specimens from long-infected subjects were identified (specimens drawn at least T after the subject's first recorded HIV-positive visit and estimated Western blot infection time), and the proportion of 'recently' infected subjects estimated in each of the specimen sets described above. To capture subject-level clustering, when a subject provided more than one result to any FRR estimate, the most frequent classification was used (few subjects had equal numbers of 'recent' and 'non-recent' results, and each such subject contributed half to the aggregate count of subjects with majority 'recent' classifications). Exact Clopper-Pearson 95% CIs [210] are provided.

5.1.3 Results

The incidence assay dynamics, excluding specimens from treated subjects and SCOPE elite controllers, are shown in Figures 5.1 to 5.3. The evolution of assay readings by time since infection is shown in Figure 5.1. The distribution of results for specimens drawn more than $T = 2$ years after infection is shown in Figure 5.2. In Figure 5.3, the proportion of 'recent' results (assay measurements below the 'recent'/'non-recent' threshold) is plotted by time since infection and stratified by HIV subtype (A1, B, C and D). Note that (i) there is natural variability in biomarker maturation, leading to a significant number of subjects reaching the standard 'recent'/'non-recent' threshold more than one year, but often less than two years, post infection; and (ii) there is significant delay or failure to achieve maturation to 'non-recent' status among specimens of subtypes A1 and D.

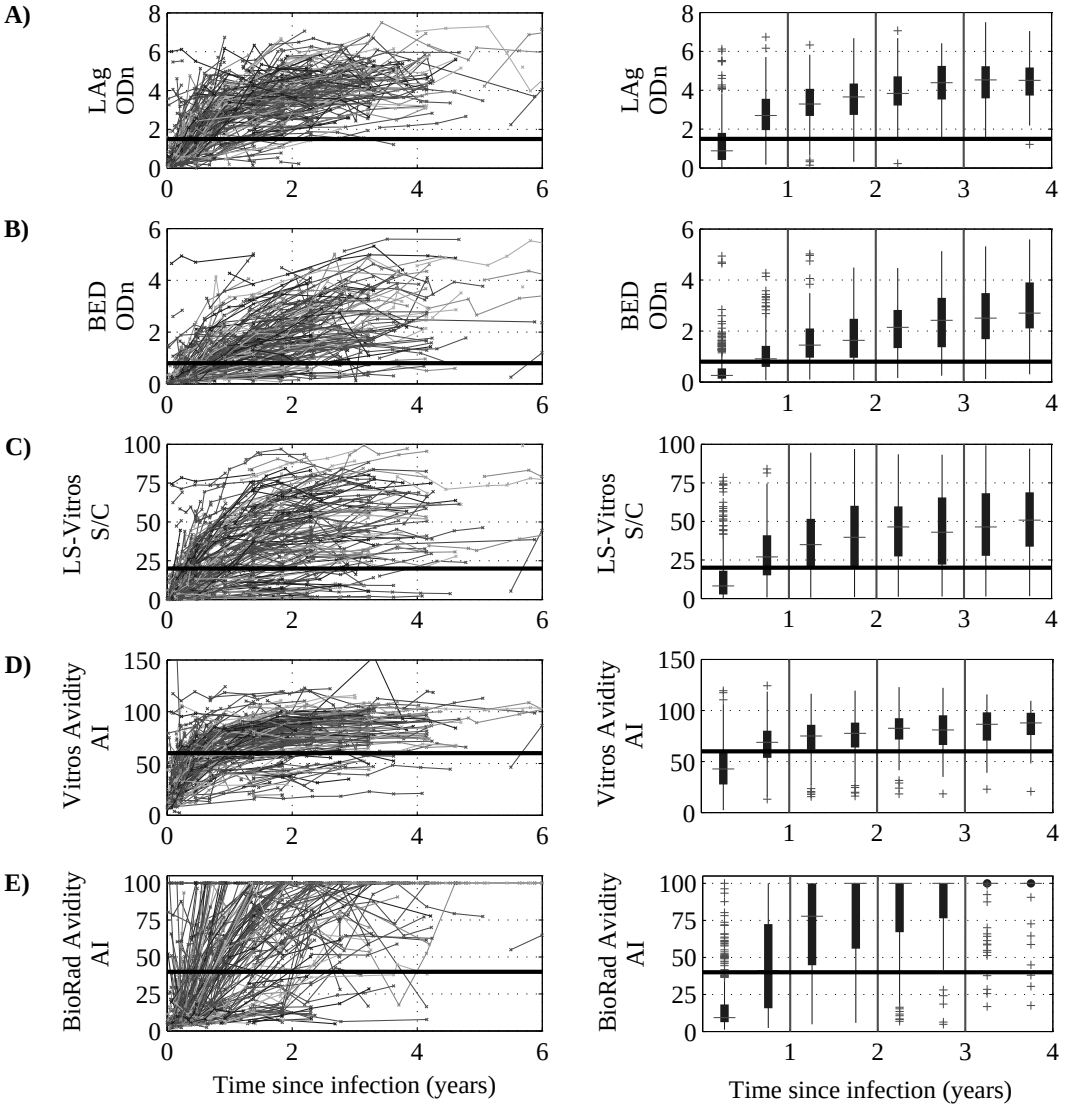


Figure 5.1: Spaghetti plots and box-and-whisker plots of incidence assay measurements observed over time after infection for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity

Incidence assay measurements are shown as a function of (estimated) time since infection (years), excluding treated subjects and identified elite controllers, for A) LAg, B) BED, C) LS-Vitros, D) Vitros Avidity and E) BioRad Avidity (1376 data points from 418 subjects). A spaghetti plot (left) shows subjects' trajectories, and box-and-whisker plots (right) show percentiles of measurements in 6-monthly intervals of time post infection (the central 50% and median of measurements are captured by the box and dividing line respectively, and whiskers and '+' symbols capture remaining measurements and outliers respectively; there are 40-450 data points in each group). 'Recent'/'non-recent' thresholds are shown by horizontal solid lines.

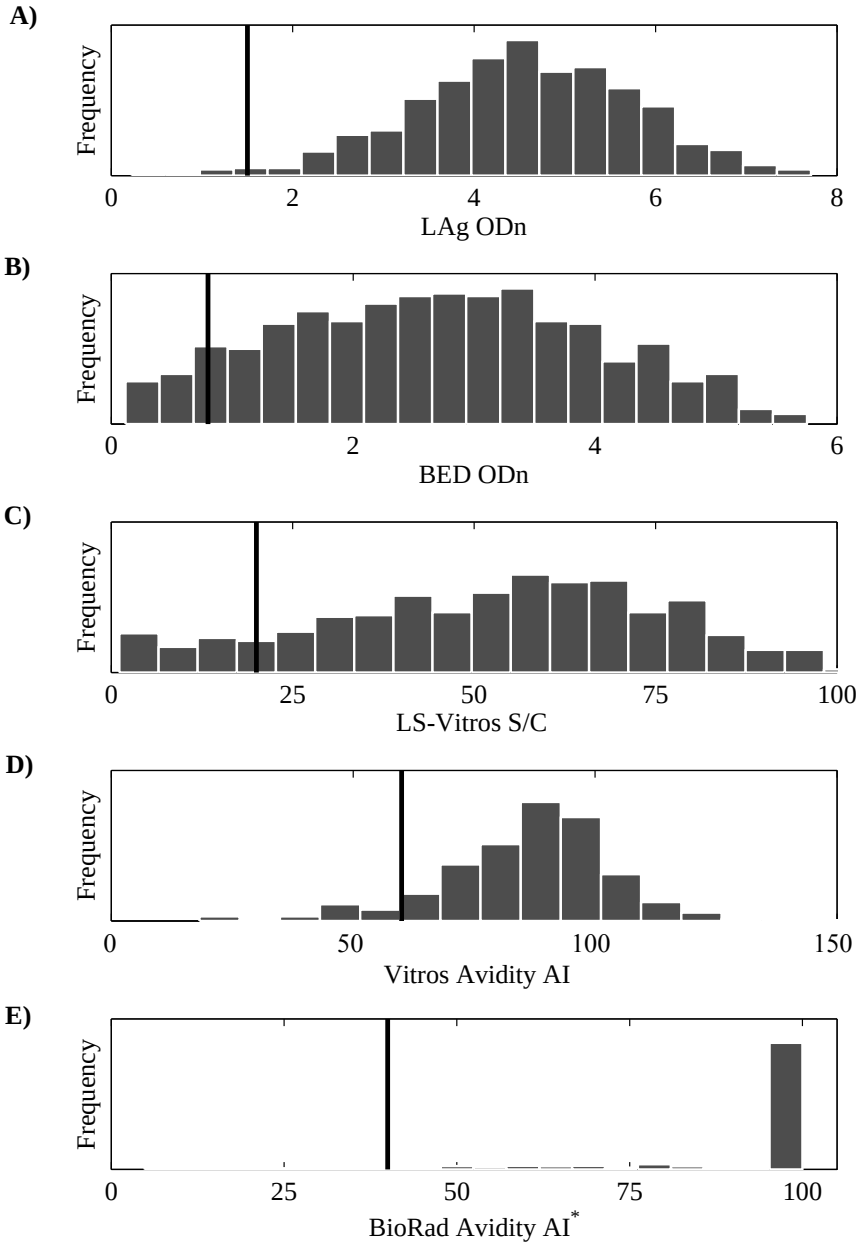


Figure 5.2: Distribution of incidence assay measurements for specimens from long-infected subjects for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity

The empirical distribution of incidence assay measurements for specimens drawn more than $T = 2$ years after infection, excluding treated subjects and identified elite controllers, is shown for A) LAg, B) BED, C) LS-Vitros, D) Vitros Avidity and E) BioRad Avidity (665 data points from 316 subjects). 'Recent'/'non-recent' thresholds are shown by vertical solid lines.

* The peak of BioRad Avidity results at 100% is due to a large proportion of (treated and untreated) aliquots returning the maximum possible S/C on the equipment used.

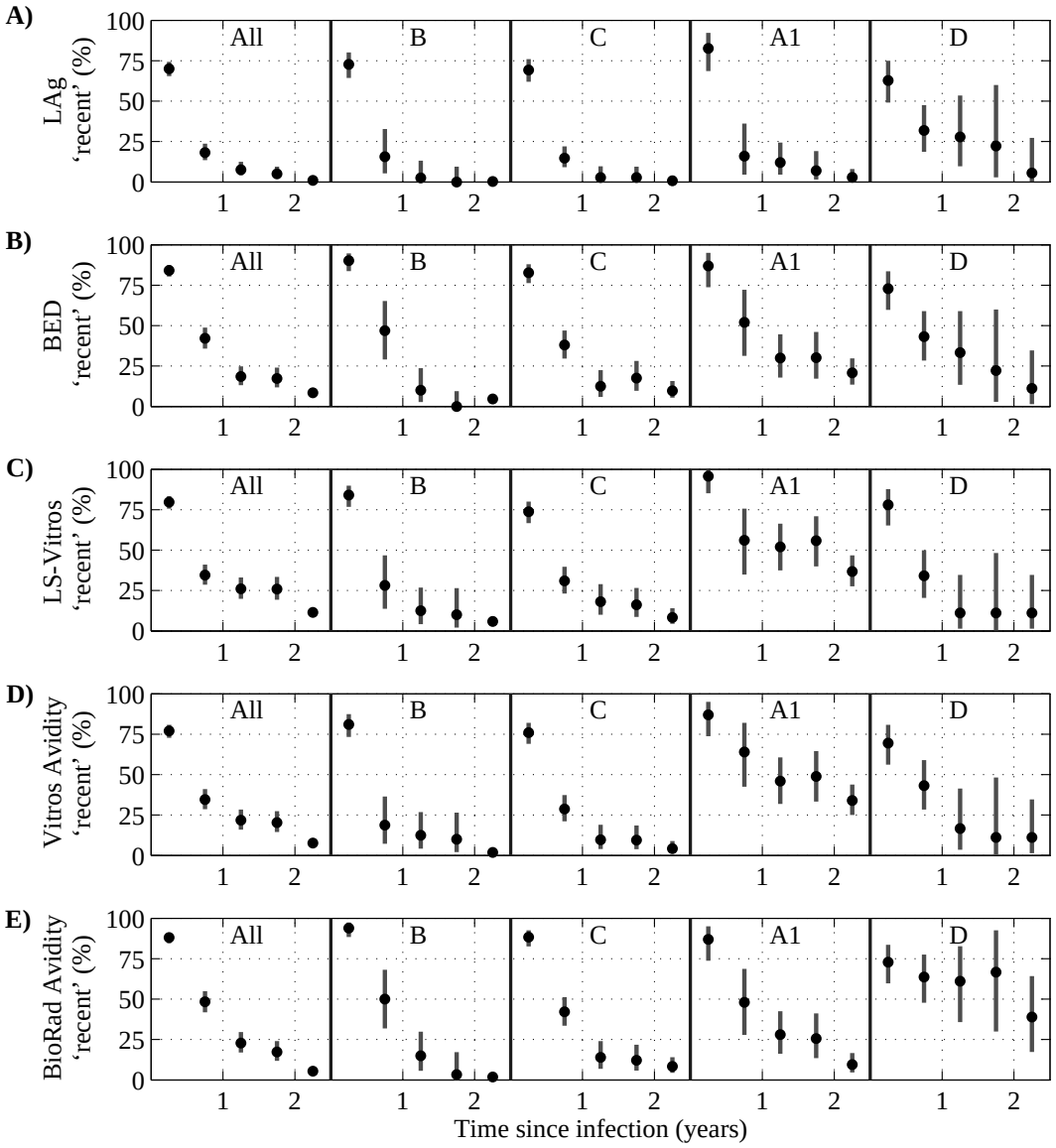


Figure 5.3: The proportion of 'recent' results stratified by time since infection and HIV subtype for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity

The proportion of 'recent' results (%) as a function of time since infection (years) and stratifying by HIV subtype (A1, B, C and D), excluding treated subjects and identified elite controllers, is shown for A) LAg, B) BED, C) LS-Vitros, D) Vitros Avidity and E) BioRad Avidity. Circles show observed proportions and lines capture 95% confidence intervals. Specimens are grouped by 6-monthly intervals of time since infection until 2 years, after which all specimens are grouped together (there are 25 to 665 data points per group, other than for subtype D, which has fewer than 20 points 1-2 years after infection).

Table 5.1 provides estimated test properties for the various specimen sets. LAg has an estimated MDRI of 188 days (95% CI: 165-211), while remaining assays have MDRI estimates of 285 to 333 days (the CI limits range from 254 to 363 days). Results were insensitive (less than a 2% change in results) to whether ARS onset dates were used to adjust estimated infection dates, a change to Data Exclusion Rule II, and the use of alternative Model C. MDRI estimates increased by 2% to 4% when changing to Model B, which was the most sensitive to the data exclusion rules (there was a 4% to 10% increase in estimates when changing to Data Exclusion Rule II).

Excluding treated subjects and SCOPE elite controllers, and analysing all remaining specimens drawn more than $T = 2$ years after infection, the measured FRR ranges from 1% (95% CI: 0.3%-3%) for LAg, to 6% to 10% (95% CIs span 3% to 14%) for the remaining assays.

When stratifying by time since infection, the varying persistence of 'recent' classifications across assays is evident, with LAg exhibiting the leanest tail of persistence of 'recent' infection.

The FRR among elite controller specimens is high for all assays, and averages 25% (minimum of 13% to a maximum of 48% across assays). The FRR among treated subjects is even higher, averaging 65% (minimum of 50% to a maximum of 76% across assays). Further stratifying treated subjects by time from infection to treatment initiation, the FRR decreases as the time to treatment initiation increases: for early treatment initiation (within 6 months of infection) the average FRR is 84% (64% to 93%), while for later treatment initiation (more than 6 months after infection) it is 41% (27% to 57%).

The FRR for subjects with low viral loads, here defined as below 75 copies/ml, is high, averaging 55% (41% to 69%). This is consistent with results above, as 92% of this specimen set is made up of specimens from the identified elite controllers and treated subjects (and 94% of specimens from SCOPE elite controllers and treated subjects have a low viral load).

Lastly, the FRR among subjects with low CD4 cell counts, namely less than 200 cells/µl and acting as a proxy for AIDS identification, was relatively low, averaging 2% (0% to 4%). Further stratifying this group by CD4 cell count (not shown) did not reveal any patterns.

	# Subjects (data points)	LAg	BED	LS-Vitros	Vitros Avidity	BioRad Avidity
'Recent'/'non-recent' threshold (unit)		1.5 (ODn)	0.8 (ODn)	20 (S/C)	60 (AI as %)	40 (AI as %)
MDRI (days)[i]						
All specimens[ii]	400 (1032)	188 (165-211)	302 (274-331)	306 (274-338)	285 (254-316)	333 (302-363)
FRR (%)[i]						
All specimens[ii]	316 (665)	1.3 (0.3-3.2)	7.4 (4.8-10.9)	9.7 (6.6-13.5)	6.5 (4.0-9.8)	6.2 (3.8-9.4)
By time since infection (years)[ii]						
(2,3]	140 (208)	2.5 (0.6-6.6)	12.5 (7.5-19.1)	17.5 (11.6-24.8)	12.5 (7.5-19.1)	12.5 (7.5-19.1)
(3,4]	77 (110)	0.6 (0.0-5.9)	7.1 (2.5-15.4)	14.9 (7.8-24.9)	14.3 (7.4-24.1)	6.5 (2.1-14.5)
(4,5]	35 (45)	0.0 (0.0-8.2)	7.1 (1.2-21.1)	5.7 (0.7-19.2)	5.7 (0.7-19.2)	8.6 (1.8-23.1)
>5	112 (193)	0.0 (0.0-2.6)	6.7 (2.8-13.0)	3.1 (0.8-8.3)	1.3 (0.1-5.6)	0.0 (0.0-2.6)
Elite controllers[iii]	31 (89)	12.9 (3.6-29.8)	19.4 (7.5-37.5)	48.4 (30.2-66.9)	29.0 (14.2-48.0)	12.9 (3.6-29.8)
Treated subjects[iv]	113 (185)	58.8 (49.2-68)	65.9 (56.4-74.6)	76.1 (67.2-83.6)	72.6 (63.4-80.5)	50.0 (40.4-59.6)
By time from infection to treatment (years)[iv]						
[0,0.5)	53 (90)	84.9 (72.4-93.3)	86.8 (74.7-94.5)	92.5 (81.8-97.9)	92.5 (81.8-97.9)	64.2 (49.8-76.9)
≥0.5	53 (88)	27.4 (16.0-41.3)	40.6 (27.3-54.9)	56.6 (42.3-70.2)	49.1 (35.1-63.2)	31.1 (19.1-45.3)
Low viral load[v]	154 (273)	47.1 (39.0-55.3)	56.5 (48.3-64.5)	68.5 (60.5-75.7)	62.7 (54.5-70.3)	40.6 (32.8-48.8)
Low CD4 cell count[vi]	124 (214)	0.0 (0.0-2.4)	4.0 (1.3-9.2)	2.4 (0.5-6.9)	0.0 (0.0-2.4)	1.6 (0.2-5.7)

[i] Using an HIV viral lysate-based Western blot assay to identify HIV-positive subjects, and $T = 2$ years

[ii] Excluding treated subjects and SCOPE elite controllers

[iii] Identified as elite controllers in the SCOPE cohort (virally suppressed in the absence of treatment)

[iv] No previous treatment interruptions and treated for at least 3 months

[v] Viral load at draw <75 copies/ml

[vi] CD4 cell count at draw <200 cells/μl

Table 5.1: Estimated test properties for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity for various specimen sets

Estimates of the mean duration of recent infection (days) and false-recent rate (%), and 95% confidence intervals, are provided for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity, for various sets of specimens

Table 5.2 lists MDRI and FRR estimates by subtype. The most significant pairwise differences in the MDRIs were between subtype A1 and any other, on the Vitros platform. With one exception, notably small p-values for subtype pairwise differences in the FRRs involve A1 or D and a non-A1, non-D subtype, dominated by LS-Vitros, Vitros Avidity and BioRad Avidity results. While these initial results highlight potential subtype differences, a more definitive analysis (beyond the present scope) should be based on a large number of subtype D and A1 specimens and estimation procedures specifically adapted to this stratification.

5.1.4 Discussion

The application of cross-sectional HIV incidence surveillance, utilising tests for recent infection, has been hampered by the lack of high performance incidence assays and the lack of independent, rigorous and consistent evaluations of candidate assays [9, 10, 12-14]. Over the last four years, CEPHIA [47] has developed a substantial repository of precious specimens, and begun using these specimens to characterise the most promising incidence assays. Results for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity are presented above.

Assays can be evaluated against a 'Target Product Profile' (TPP) [13, 14, 188]: Not only should the technology be affordable, practical and transferable to other laboratories, but the mean duration of recent infection (MDRI) should be sufficiently long (of around one year) and the false-recent rate (FRR) small (ideally zero, and less than 2%). Results suggest that incidence assays continue to struggle to simultaneously achieve these two test property goals, with no single assay unequivocally meeting the criteria set out in the TPP. Compared to the increasingly used LAg assay, the other assays provide larger MDRIs but also higher FRRs.

	# Subjects (data points)	LAg	BED	LS-Vitros	Vitros Avidity	BioRad Avidity
'Recent'/'non-recent' threshold (unit)		1.5 (ODn)	0.8 (ODn)	20 (S/C)	60 (AI as %)	40 (AI as %)
MDRI (days)[i,ii,iii]						
All specimens	400 (1032)	188 (165-211)	302 (274-331)	306 (274-338)	285 (254-316)	333 (302-363)
Subtype A1	80 (166)	211 (156-275)	363 (288-442)	473 (387-560)	451 (362-539)	364 (289-442)
Subtype B	90 (246)	153 (117-196)	255 (208-308)	232 (173-299)	210 (157-270)	299 (233-371)
Subtype C	181 (454)	177 (150-206)	287 (248-328)	265 (221-311)	240 (202-280)	298 (262-338)
Subtype D	38 (131)	273 (170-387)	328 (227-433)	264 (190-339)	276 (194-357)	467 (354-581)
FRR (%)[i,ii,iv]						
All specimens	316 (665)	1.3 (0.3-3.2)	7.4 (4.8-10.9)	9.7 (6.6-13.5)	6.5 (4.0-9.8)	6.2 (3.8-9.4)
Subtype A1	37 (106)	2.7 (0.1-14.2)	18.9 (8.0-35.2)	35.1 (20.2-52.5)	35.1 (20.2-52.5)	12.2 (3.8-27.1)
Subtype B	190 (388)	0.5 (0.0-2.9)	4.7 (2.2-8.8)	4.7 (2.2-8.8)	1.3 (0.2-4.2)	2.1 (0.6-5.3)
Subtype C	75 (144)	1.3 (0.0-7.2)	7.3 (2.6-15.7)	8.7 (3.4-17.5)	4.0 (0.8-11.2)	6.7 (2.2-14.9)
Subtype D	11 (18)	9.1 (0.2-41.3)	18.2 (2.3-51.8)	18.2 (2.3-51.8)	18.2 (2.3-51.8)	54.5 (23.4-83.3)

[i] Using an HIV viral lysate-based Western blot assay to identify HIV-positive subjects, and $T = 2$ years

[ii] Excluding treated subjects and SCOPE elite controllers

[iii] In a test for pairwise differences in MDRIs by subtype, using a z-test, the following pairs provided p-values below 0.05: LAg – B&D; BED – A1&B; LS-Vitros – A1&B, A1&C, A1&D; Vitros Avidity – A1&B, A1&C, A1&D; BioRad Avidity – B&D, C&D. Estimated standard deviations of the MDRI estimators were used as proxies for true values, and therefore tests were anticonservative (particularly when sample sizes were small).

[iv] In a test for pairwise differences in FRRs by subtype, using the Fisher-Boschloo test [211, 212] the following pairs provided p-values below 0.05: BED – A1&B; LS-Vitros – A1&B, A1&C; Vitros Avidity – A1&B, A1&C, B&D; BioRad Avidity – A1&B, A1&D, B&D, C&D.

Table 5.2: Estimated test properties for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity by HIV subtype

Estimates of the mean duration of recent infection (days) and false-recent rate (%), and 95% confidence intervals, are provided for LAg, BED, LS-Vitros, Vitros Avidity and BioRad Avidity, stratified by HIV subtype

While a stable, high-performance incidence assay should ideally produce a consistently small FRR, regardless of the study population, data from this work helps to understand some of the reasons why an assay's performance could be unstable and FRRs may be large. All assays produce particularly high FRRs among elite controllers (>10%) and treated subjects (>50%), and the size of these subpopulations will vary by region and time. In a surveillance study, identifying these subjects is problematic, as there is no universal definition of, or test for, elite controllers, and self-reported treatment status may be unreliable. Furthermore, earlier initiation of treatment is associated with a higher FRR, in line with varying impacts of treatment on immune responses by treatment timing [72, 78, 213]. Context strongly affects when patients begin treatment – for example, in some states in the USA, patients are offered treatment immediately following HIV diagnosis [214], while in South Africa most HIV-positive patients are unable to access treatment until CD4 cell counts drop below 350 copies/µl [215]. In this study, 94% of specimens from elite controllers and treated subjects also had a low viral load (<75 copies/ml), and so viral load testing provides a potential tool to screen for these high-FRR subjects – specimens with viral loads below an optimised threshold would be classified as 'non-recent' (regardless of the incidence assay measurement). Note that such a change in the 'recent' infection classification rule will also impact (reduce) the MDRI. Surveys could also directly test for the presence of antiretroviral drugs to identify treated subjects [216].

Properties for each assay have been estimated here on the standardised basis of a Western blot being used to identify HIV-positive subjects. However, other diagnostic screening tests are likely to be used in incidence studies, and the time between HIV exposure and reactivity on these tests can differ by several weeks [180, 181, 185]. Therefore, for application to incidence studies, the base case MDRI reported here would need to be *increased or decreased* – depending on the particular screening test or algorithm used in the study to classify a specimen as HIV-positive, and hence eligible for 'recent' infection testing.

The results presented here should not be viewed as discouraging, as they provide a consistent, independent characterisation of these candidate incidence assays. Large FRRs continue to limit the utility of single incidence assays, and subtype-specific test behaviour should be further explored. This analysis provides the basis for exploring optimisation through such adjustments as variation of 'recent'/'non-recent' thresholds, inclusion of supplemental tests (in particular, viral load), and the use of multiple incidence assays, all of which is the subject of ongoing work within and beyond CEPHIA [10, 12, 13, 86, 88,

101, 102]. Optimisation should also consider the time cut-off T, to distinguish 'true-recent' from 'false-recent' results. Although T should be not be too large, the value of T was increased from 1 year, as used in preliminary analyses [42], to 2 years in this analysis, to better capture the tails of persisting 'recent' results and thus reduce FRRs. Ongoing analyses also include the evaluation of tests for recent infection, using the precision of the incidence estimator as a summary performance metric [30].[40] In addition, efforts are being made to capture more detailed information on cohorts' diagnostic testing protocols and more complete testing histories of subjects, thus providing the required data to further refine estimated infection dates for later analyses of assay results.

The repository of specimens and data that has been generated by CEPHIA provide a unique opportunity to further advance the investigation and refinement of markers of 'recent' HIV infection. Specimens and datasets are well-maintained, samples sizes are large, specimen background information is recorded, and multiple incidence assays and potential supplemental tests have been applied to the same specimens.

CEPHIA has begun testing the 'evaluation panel' using other assays, with the aim of evaluating ten incidence assays in its first phase. A second phase of CEPHIA, known as CEPHIA II and also funded by the *Bill and Melinda Gates Foundation* (BMGF), was launched in the beginning of 2013. Under CEPHIA II, the repository is being expanded to include non-plasma specimens (such as linked whole blood, oral fluid, urine and stool) that are being prospectively collected through collaborations with various study sites. CEPHIA is also supporting biomarker discovery projects funded by the BMGF and *US National Institute of Health* (NIH), with a focus on earlier steps in the development pathway. Further updates on CEPHIA activities can be found on the project website (http://www.incidence-estimation.com/page/cephia).

[40] The use of the precision of the incidence estimator, as a standard metric for assessing test performance, was formally outlined in earlier work, and is presented in Chapter 6.

Chapter 6

Measuring and Optimising Test Performance

In principle, any test for recent infection can be characterised and applied in an incidence surveillance study. Earlier chapters of this work have both developed the theoretical framework and explored the practical methodology for doing this. However, the lack of standard metrics for assessing the utility of recent infection tests has been an obstacle to test development, and therefore the measurement and optimisation of test performance are the focus of this final contribution of the thesis.

While it has been increasingly recognised that tests should provide an enduring state of 'recent' infection and rarely (ideally, never) return 'recent' results at large times post infection, formal metrics for trading off these two test properties have not been widely adopted. An appropriate metric for assessing and optimising tests for recent infection, namely the precision of the incidence estimator, is presented in Section 6.1, which is a reproduction of a published article providing guidance to developers [30].[41]

As the discourse increasingly moves towards practical application, efforts are being made to develop high-performance biomarkers for recent infection. Some important practical considerations are highlighted in Section 6.2, which explores the scope of the test optimisation, the context-dependence of test performance, and other test design criteria of relevance.

---

[41] The contents of Section 6.1 have been published as: 'Kassanjee R, McWalter TA, Welte A. Defining optimality of a test for recent infection for HIV incidence surveillance. *AIDS Res Hum Retroviruses*. 2014; 30(1): 45-49'. The article was reproduced with permission from *AIDS Research and Human Retroviruses*, published by *Mary Ann Liebert, Inc.,* New Rochelle, NY. The manuscript was primarily written by RK, who also performed all analyses. AW and TAM helped conceive the ideas, reviewed the analyses, and assisted in writing the manuscript.

6.1 Defining Optimality of a Test for Recent Infection for HIV Incidence Surveillance

The measurement of HIV incidence, the rate of new infections, is essential in most surveillance and intervention contexts. Recognising the practical challenges presented by longitudinal studies, the estimation of incidence from cross-sectional surveys using tests for recent infection has attracted considerable interest [8-11, 14, 147, 217]. However, the performance and optimisation of a test that aims to categorise infections as 'recent' or 'non-recent', specifically for *population-level surveillance,* requires a shift from conventional diagnostic thinking about test performance.

When *individual-level detection* of a condition is of primary interest, sensitivity, specificity and predictive values are appropriate metrics of performance. These metrics improve as inter-subject variability decreases. However, when estimating a *population-level summary parameter*, such as incidence, the appropriate performance metrics are accuracy and precision of the statistic measured. Here, incidence estimation utilises information on the *average* behaviour of biomarkers, and is relatively insensitive to the variability underlying this averaging. While the appropriate optimisation of tests for recent infection has been noted in passing [10, 11, 14, 147, 217], there is neither consensus nor guidance for developers.

As with any diagnostic, elements of a test for recent infection may be adjusted to alter its performance. In the context of HIV recent infection tests, typically some quantitative host or viral biomarkers are measured, and the infection is categorised as 'recent' or 'non-recent' by reference to thresholds [8-10]. For example, the widely used BED assay measures the proportion of HIV-specific immunoglobulin G (IgG) antibodies in total IgG, and a measurement below some threshold classifies the infection as 'recent' [107]. While a test may be comprised of many elements that can be varied, from the underlying biological processes measured to the assay platforms and specific kits, ultimately, the optimisation will involve the fine tuning of thresholds.

It is increasingly recognised that the lack of high performance recent infection tests poses a major obstacle to the widespread implementation of cross-sectional incidence surveillance [14, 217]. The *World Health Organisation* (WHO) has maintained a *WHO*

Working Group on HIV Incidence Assays since 2006, the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (*CEPHIA*) was established in 2010, and both the *Bill and Melinda Gates Foundation* and the *National Institutes of Health* have provided substantial funding for the development of better tests [13, 47, 52, 188, 218]. Given the current surge in the development of candidate tests for recent infection, it is important to have clarity and consensus on robust metrics of performance, and in particular to avoid the pitfalls of traditional diagnostic thinking.

Prevalence, the fraction of a population with a condition, can at times be substantially informative about incidence. For example, for transient conditions, such as influenza, it is well known that near demographic equilibrium:

$$\text{Incidence} \approx \frac{\text{Prevalence}}{\text{Mean duration of condition}}, \tag{6.1}$$

where incidence is expressed as a rate of cases per person time in the entire population, not just per person time at risk. However, when a condition is enduring, and survival in the state is poorly known and evolving, as is the case with HIV, prevalence becomes uninformative about incidence. In this case, it makes sense to find ways of defining and detecting a robust early phase post infection, and using a more refined version of the above heuristic to infer incidence from the prevalence of 'recent' infection.

Under simplistic assumptions, HIV incidence, expressed as a rate of infection per person time at risk, is then formally estimated, in a cross-sectional setting, by [7]:

$$\hat{I} = \frac{p_R}{p_S \Omega}, \tag{6.2}$$

where p_R and p_S are the proportions of 'recently' infected and susceptible or HIV-negative subjects in the sample, and Ω is the mean duration of recent infection. Currently available tests (and perhaps all conceivable tests) for recent infection present a subtle problem in that some individuals who have been infected for long periods of time may nevertheless yield spurious 'recent' results [77, 78, 213]. With some simplifying assumptions, it has been shown how this 'false-recent' phenomenon can be intuitively understood as requiring a 'subtraction' of the estimated number of 'false-recent' results from the observed number of 'recent' results [16, 17, 25, 29].

More recently, a very general analysis has been obtained by introducing a convenience time cut-off T, which represents the time, post infection, after which a 'recent' test result is a 'false-recent' result [29]. The test properties then are (i) a false-recent rate (FRR), β_T,

which is the (population-dependent) proportion of those individuals infected for more than time T who produce 'recent' test results, and (ii) a somewhat subtly-defined mean duration of recent infection (MDRI), Ω_T, which is the average time spent 'recently' infected while infected for less than T [29].[42] Note that $1-\beta_T$ is the (population-dependent) specificity of the test *if* it aimed to identify infections that have occurred within the preceding period T. This leads to the following incidence estimator [29]:

$$\hat{I}_T = \frac{p_R - \beta_T p_+}{p_S \cdot (\Omega_T - \beta_T T)}, \tag{6.3}$$

which depends on the proportions of subjects in the sample who are classified as 'recently' infected, HIV-positive and HIV-negative, denoted by p_R, p_+ and $p_S = 1 - p_+$, respectively; and the test properties, Ω_T and β_T, for a chosen time cut-off T. When there are no 'false-recent' results ($\beta_T = 0$), Equation (6.3) reduces to Equation (6.2), noting that $\Omega = \Omega_T$ in the absence of an explicit time cut-off T. In terms of epidemiological and demographic context, the applicability of Equation (6.3) requires only that the susceptible population size does not vary substantially over a period of duration T [29]. The biomarkers underlying the test for recent infection should mainly capture stable biological, as opposed to environmentally-dependent, factors over the period T post infection (that is, the MDRI should not vary significantly by context, but be a true property of the test) [29]. It is understood that the FRR will have contextual variability.

Uncertainty in the incidence estimate arises from statistical fluctuations of the proportions of subjects in the sample in the various classes as well as from uncertain test properties. The uncertainty of the incidence estimator, described here by its coefficient of variation (ratio of standard deviation to mean), c, can be approximated using the delta method [29]:

$$\begin{aligned} c^2 = {} & \frac{1}{nP_+} \cdot \left(\frac{1}{P_S} + \frac{P_R P_{NR}}{(P_R - \beta_T P_+)^2} \right) \\ & + \sigma^2_{\hat{\Omega}_T} \cdot \left(\frac{1}{\Omega_T - \beta_T T} \right)^2 \\ & + \sigma^2_{\hat{\beta}_T} \cdot \left(\frac{P_{NR}\Omega_T - P_R \cdot (T - \Omega_T)}{(P_R - \beta_T P_+) \cdot (\Omega_T - \beta_T T)} \right)^2, \end{aligned} \tag{6.4}$$

[42]The general framework for incidence inference, producing these definitions of the MDRI and FRR, was developed as part of earlier work, and is presented in Chapter 2.

where P_R and P_{NR} are the proportions of 'recently' and 'non-recently' infected individuals in the study population; $P_+ = P_R + P_{NR}$ and $P_S = 1 - P_+$ are the proportions of infected (or HIV-positive) and susceptible (or HIV-negative) individuals in the population; and $\sigma_{\widehat{\Omega}_T}$ and $\sigma_{\widehat{\beta}_T}$ are the uncertainties (standard deviations) with which the test properties Ω_T and β_T are measured. It is certainly possible that the normality assumptions intrinsic to deriving Equation (6.4) could be violated in practice, in which case the same underlying theory can be used as a basis for a numerically more complex calculation of the variance of incidence estimates, such as by bootstrap resampling methods [174]. Whether or not one uses Equation (6.4) to estimate the coefficient of variation is not fundamental to the present discussion about optimisation.

The familiar statistical objective when estimating any parameter is to find a sample statistic that estimates the parameter with the greatest accuracy (least bias) and greatest precision (smallest variance). As shown in the general derivation of Equation (6.3) [29], the bias is negligible compared to variance in the epidemiologically and demographically relevant regime, using any reasonable test for recent infection. The remaining goal is therefore the minimisation of variance. The apparent bias reported by other researchers [27] is a result of an alternative summary parameterisation of biomarker dynamics, which declares the FRR to be zero. This leads to an MDRI which is complex, context-dependent and difficult to estimate; and hence produces a similarly context-dependent implicit weighting over historical incidence.

To minimise variance, the state of 'recent' infection should not be too transient, so that a realistically sized cross-sectional survey can capture a sufficient number of 'recent' cases for the estimation of the 'recent' proportion to be statistically robust (and therefore the MDRI should be large). The larger the adjustment for 'false-recent' results, the greater the overall uncertainty arising from fluctuations of the sample proportions (and therefore the FRR should be small). However, the two test properties cannot be independently adjusted, as test modifications that increase the MDRI typically also increase the FRR. Hence, the central goal of test design is an optimal balance between these two properties. An ideal test would have a near-zero FRR and an MDRI of around a year, if considering a setting where one is interested in the average incidence over approximately the last year.

Figure 6.1 illustrates this trade-off between the MDRI, Ω_T, and FRR, β_T. The contours show the coefficient of variation (CoV) of the incidence estimator as a function of the MDRI and FRR, in an example context. In the context, HIV incidence is 1% per annum

and HIV prevalence is 10%, Ω_T and β_T are estimated with a 5% and 30% CoV respectively and $T = 1$ year, and incidence is measured in a cross-sectional survey of 5 000 subjects. Moving to the right (to a large MDRI) and down (to a low FRR) in the contour plot, the CoV of the incidence estimator decreases (that is, precision increases). For a test to begin to move into a regime of usefulness, the MDRI should be at least 6 months and the FRR below 2% [11, 14, 188]. In the context described by the figure, this implies a CoV of the incidence estimator of 30%, which implies that one can be 95% confident of estimating the true incidence of 1% per annum as a point estimate between 0.4% and 1.6% per annum.

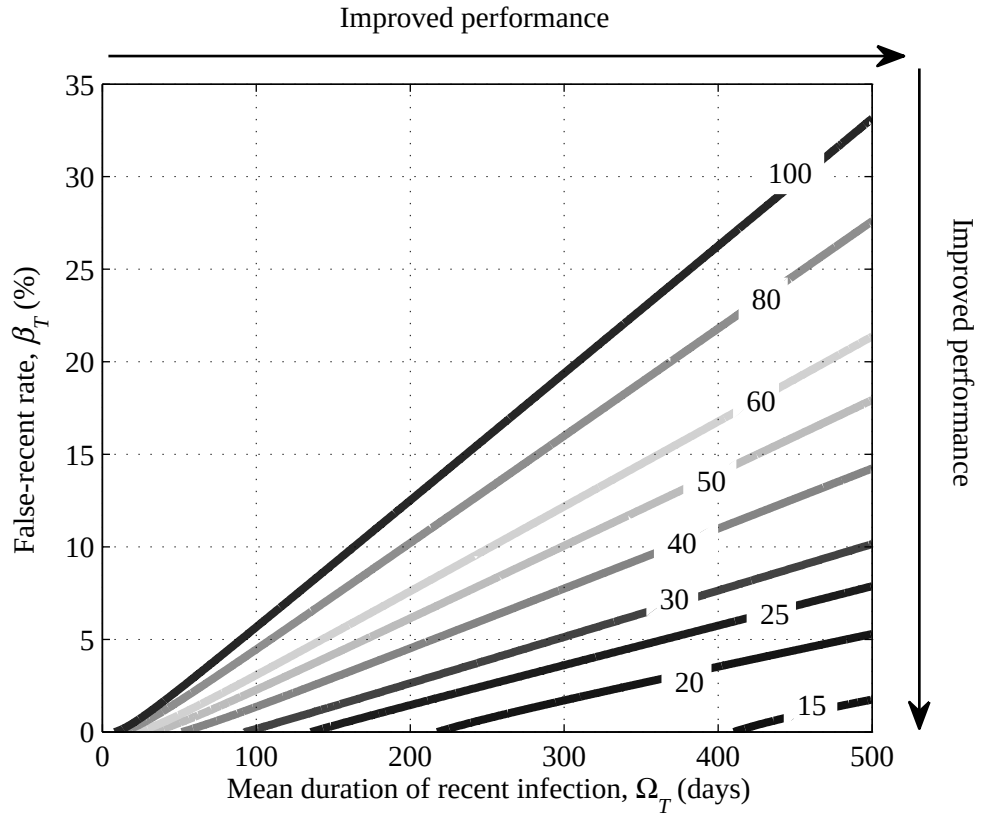


Figure 6.1: Performance of a test for recent infection as a function of test properties, in an example context

The contour plot captures the coefficient of variation (CoV) of the incidence estimator (%), as a function of the mean duration of recent infection, Ω_T (days), and false-recent rate, β_T (%), for an example context. HIV incidence is 1% per annum and HIV prevalence is 10%, Ω_T and β_T are estimated with a 5% and 30% CoV respectively and $T = 1$ year, and incidence is measured in a cross-sectional survey of 5 000 subjects.

To illustrate the optimisation of test design in a simplistic scenario, consider a biomarker where the reading at time t after infection (years) is

$$y(t) = 1 - \exp(-2t) + 0.2 \cdot \varepsilon, \tag{6.5}$$

where ε captures noise and is a standard normal random variable, and readings below some chosen threshold indicate 'recent' infection. Here, the only source of variability is noise in the biomarker, while in reality there is often substantial inter-subject variability, and effects due to immune system decline and treatment. Nevertheless, this simple model captures the same trade-off – that increasing the threshold increases both the MDRI and FRR – observed with more complex biomarker dynamics. In Figure 6.2 the precision of the incidence estimator is plotted as a function of test threshold, and the optimal threshold indicated, for an example context. In the context, there is epidemiological and demographic equilibrium of 1% per annum HIV incidence and 10% HIV prevalence, Ω_T and β_T are exactly known, $T = 1$ year, incidence is measured in a survey of 10 000 subjects, and all individuals survive for (exactly) 10 years from infection. Note that the FRR, β_T, depends on the survival dynamics and epidemiological and demographic history of the population (up until the maximum post-infection survival time). A tool to calculate the CoV of the incidence estimator is available at http://www.incidence-estimation.com/page/tools.[43]

Unfortunately, there is no single test design that will be optimal in all settings. This is because uncertainty in incidence estimation is determined by both the dynamics of the recent infection test and the context-specific epidemiological and demographic history (captured by HIV incidence and prevalence). Therefore, a range of anticipated contexts should be considered in evaluating a candidate test, or in fine tuning test design. This context-specific performance may be discouraging and regrettably complicated, but it is not unique to this surveillance application: even in a conventional simplistic diagnostics setting, the sensitivity and specificity of a test, if these can be assumed to be context-independent, must be combined with a contextual prevalence to determine the predictive value performance of the test.

The minimisation of the variance of the incidence estimator, or maximisation of precision, by trading the MDRI off against the FRR, provides a completely general

[43]The *Test Performance Calculator* is part of the online ABIE v2.0 tool suite, and is described in Section 2.3 and Appendix A.

criterion for optimising test design, regardless of the complexity of the test. For example, there is a trend towards using multiple biomarkers in a single test for recent infection, where various approaches for combining the individual biomarker results to produce a classification could be employed. For example, the final classification could be based on the sum of biomarkers readings, or on the number of individual readings below biomarker-specific thresholds [86, 88, 95, 101, 102]. The optimal test design is that which provides the lowest variance of incidence estimates across intended contexts.

Obtaining the most precise incidence estimates also consistently captures the optimisation that would be appropriate in studies that aim to test for differences in incidence or identify risk factors for HIV acquisition. Statistical tests for differences among groups (for example, capturing different ages, genders or social and sexual behaviours) are more highly powered when incidence is more precisely estimated in each group.

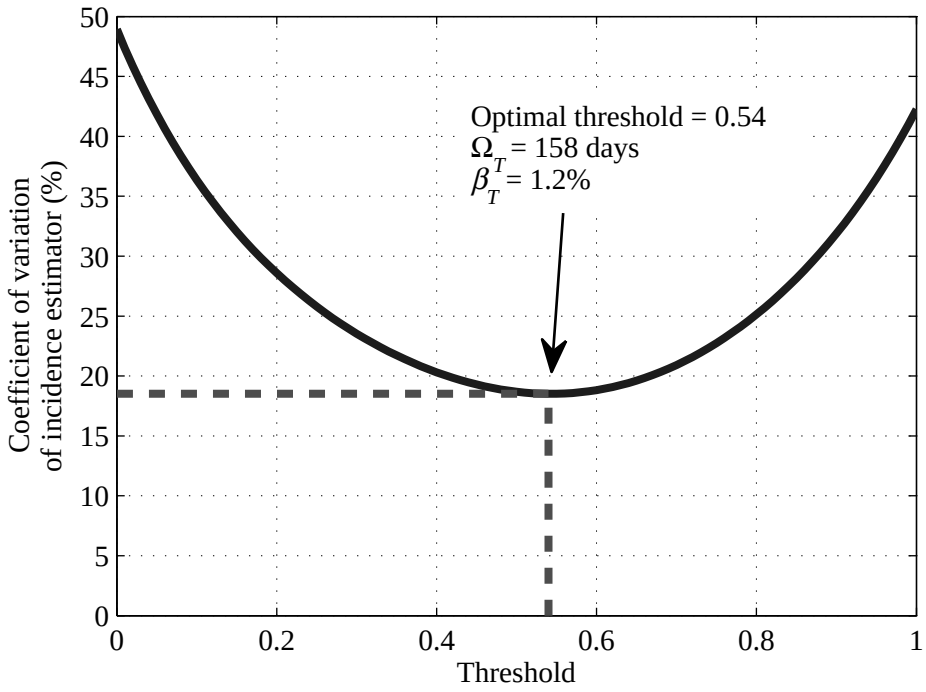


Figure 6.2: Optimal threshold for a hypothetical biomarker for recent infection, for an example context

The coefficient of variation (CoV) of the incidence estimator (%) is shown, as a function of the threshold used to distinguish between 'recent' and 'non-recent' infection, for a hypothetical biomarker and example context. A biomarker measurement at time t (years) after infection is given by $1 - \exp(-2t) + 0.2 \cdot \varepsilon$, where ε is a standard normal random variable. HIV incidence and prevalence have remained at 1% per annum and 10% respectively, Ω_T and β_T are exactly known, $T = 1$ year, incidence is measured in a cross-sectional survey of 10 000 subjects, and post-infection survival is 10 years.

Much of the literature introducing new tests for recent infection has attempted to assess their utility in terms of sensitivity and specificity [92, 109, 111, 112, 125, 187, 219]. As would be appropriate in the more familiar diagnostic applications, values close to 100% have been regarded as realistic targets, with these two measures summarised into, for example, Receiver Operating Characteristic (ROC) curves and the overall classification accuracy. However, three major obstacles are encountered when extending the use of these diagnostic metrics to this surveillance application. (i) Any workable definition of sensitivity and specificity requires a notion of truly recent infection that is defined by a strict threshold on time since infection, and a fully specified distribution of times since infection in a population. (ii) Even if thus defined, sensitivity and specificity cannot be accurately estimated from interval censored seroconverter data sets. (iii) Inter-subject variability of infection-related biomarkers naturally increase with time post infection; and therefore diagnostic optimisation will tend to motivate for a category of 'recent' infection restricted to the lower variability period close to infection, whereas incidence estimation requires the most enduring notion of 'recent' infection which does not bring a substantial false-recent rate.

In a clinical setting there may be substantial value in having some evidence of time since infection, at the time of HIV diagnosis. This opens up a multitude of new questions beyond the scope of this discussion. Most importantly, there needs to be further work to support reporting and interpreting individual biomarker values, beyond a 'recent' or 'non-recent' categorical result, and the appropriate optimisation of a test may not be the fine tuning of a threshold for this clinical context.

Any evaluation and optimisation of a test for recent infection should be based on the specific purpose for which the test is to be used, with the current work focusing on incidence estimation. The goal of HIV incidence estimation from cross-sectional surveys, using tests for recent infection, has attracted the interest of test developers. However, the assessment and optimisation of these tests, for purposes of estimating a population-level average, requires a fundamental shift from traditional criteria for measuring performance. The relevant performance metric of such tests is the precision of incidence estimates produced in an intended context. The central goal of the test developer, then, is the minimisation of the variance of the incidence estimator through a trade-off between the mean duration of recent infection and false-recent rate.

6.2 Some Important Practical Considerations for Test Optimisation

Following the development of the theoretical and methodological foundations for characterising and applying tests for recent infection, discourse in the field has shifted towards the practical application of biomarker-based incidence surveillance. Topics of growing interest are therefore the identification and optimisation of the most promising tests, and the development of improved tests. In support of these efforts, a single summary measure of test performance was introduced in the previous section. This metric, namely the precision of the incidence estimator, provides a formal framework for balancing the needs for a large mean duration of recent infection (MDRI) and small false-recent (FRR). When utilising this metric in practice, a number of important considerations arise. The scope of the optimisation, the context-dependence of test performance, and other criteria for assessing tests are briefly discussed below.

Defining the scope of the test optimisation

When optimising the design of a test for recent infection, one needs to choose which rules or parameters capturing the test design may be varied and how. While it is important to allow for sufficient flexibility to achieve high test performance, this goal needs to be balanced with practical limitations. For one, tests based on very complex schemes may be difficult to understand and consistently apply in surveillance studies. Also, if there are many parameters that can be tuned, an exhaustive exploration of possible test designs may become computationally daunting and the test design becomes vulnerable to being over-fitted to the data at hand. Some aspects to consider when defining the scope of the test optimisation are therefore briefly explored.

For an incidence assay consisting of a single quantitative measurement that increases (or decreases) over time after infection, one may reasonably restrict the test design to rely on only one parameter – namely a threshold for distinguishing between ‘recent’ and ‘non-recent’ infections. The optimisation of the test is then straightforward, and entails finding the value of the threshold that maximises the precision of incidence estimates. However, the uncertain or poor performance of existing incidence assays has led to increasing

interest in tests for recent infection based on multiple markers, in which case the optimisation search space can become high-dimensional, large and complex.

To illustrate this expansion of the search space, Figure 6.3 shows some possible test designs when using just two quantitative biomarkers, each increasing over time after infection. A test design specifies how to map the two quantitative readings onto dichotomous 'recent' and 'non-recent' classifications, and relies on a number of tuneable parameters. The optimisation procedure would need to select both an approach for mapping the measured markers onto classifications and the particular parameter values.

More generally, for m biomarkers (quantitative or otherwise), a region (possibly made up of a number of disjoint regions) in the m-dimensional space describing the m results needs to be defined as representing 'recent' infection. The boundary of this region could be described by an arbitrarily large number of parameters, and should be optimised so that individuals remain within it for sufficiently enduring times after infection on average (large MDRI), but are rarely still within it once infected for some time, such as a couple of years (small FRR). Some examples of test designs for multiple biomarkers can be found in the literature: proposed 'multi-assay algorithms', which utilise a number of incidence assays and measures of viral load and CD4 cell counts, classify infections as 'recently' acquired if each biomarker is below or above some biomarker-specific threshold [13, 61, 88, 89, 95, 101, 126]. Generalising this design, 'recent' infection may be identified by at least a specified number of all biomarkers being below (or above) their respective thresholds, as applied in the interpretation of the multiple immune responses produced by the Bio-Plex platform [102].

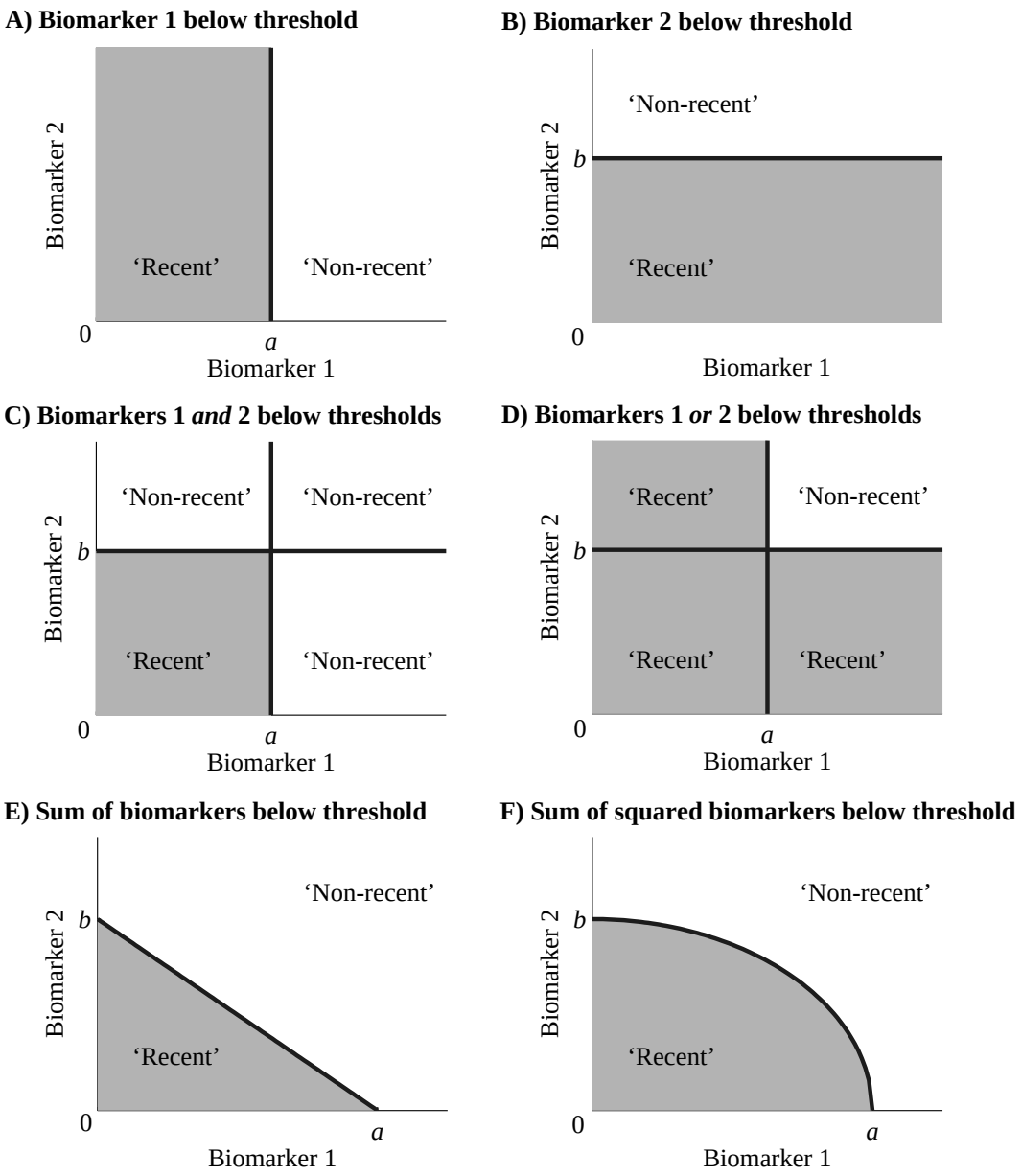


Figure 6.3: Examples of test designs for two biomarkers that increase over time after infection

Six possible test designs are shown for creating 'recent' and 'non-recent' infection classifications from two (non-negative) quantitative biomarkers (Biomarkers 1 and 2) that increase over time after infection. Each mapping of measurements onto classifications relies on at most two parameters, a and b. In A), a Biomarker 1 measurement below a produces a 'recent' result, while in B), a Biomarker 2 measurement below b produces a 'recent' result. In C), *both* biomarker measurements must be below their corresponding thresholds for a 'recent' classification, while in D), *at least one* biomarker measurement must be below its corresponding threshold. In E), a linear combination of the readings below a chosen threshold indicates 'recent' infection (summarised here by specifying the two intercepts of the dividing line shown). Lastly, in F), a linear combination of the squares of the two readings below a chosen threshold indicates 'recent' infection (summarised here by the two intercepts of the dividing elliptical boundary).

It is worth noting that some familiar tests for recent infection are based on quantities that are themselves already summaries of multiple biomarkers. Such a test therefore captures a specific mapping of multi-dimensional space onto a one-dimensional quantitative metric, which is then used to classify infections based on a single threshold parameter. For example, this is the case for antibody avidity assays which measure the ratio of antibody signal in a sample treated with a chaotropic agent to that in an untreated sample [100, 113, 115, 116, 119, 121]. A ratio, or 'avidity index', below a chosen threshold indicates 'recent' infection. The interpretation of the treated and untreated sample antibody signals, as a region in two-dimensional space representing 'recent' infection, is shown in Figure 6.4. Also, genetic diversity assays provide summary measures of very large amounts of complex data that collectively describe virus heterogeneity – for example, in the work of Park et al [125], the assay reading is the lower 10^{th} percentile of the Hamming distance distribution of sequences; and in the work of Cousins et al [220], a High Resolution Melting (HRM) score captures the melting peak width.

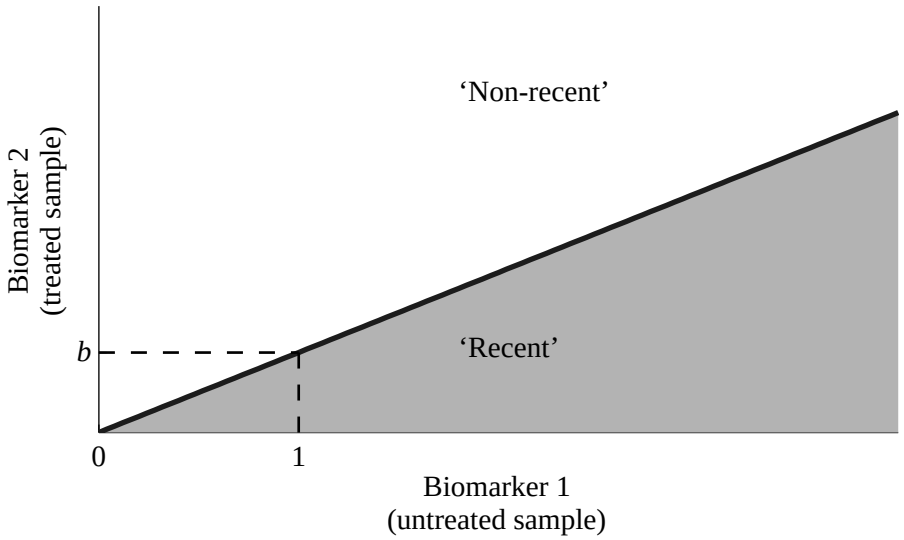


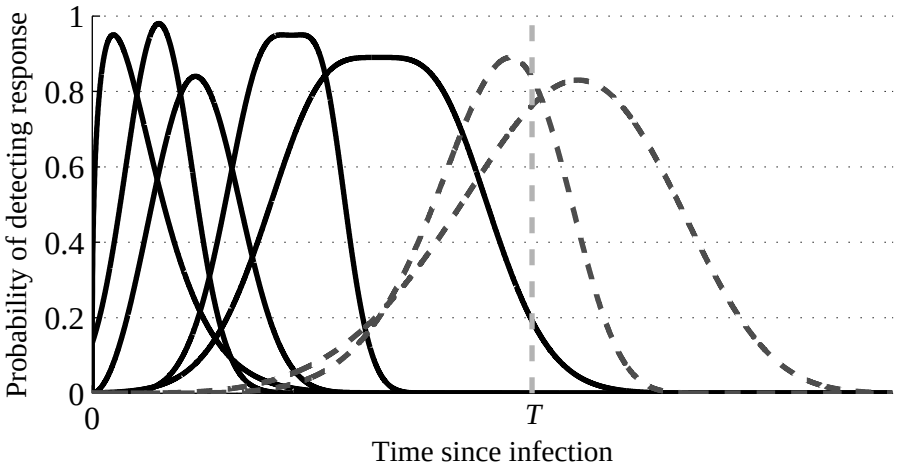
Figure 6.4: The 'recent' infection region implied by a conventional 'avidity index' interpretation of two antibody signals
A conventional avidity assay produces an 'avidity index', which is the ratio of antibody signal in a sample treated by a chaotropic agent (Biomarker 2) to that in a sample not thus treated (Biomarker 1), and a value of the avidity index below a chosen threshold b is interpreted as indicating 'recent' infection. The figure shows the 'recent' and 'non-recent' infection regions of the two-dimensional space that describes the two biomarker measurements.

Importantly, an understanding of the biomarker dynamics and the relationship among biomarkers should be used to inform which test designs should be considered in the optimisation. For example, in the extreme case that one biomarker can be expressed as a monotonic transformation of another, there can be no benefit from using both biomarkers. In the case that the two biomarkers are sufficiently different (uncorrelated or orthogonal), the performance of a test utilising both biomarkers will depend on the test design and the relationship between the biomarkers.

Continuing this example, first consider a test based on one biomarker, where a measurement below threshold a is interpreted as indicating 'recent' infection (Part A in Figure 6.3). Then stipulating that the second biomarker measurement must also be below threshold b for the 'recent' classification to be retained (Part C in Figure 6.3) would decrease both the MDRI and FRR (as the size of the 'recent' infection region is reduced). This test design would be best suited to two biomarkers that are closely related soon after infection (little impact on the MDRI), but poorly associated at large times post infection (large reduction in the FRR). Alternatively, stipulating that an infection is considered to be 'recent' if the first biomarker measurement is below a *or* the additional second biomarker measurement is below b (Part D in Figure 6.3) would increase both the MDRI and FRR. This design would be suited to biomarkers that are independent soon after infection (large increase in the MDRI) but highly related at large times post infection (small increase in the FRR). In practice, the optimal choice of threshold for the first biomarker is likely to change once the second biomarker is introduced.

As another example of how knowledge of individual biomarker dynamics can be used to select a plausible test design, consider a set of biomarkers where each aims to detect a different type of host immune response (capturing a particular type of antibody responding to a specific component of the virus). These responses may be transient, each growing and waning over a short and different time after infection, as captured in Figure 6.5 (Part A). By identifying which of these responses occur within a year or two after infection, and which occur later, it may be possible to construct a rule that provides a useful definition of 'recent' infection. A suitably enduring state of 'recent' infection could be obtained by requiring the detection of at least one early response, while the FRR could be kept low by also requiring that none of the later responses are detectable (Part B of Figure 6.5).

A) Probability of detecting immune response



B) Probability of 'recent' infection

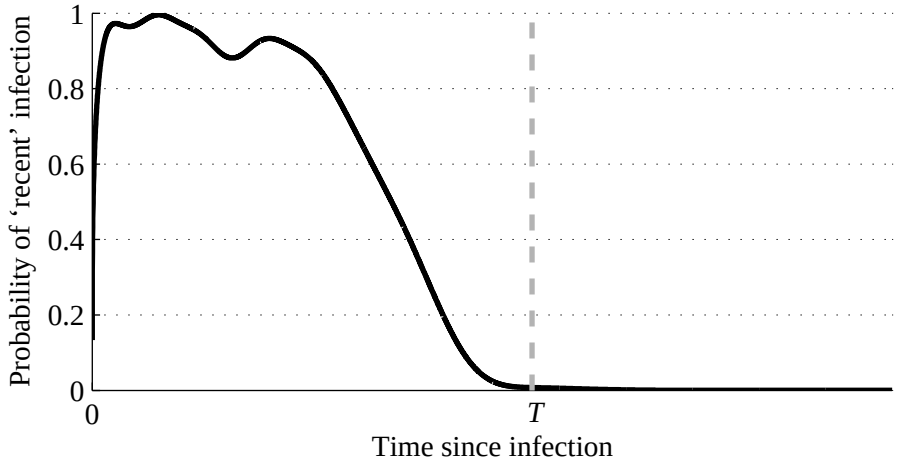


Figure 6.5: The dynamic of a hypothetical test for recent infection based on the detection of transient immune responses observed at different times after infection
The dynamic of a hypothetical test for recent infection, based on the detection of each of seven transient responses, is shown. In A), the probability of detecting each response is shown as a function of time since infection. In B), the dynamic of the test for recent infection is summarised by the probability of testing 'recently' infected as a function of time since infection. A 'recent' classification is produced if at least one of the five earlier responses (solid lines in Part A) is observed, and neither of the two later responses (dashed lines) is observed.

When considering a large number of complex biomarkers that are not well understood, more data-driven approaches may be required to inform test designs. Formal supervised learning methods may be considered, such as support vector machines, discriminant analysis, classification trees and generalised additive models. These machine learning techniques are used to find predictive rules – in this case, to classify an infection as

probably recently or non-recently acquired, based on the individual biomarkers. In addition to being critical of assumptions and the interpretability of the models, it is essential that the methods for fitting these models are appropriately adapted to this surveillance application.

Supervised learning approaches conventionally seek model parameters that maximise the probability of a 'correct' classification. This frames the optimisation in terms of metrics that are useful for individual diagnostic settings (namely sensitivity, specificity and predictive values), rather than those that are directly meaningful for incidence surveillance (namely those capturing the precision of the incidence estimator). For argument's sake, suppose that a test for recent infection is viewed as a test for identifying infections that have occurred in the preceding period T. The ratio of the MDRI to T could then be described as the average sensitivity of the test, specifically when the *true time since infection is uniformly distributed between 0 and* T, and the FRR could be described as one minus the *context-dependent* test specificity. For incidence surveillance, a high-performance test should then have a specificity close to 100%, but, critically, *does not require* a sensitivity close to 100%, and values larger than 50% would probably be difficult to achieve. This performance regime, though excellent for surveillance purposes, would be considered far from optimal by machine learning approaches that aim to maximise overall classification accuracy (and also conventionally assign the same 'costs' to 'false-positive' and 'false-negative' results).

Another aspect of defining a test for recent infection is choosing the post-infection time cut-off T, separating 'true-recent' from 'false-recent' results. A larger T would typically produce better test performance, as a larger T implies a larger (or unchanged) MDRI and (typically, but not always) a smaller FRR. However, T determines the maximum period in the past over which incidence is averaged, and allows for the decoupling of early test dynamics (summarised into the MDRI) from later dynamics which become convolved with the population's epidemiological and demographic history (captured by the FRR).[44] Therefore, while T should be large enough for the overwhelming majority of individuals to no longer return 'recent' results beyond T after infection, it should not be chosen much larger than necessary to achieve this. While not envisioned as being a parameter that is

[44] The decoupling of short-term and long-term test dynamics, achieved by introducing the post-infection time cut-off T into the framework for incidence estimation, is explored in Chapter 2.

finely tuned during test optimisation, choosing a suitable value of T is an important aspect of test design. For example, a value of one, two or three years might be chosen, depending on the test dynamics, surveillance study objectives, and availability of test characterisation data.

All of the discussion above has focused specifically on the interpretation of quantitative biomarker measurements to produce 'recent' and 'non-recent' classifications. Stepping further upstream into the development process, the optimisation of the technology that produces the measurements themselves should be considered. For example, tests for recent infection based on 'detuned' or 'less-sensitive' versions of standard HIV diagnostic assays were developed by modifying laboratory procedures, such as dilution specifications and incubation times. The modifications were introduced so that the measured biomarker grew more slowly over time post infection and produced a dynamic that supported the introduction of a threshold to distinguish 'recent' from 'non-recent' infection [91, 100, 104-106]. The laboratory procedures for these and other biomarkers for recent infection could be further modified to provide more optimal evolutions of quantitative measurements.

In principle, the optimisation of a test for recent infection, aiming to maximise the precision of incidence estimates, could be pushed arbitrarily far back into the development process. However, to be practical, early test development would simply seek biomarkers that appear predictive of the timing of infection, and that could potentially produce tests for recent infection with large MDRIs and small FRRs. It is in these earlier development stages that simple performance targets, such as those provided by the 'Target Product Profile' [13, 14, 188], are particularly useful. In fact, funding agencies are currently supporting the development of a collection of 'Target Product Profiles', each tailored to a particular use of a biomarker for recent infection (for example, as a stand-alone test for recent infection, or as part of a 'multi-assay algorithm' [88, 101]). It is only later on in the development process, when a limited number of well-defined parameters are to be tuned and sufficient data on test dynamics have been captured, that statistical analysis of the precision of the incidence estimator would be most practical and likely to add the most value. Even at this late stage, the scope of the optimisation should be carefully chosen, balancing sufficient flexibility of test design with practical needs.

The context-dependence of test performance

A subtlety when assessing the performance of tests for recent infection, based on the precision of the incidence estimator, is the dependence on context. The precision of the incidence estimator depends on (i) the size of the incidence surveillance survey, (ii) the proportions of HIV-negative, 'recently' infected and 'non-recently' infected subjects in the population, and (iii) the mean duration of recent infection (MDRI) and false-recent rate (FRR) and the uncertainties with which these test properties are measured (as well as T). The population proportions, and FRR, depend on the full epidemiological and demographic history of the population, convolved with the complete dynamics of the test for recent infection. Therefore, to assess the performance of any given test, hypothetical contexts need to be chosen, and should closely resemble the real-world contexts of interest.

Given any hypothetical context, various approaches could then be considered for calculating the precision of the incidence estimator. In the extreme case, when given complete knowledge of the population history and test dynamics, analytical or simulation-based approaches could be used to translate this information into inputs to a calculation of the precision of the incidence estimator, which could be performed using general methods for estimating uncertainties, such as bootstrap resampling [174]. However, such approaches are impractical for a number of reasons – they require unrealistic inputs, the calculations are unnecessarily complicated and computer intensive, and it is likely that highly uncertain results will be produced (given the uncertainties in the inputs). Therefore, various simplifications could be used, guided by what information is available and the applicability of analytical approximations in a particular setting.

For example, in the test performance calculations presented in Section 6.1 above, two simplifications were used. Firstly, population histories were summarised into the HIV prevalence and HIV incidence of the population at the time of the incidence study, where HIV incidence refers to the average incidence measured in the survey. Using the general expression for incidence (that leads to the estimator in Equation (6.3)), the population proportions were then calculated from the input test properties (and T), HIV incidence and HIV prevalence. Secondly, a closed form approximation for the precision of the incidence estimator was used, obtained by applying the delta method and assuming Gaussian uncertainty in all inputs (see Equation (6.4)).

In the calculation of precision, the context-specific (estimated) test properties need to be used. While the MDRI should capture only biological dynamics of the test over T after infection, and therefore ideally not rely on context, the FRR is expected to change by time and region [29, 189]. If a direct measurement of the FRR is not available for the population of interest (or one suitably similar), then knowledge about the population and test dynamics could be used to inform an FRR input. More specifically, the proportion of 'recent' results produced by a test will vary by subpopulation, as highlighted by the results in Section 5.1. Based on views about the prevalence of these subpopulations in the long-infected population at the time of the survey, and the probability of 'recent' results in each subpopulation, an appropriately averaged context-specific FRR can be calculated.

The stability of the test dynamics will determine how context-dependent the test properties are. For example, if test behaviour varies by HIV subtype, then the subtype-composition of the study population needs to be taken into account. This has the consequence that even the MDRI can become context-dependent, and, since the distribution of subtypes in a population may change over time, even time-dependent. This limitation on transferability of test characteristics undermines key benefits of biomarker-based cross-sectional surveillance. On the other hand, if a test for recent infection is not affected by factors such as subtype, and has a zero FRR in all subpopulations that could be considered, then the MDRI and (zero) FRR would be same in any context. Therefore, as more robust tests for recent infection are developed, with smaller FRRs, it is anticipated that the context-dependence of test properties will be reduced, making test performance assessment more straightforward.

While the context-dependence of the precision of incidence estimation is unavoidable, a test for recent infection would be most practical if it performs well across a range of epidemiologically relevant contexts. While context will affect the exact value of the incidence estimator precision, it is expected that context will only mildly affect determination of the optimal design for a test (such as the choice of threshold for distinguishing between 'recent' and 'non-recent' infections). Also, it is important that all relevant uncertainties are propagated through the calculation of test performance, to avoid misguided over-commitment to very particular test designs – for example, a confidence interval for the coefficient of variation of the incidence estimator should be reported. It is anticipated that the optimisation procedure will indicate a space of optimal test designs (for a range of contexts), rather than advocate exact (context-specific) test designs, given

the inevitable uncertainties of and sensitivities to inputs (for example, the precision of the incidence estimator is sensitive to the input uncertainties in test properties).

To demonstrate some of the ideas presented above, potential designs for a test for recent infection based on biomarkers measured by the Bio-Plex platform [86, 102] are compared in Table 6.1. The platform provides a number of measures of titre and avidity for antibodies responding to each of a number of specific HIV proteins. Published estimates of test properties (in treatment naïve populations) [102], for thirteen possible test designs (Tests 1 to 13), were used to estimate the coefficient of variation (CoV) of the incidence estimator in four hypothetical contexts (Contexts A-D in Table 6.1). The contexts capture different incidence and prevalence values, incidence study sizes, and uncertainties in test properties (and it is assumed that the published test properties can be recycled in these contexts).

The relative ordering of the candidate tests, according to their performances, remains similar across the contexts considered. The tests based on multiple biomarkers generally perform similarly to one another and outperform the tests based on single biomarkers, with the exception of Test 1, which provides relatively high performance yet utilises only a single biomarker (measuring the avidity of antibodies responding to the gp160 HIV protein). Sometimes, when moving from one test design to another, both test properties improve (or worsen), and the decrease (or increase) in the CoV is then useful for quantifying the size of the performance gain (or loss), which can be weighed against factors such as changes in costs and test complexity. In other cases, the changes in the MDRI and FRR counteract each other (they both increase or decrease), and the direction of the change in the CoV then indicates which of the two effects is stronger.

In Context D, which is the only context that accounts for uncertainties in test properties, the CoVs for the different test designs become more dispersed, highlighting the importance of correctly accounting for test property uncertainties. While a high, but known, FRR decreases the precision of the incidence estimator, it is the uncertainty about the test property that substantially penalises a test with a high FRR.

	Biomarkers included[i]	Threshold/ rule[ii]	MDRI (days)	FRR (%)	Coefficient of variation of incidence estimator (%), by context			
					A	B	C	D
1	160a	25	235	1.1	22	15	15	26
2	120a	20	265	4.5	27	19	18	35
3	41a	35	224	3.4	28	20	19	37
4	66a	10	279	27.8	61	43	38	92
5	160n	5	164	0.3	24	17	17	26
6	120n	7	176	8.4	50	35	32	72
7	160n, 120n, 66a, 120a, 160a, 41a	4/6	228	0.3	20	14	15	22
8	160n, 66a, 120a, 160a, 41a	3/5	257	1.1	21	15	15	24
9	120n, 66a, 120a, 160a, 41a	3/5	264	2.3	23	16	16	28
10	160n, 120n, 120a, 160a, 41a	3/5	239	0.3	20	14	14	21
11	120n, 120a, 160a, 41a	2/4	278	3.1	23	16	16	30
12	160n, 120a, 160a, 41a	2/4	266	1.4	21	15	15	25
13	120a, 160a, 41a	2/3	250	1.4	22	15	15	26

[i] Lists which biomarkers are included in the test for recent infection, where each biomarker identifier consists of a number that represents the HIV protein used and letter that indicates whether antibody titre (n) or avidity (a) is measured

[ii] If a single biomarker is used, this provides the biomarker-specific threshold (below which a measurement indicates 'recent' infection); if multiple biomarkers are used, this indicates the minimum number of biomarker measurements that must be below their biomarker-specific thresholds to obtain a 'recent' result (out of the total number of biomarkers considered) – for example, for Test 7, '4/6' indicates that at least 4 out of the 6 biomarkers must return measurements below their thresholds to produce a 'recent' result

Table 6.1: Coefficient of variation of the incidence estimator, for thirteen potential tests for recent infection based on multiple biomarkers produced by the Bio-Plex platform, for different contexts

The coefficient of variation of the incidence estimator (%) was calculated for each of four hypothetical contexts (A-D) using previously published estimates of test properties [102]. In Context A, HIV incidence was 1.5% per annum, HIV prevalence was 15%, the incidence study contained 5 000 subjects and there was no uncertainty in input test properties. Contexts B to D are each obtained by changing specific aspects of Context A. In Context B, the sample size was increased to 10 000 subjects. In Context C, HIV incidence and prevalence were increased to 2% per annum and 20% respectively. In Context D, non-zero uncertainties in test properties were accommodated – assuming a 5% coefficient of variation for the MDRI estimator, and that the FRR was measured as a binomial proportion in a sample of 500 subjects. The delta method approximation for the coefficient of variation of the incidence estimator was used.

Given the similarity of many of the CoVs, it is likely to be challenging to distinguish between many of these test designs (especially once uncertainties in the CoVs are formally accounted for). However, the scope of the optimisation could be vastly broadened, by allowing biomarker-specific thresholds to be tuned, considering all possible subsets of biomarkers for inclusion in the algorithm, and exploring other rules for defining the 'recent' region in the multi-dimensional space describing the multiple biomarkers.

Throughout the discussion above, test dynamics were combined with features of hypothetical populations and incidence study sizes to produce measures of precision of the incidence estimator. It is worthwhile keeping in mind that this relationship could instead be used in the opposite direction, by specifying the precision and making any of the inputs the subject of the equation to then be solved. For example, sample sizes required to produce suitably powered incidence studies could be calculated to inform study design. Alternatively, one could solve for the test properties that provide sufficient precision of incidence estimation, in chosen contexts, to inform guidelines for test development. Similarly, one could explore in which epidemiological contexts this surveillance approach is of utility, given currently available tests. The context-dependence of the performance of recent infection tests recent infection is unavoidable, and understanding this analytical subtlety and correctly accounting for it is an important aspect of optimising test design and understanding the utility of biomarker-based incidence surveillance.

Other criteria for assessing recent infection tests

While the variability of incidence estimates provides a sound metric for evaluating tests for recent infection for incidence surveillance from a theoretical point of view, this is not the only measure that would be considered in practice. A few examples of other criteria that are of importance are listed below.

The deployment of a test for recent infection into surveillance systems should be logistically feasible. This is also reflected in the 'Target Product Profile' [13, 14, 188], which stipulates objectives related to the type of sample and method of sample collection that are required, infrastructure and training needs, and test costs.

A central concept behind biomarker-based incidence surveillance is the transferability of test properties. Results should therefore be robust with regard to small variations in

procedures that may inevitably occur in practice. Biomarker dynamics should also be insensitive to factors such as HIV clade or gender, which would substantially vary across populations and in time, and could limit the transferability of even the MDRI. The FRR should ideally be close to zero in all possible subpopulations (such as treated subjects and elite controllers), as this would limit its variation by context.

Results from the analysis of the precision of the incidence estimator would need to be weighed against practical requirements. For example, the improvement in precision obtained by moving from one test design to another would need to be viewed alongside any increase in costs or test complexity.

Also, once a test that produces high precision incidence estimates has been identified, other aspects of operations research could be explored. For example, for tests that are based on algorithms of multiple biomarkers, where not all biomarker measurements always need to be known to unambiguously classify an infection, the order in which the biomarkers are applied could be optimised to minimise costs.

As highlighted by the examples above, while there is now a clear framework for defining and estimating test properties of relevance for incidence estimation, and converting these into direct metrics of test performance, there are numerous operational issues (mainly summarised by cost and robust transferability) which also influence test selection and optimisation. Ongoing developments in biomarker discovery and optimisation can reasonably be expected to lead to meaningful improvements in availability and practicality of recent infection tests for surveillance.

Chapter 7

Conclusion

As HIV epidemics continue to sweep across nations, typically disproportionately affecting those most in need, governments and public health organisations continue to commit themselves to the fight against the virus. Monitoring the spread of HIV and assessing the impact of interventions require reliable and practical methods for measuring infection rates. In the real world, new infections occur in a population at various points in time through a time-dependent, stochastic process. For the purpose of mathematically analysing the spread of the virus, *incidence* is thus defined in a continuum model world as the risk or hazard of infection per unit time per susceptible individual.

The measurement of this parameter is challenging as incidence is a *rate*, the consequences of which are only observable over time (rather than a *state*, which has a prevalence that can be directly estimated at any point in time). The estimation of incidence from single cross-sectional surveys has therefore attracted much interest: by application of a test for recent infection, to distinguish between 'recently' and 'non-recently' acquired infections, the prevalence of 'recent' HIV infection, together with estimated properties of the test, may be used to infer (average past) incidence.

This work presents a number of important theoretical and methodological advances, together with some practical applications and related corollaries, and addresses key obstacles that have been hampering the widespread use of this surveillance approach.

Firstly, a general theoretical framework was derived for inferring incidence from the cross-sectional application of a test for recent infection.

In the derivation of the general incidence estimator, assumptions about test dynamics were relaxed. These assumptions were used to develop earlier estimators although they are known to be violated in practice. Previously used assumptions of equilibrium conditions were also avoided by defining the measured incidence as a particular weighting of past incidence, determined by the early post-infection test dynamics and variation in the population size over recent times. Any residual biases were explicitly summarised into a set of mathematical terms, which an experimenter could examine to assess whether a study is in a regime of utility.

Through the introduction of a post-infection time cut-off T, completely general definitions of the two test properties that are required for incidence inference emerged from the analysis. The mean duration of recent infection (MDRI) is the average time alive and classified as 'recently' infected while infected for less than T, and the false-recent rate (FRR) is the probability that an individual who is infected for longer than T will produce a 'recent' result.

The general framework was compared to that of the competing school of thought, in which there is no cut-off T and test dynamics are summarised into a single property, namely the MDRI [7, 15, 27]. The primary benefit of introducing T and the FRR is that this creates a clear separation between (i) a carefully defined weighting of only recent incidence measured in the survey, which is loosely captured by the (stable) MDRI; and (ii) the full population history, which is neatly captured by the (context-dependent) FRR rather than obscuring the meaning of the incidence that otherwise would be measured.

Secondly, attention was turned to estimation of a test's MDRI: new approaches, current practices, and some ideas for reducing artefacts in analyses were investigated.

To address the bottleneck in test development posed by the reliance on longitudinal specimens that are collected by following subjects over time, the use of alternative specimen sets to obtain preliminary estimates of MDRIs was formalised and demonstrated. Only the classification of a subject's infection (as 'recent' or 'non-recent') at the time of HIV diagnosis is required, together with the time since the last HIV-negative test. The approach is suited to settings where infection and testing times are approximately independent and the times between HIV tests are large. In this work, two

previously untapped sources of specimens, namely blood donors and subjects tested as part of routine surveillance, were used to perform initial test characterisations.

Moving beyond preliminary characterisations of tests, accurate and precise MDRI estimates are required for use in surveillance studies. Therefore a detailed benchmarking of approaches for estimating the MDRI from longitudinal data was performed. Using simulated data, the performance of a range of methods that could be used in practice was thoroughly assessed in a number of modelled scenarios that capture essential features of what could be encountered in reality, including various underlying test dynamics, study designs, and subject behaviours. The dangers of neglecting noise in biomarkers or choosing inappropriate parametric forms were highlighted; and, while formal methods for analysing repeated-measures data most comprehensively captured data structures, simpler interpolation and regression approaches were also found to be of practical value. High performance of MDRI estimation was achieved across a range of realistic scenarios.

As highlighted by the results of the benchmarking exercise, the unknown infection times of subjects pose a particular challenge to the accurate estimation of the MDRI from longitudinal data. An analysis was presented that demonstrates some benefit from artificially controlling, by definition, entry into the 'HIV-positive and recently infected' state, provided that times of entry into this state can be accurately estimated given typically available data. Also, a straightforward, general approach was outlined for leveraging potentially complex testing history data to obtain the most informative possible estimates of infection times.

Thirdly, an important practical application, entailing the analysis of the characteristics of five prominent tests for recent infection, was presented.

The analysis was performed as a member of the *Consortium for the Evaluation and Performance of HIV Incidence Assays* (CEPHIA). CEPHIA was established in 2010 as an independent body, tasked with fostering consensus in the field and providing clear guidance to the public. The analysis of the first five tests to undergo a full evaluation by CEPHIA represents an important milestone, and captures more than three years of collaborations, specimen collection, laboratory testing, and data management and analysis by the team – all applying stringent quality control measures which it is hoped will guide standards in the field.

The results represent the first independent and consistent characterisation of multiple candidate tests. The analysis provides the basis for exploring the optimisation of the tests (an ongoing exercise by CEPHIA and other groups), as the struggle continues to achieve simultaneously, for all five tests, the goals of a large MDRI and small FRR, particularly among virally suppressed subjects.

Lastly, to support the ongoing development of tests for recent infection, a framework for measuring and optimising test performance was outlined.

In principle, any test can be consistently characterised and applied in a surveillance study to obtain valid incidence estimates, and therefore a standard metric for assessing the *utility* of tests for this surveillance application was needed. The precision of the incidence estimator was presented as a measure for formally and consistently balancing the needs for a large MDRI and small FRR. The guidance aimed to counter the use of more familiar metrics (sensitivity, specificity and predictive values) that are suited specifically to *individual diagnostics* settings and would lead to spurious evaluations of tests for estimating *population-level average* incidence rates.

When optimising the design of a test by maximising the precision of incidence estimation, a number of practical considerations arise, and some key aspects of these were briefly explored. The scope of any optimisation should be carefully defined – while it should be broad enough for high performance designs to be discovered, it should be suitably restricted so that the analysis remains tractable and tests are not overly complicated. Also, test performance is unavoidably dependent on context, and the ideal test would perform well across a range of epidemiologically relevant contexts. Importantly, the selection and optimisation of a test to be integrated into surveillance studies would also be influenced by operational factors, primarily related to cost and transferability of the technology.

*Collectively***,** the contributions made within this work provide essential theoretical and methodological foundations for the application, characterisation and optimisation of recent infection tests for HIV incidence surveillance. An important aspect of this research was the sharing of ideas – in journals, at conferences, through collaborations and at meetings of working groups, by providing training, and by developing online resources [29-50, 52, 53]. Through these efforts, this research has already had demonstrable impact in shaping discussion in the area and building consensus.

Focus in the field has now shifted towards the practical application of cross-sectional incidence surveillance. In March 2014, the *Centers for Disease Control and Prevention* (CDC) and US *President's Emergency Fund for AIDS Relief* (PEPFAR) announced a funding opportunity worth 125 million US dollars for improving the monitoring of HIV and assessment of interventions [221]. One of the funding goals is the complete integration of cross-sectional HIV incidence estimation, using tests for recent infection, into general population-based surveillance systems in a number of resource-limited countries.

Ongoing endeavours to develop and optimise candidate tests for recent infection should continue to be supported, as well as efforts to meticulously characterise and deploy chosen tests. There have been increasing funding opportunities for test developers over recent years [188, 218], and CEPHIA is currently supporting the work of a number of biomarker discovery groups. Also, more general and sophisticated analysis tools will need to be developed and made available, and individuals trained in their use. The tools should support a range of analyses that may be required, from the characterisation of recent infection tests to sophisticated power calculations for detecting incidence trends by repeating incidence studies over time.

Tests for recent infection find application in a number of closely related areas of work, which are beyond the scope of this thesis. In some countries, incidence is estimated by applying recent infection tests to subjects who are diagnosed with HIV in health-care settings (rather than to HIV-positive subjects identified in a survey), using inputs about testing behaviours and at-risk population sizes [134-138]. Also, an area that is currently largely undeveloped, but is starting to attract interest, is the estimation of incidence from multiple sources of data – that is, by combining measurements of the prevalence of 'recent' HIV infection with, for example, sentinel surveillance data and national household survey data. Countries have also begun using tests for recent infection in clinical settings, to provide individual feedback on the likely time since infection at the time of diagnosis [139-141]. While approaches for interpreting and reporting results still require much refinement, this clinical application is important as staging information could be used effectively to support contact-tracing and the tailoring of counselling and treatment plans. As all of these applications of recent infection tests, including cross-sectional incidence surveillance, are further developed and their uptake increases, tests will become more marketable. Perceived profitability would be an important driver of the required product development.

This thesis has focused on estimating HIV incidence using tests for recent HIV infection. As a closing remark, it is possible that transient early stages of other conditions could be defined and their prevalence measured in surveys to infer incidence. The nuances and limitations of each application would need to be carefully investigated, and it will be interesting to explore how other conditions' biomarkers for recent infection fit into the framework for application and characterisation that has been established for HIV. A practical challenge will be the development of suitable tests for recent infection in this wider context. Even for HIV, high-performance and practical recent infection tests continue to be sought, despite HIV being the focus over the last two decades. The community needs to continue to translate knowledge into practice, so that the impact of the presented advances in incidence surveillance can be fully realised.

Appendix A

Statistical Expressions Underlying Online Tools

A suite of analysis tools, called Assay-Based Incidence Estimation (ABIE) v2.0, has been made available online to support application of the theoretical and methodological frameworks presented in this work (www.incidence-estimation.com) [48].The toolset consists of a number of 'calculators', each designed to perform a particular analysis related to the application of tests for recent infection for incidence surveillance. The calculation of outputs from users' inputs is described in this appendix. Uncertainties and distributions of estimators appearing in the tools are approximated using the delta method and other large-sample results, which are summarised in Section A1. Each of the remaining sections, A2 to A7, outlines the objective, inputs and outputs of a single calculator in the tool suite.

The calculators in this version of the ABIE toolset are in the form of Microsoft Excel spreadsheets, and were designed to be particularly uncomplicated by utilising closed-form approximations (if not exact analytical expressions) for all calculations and the most straightforward test statistics (differences and ratios of incidences estimates), as discussed in Section 2.3.

All inputs that are measurements of time, namely the post-infection time cut-off T and the mean duration of recent infection (MDRI), are in units of years. The incidence rate is expressed as the hazard of infection per person year. A sample count is in units of subjects. The false-recent rate (FRR), prevalence, and the risk of infection over a specified period, are all dimensionless and bounded by 0 and 1. Any coefficient of variation (CoV) is also dimensionless.

A1 Uncertainties and Distributions of Estimators

The approximations for the uncertainties and distributions of statistics used in the calculators rely on results obtained from Taylor polynomial approximations, the delta method and maximum likelihood theory. The main results used are summarised below.

Three statistics appear in the calculators: the incidence estimator, the difference between two incidence estimators, and the ratio of two incidence estimators. Each of these statistics can be expressed in terms of (i) random variables contained in the vector $\underline{\alpha} = [\alpha_1, \alpha_2, \ldots \alpha_A]^T$, which capture sampling variability, and (ii) fixed parameters which capture properties of the study population(s), test for recent infection and study design. The stochastic nature of the scalar statistic can therefore be fully captured by its representation as $g(\underline{\alpha})$.

In all cases considered in the calculators, $\underline{\alpha}$ captures variability in sample counts (of uninfected, 'recently' infected and 'non-recently' infected subjects) and variability in estimated test characteristics (the MDRI and FRR). By construct (by using conditional distributions), the α_i are always independent. In particular, the α_i are defined to be (identically and independently distributed) standard normal random variables, and this is expected to provide a reasonable distributional approximation, in most regimes, when sample sizes for the incidence study and test characterisation studies are large. It is recognised that a Gaussian distribution for the estimated FRR may have limited applicability as improved tests for recent infection are developed and FRRs become very small.

To begin, the properties of the statistic can be explored by approximating the statistic by a first-order Taylor polynomial in α_i. A multivariate Taylor series expansion of $g(\underline{\alpha})$ around $E[\underline{\alpha}] = \underline{\alpha}_0$ is given by

$$g(\underline{\alpha}) = g(\underline{\alpha}_0) + (\underline{\alpha} - \underline{\alpha}_0)^{\mathrm{T}} \cdot Dg(\underline{\alpha}_0) + \frac{1}{2!}(\underline{\alpha} - \underline{\alpha}_0)^T \cdot D^2 g(\underline{\alpha}_0) \cdot (\underline{\alpha} - \underline{\alpha}_0) + \cdots, \quad \text{(A1)}$$

where $Dg(\underline{\alpha}_0) = \nabla g(\underline{\alpha}_0)$ is the gradient of $g(\underline{\alpha})$ evaluated at $\underline{\alpha} = \underline{\alpha}_0$,

$$\nabla g(\underline{\alpha}_0) = \left[\frac{\partial g(\underline{\alpha})}{\partial \alpha_1}, \frac{\partial g(\underline{\alpha})}{\partial \alpha_2}, \ldots, \frac{\partial g(\underline{\alpha})}{\partial \alpha_A}\right]^T \Bigg|_{\underline{\alpha}=\underline{\alpha}_0}, \tag{A2}$$

and $D^2 g(\underline{\alpha}_0) = H\big(g(\underline{\alpha}_0)\big)$ is the Hessian matrix of $g(\underline{\alpha})$ evaluated at $\underline{\alpha} = \underline{\alpha}_0$,

$$\big[\mathrm{H}\big(g(\underline{\alpha}_0)\big)\big]_{i,j} = \frac{\partial g^2(\underline{\alpha})}{\partial \alpha_i \partial \alpha_j}\Bigg|_{\underline{\alpha}=\underline{\alpha}_0}. \tag{A3}$$

Retaining only the first two terms of the Taylor series expansion yields the following approximation for $g(\underline{\alpha})$:

$$g(\underline{\alpha}) \approx g(\underline{\alpha}_0) + (\underline{\alpha} - \underline{\alpha}_0)^{\mathrm{T}} \cdot \mathrm{D}g(\underline{\alpha}_0). \tag{A4}$$

Applying the mean and variance operator on both sides of Equation (A4), in turn, produces the following expressions for the moments of the statistic:

$$E[g(\underline{\alpha})] \approx g(\underline{\alpha}_0) \tag{A5}$$

$$Var\big(g(\underline{\alpha})\big) \approx \mathrm{D}g(\underline{\alpha}_0)^T \cdot Var(\underline{\alpha}) \cdot \mathrm{D}g(\underline{\alpha}_0). \tag{A6}$$

The approximations captured by Equations (A4) to (A6) will improve as $g(\underline{\alpha})$ is increasingly linear in $\underline{\alpha}$ near $\underline{\alpha}_0$. By introducing distributional properties for $\underline{\alpha}$, the distributional properties of $g(\underline{\alpha})$ can be explored. For normally distributed α_i, it follows that $g(\underline{\alpha})$ is also (approximately) normally distributed.

Formally, the delta method (for example, see [148]) shows that functions of asymptotically multivariate-normally distributed random variables are often themselves asymptotically normally distributed. More specifically, if

$$\sqrt{n} \cdot (\underline{\alpha} - \underline{\alpha}_0) \xrightarrow{d} N(\underline{0}, \Sigma) \text{ as some sample size } n \to \infty, \tag{A7}$$

then, for $g(\underline{\alpha})$ that has a non-zero differential at $\underline{\alpha}_0$,

$$\sqrt{n} \cdot \big(g(\underline{\alpha}) - g(\underline{\alpha}_0)\big) \xrightarrow{d} N\left(\underline{0}, \mathrm{D}g(\underline{\alpha}_0)^{\mathrm{T}} \cdot \Sigma \cdot \mathrm{D}g(\underline{\alpha}_0)\right) \text{ as } n \to \infty. \tag{A8}$$

Leaning on the results above, since $\underline{\alpha}$ is defined to have a $N(\underline{0}, I_A)$ distribution where I_A is an A by A identity matrix (based on the anticipated limiting distribution of $\underline{\alpha}$ as all relevant sample sizes tend to infinity), the large-sample distribution of the statistic $g(\underline{\alpha})$ is

$$g(\underline{\alpha}) \dot{\sim} N\left(g(\underline{0}), \sum_{i=1}^{A} \left(\frac{\partial g(\underline{\alpha})}{\partial \alpha_i}\right)^2 \Bigg|_{\underline{\alpha}=\underline{0}}\right). \tag{A9}$$

The large-sample approximate normality of the statistic also follows from maximum likelihood theory (for example, see [222]), namely from the results that (i) maximum likelihood estimators are asymptotically normally distributed (subject to regularity conditions), and (ii) functions of maximum likelihood estimators are themselves maximum likelihood estimators (the 'invariance property'). The incidence estimator (based on Equations (2.23) and (2.25)) can be expressed as

$$\hat{I}_T = \hat{I}_T\left(\hat{P}_S, \hat{P}_+, \hat{P}_R, \widehat{\Omega}_T, \hat{\beta}_T\right) = \frac{\hat{P}_R - \hat{\beta}_T \hat{P}_+}{\hat{P}_S \cdot \left(\widehat{\Omega}_T - \hat{\beta}_T T\right)}, \tag{A10}$$

where $\hat{P}_S$, $\hat{P}_+$ and $\hat{P}_R$ are estimators for P_S, P_+ and P_R, which are the proportions of individuals in the population who are susceptible (HIV-negative), infected and 'recently' infected respectively ($P_S + P_+ = 1$), and $\widehat{\Omega}_T$ and $\hat{\beta}_T$ are estimators for Ω_T and β_T, which are the test MDRI and FRR respectively for a time cut-off T. If the required inputs ($\hat{P}_S$, $\hat{P}_+$, $\hat{P}_R$, $\widehat{\Omega}_T$ and $\hat{\beta}_T$) are obtained using a maximum likelihood approach (as would typically be the case), it follows that the incidence estimator is a maximum likelihood estimator, and is therefore asymptotically normally distributed.

In the calculators, the expression provided in Equation (A9) is used to approximate the distribution and moments of the incidence estimator, incidence difference estimator, and incidence ratio estimator. In the descriptions of the calculators in the sections below, to obtain compact equations, uncertainty for an estimator is expressed as either a CoV (ratio of standard deviation to mean) or standard deviation (or the square thereof). The CoV or standard deviation depends on (typically) unknown parameters (population proportions, test characteristics, and uncertainties of test characteristic estimates) that appear in $g(\underline{\alpha})$, and, in all applications, the estimated (or assumed) values of these parameters are used as proxies of the true values so that uncertainties can be assessed.

A2 Incidence Prevalence Calculator

Objective

This calculator estimates HIV incidence and prevalence from the input sample counts observed in the cross-sectional survey and test characteristics.

Inputs

- Post-infection time cut-off T
- Estimated (or assumed) test characteristics
  - Estimated MDRI $\widehat{\Omega}_T$
  - Estimated FRR $\hat{\beta}_T$
  - Estimated CoV of MDRI estimator $\hat{c}_{\widehat{\Omega}_T}$
  - Estimated CoV of FRR estimator $\hat{c}_{\hat{\beta}_T}$
- Observed sample counts in the survey of n subjects
  - Number of HIV-negative subjects n_s versus HIV-positive subjects n_+, where $n_S + n_+ = n$
  - Number of 'recently' infected subjects n_R versus 'non-recently' infected subjects n_{NR}, where $n_R + n_{NR} = n_+$

HIV prevalence, P_+

The point estimate for HIV prevalence is

$$\hat{P}_+ = \frac{n_+}{n}. \tag{A11}$$

The limits of the 95% confidence interval for HIV prevalence, $[\hat{P}_{+,l}, \hat{P}_{+,u}]$, are equal to

$$\hat{P}_{+,l} = \begin{cases} 0 & \text{if } n_+ = 0 \\ \dfrac{v_{1,l} \cdot F(v_{1,l}, v_{2,l}, \frac{\alpha}{2})}{v_{2,l} + v_{1,l} \cdot F(v_{1,l}, v_{2,l}, \frac{\alpha}{2})} & \text{if } n_+ > 0 \end{cases} \tag{A12}$$

$$\hat{P}_{+,u} = \begin{cases} \dfrac{v_{1,u} \cdot F(v_{1,u}, v_{2,u}, 1-\frac{\alpha}{2})}{v_{2,u} + v_{1,u} \cdot F(v_{1,u}, v_{2,u}, 1-\frac{\alpha}{2})} & \text{if } n_+ < n \\ 1 & \text{if } n_+ = n \end{cases}, \tag{A13}$$

where $F(v_1, v_2, q)$ is the lower q^{th} percentile of an F distribution with v_1 and v_2 degrees of freedom, $\alpha = 0.05$, $v_{1,l} = 2 \cdot n_+$, $v_{2,l} = 2 \cdot (n - n_+ + 1)$, $v_{1,u} = 2 \cdot (n_+ + 1)$ and $v_{2,u} = 2 \cdot (n - n_+)$. This Clopper-Pearson 'exact' binomial proportion confidence interval provides a conservative coverage of least 95% [210].

HIV incidence rate, I_T

The point estimate for the incidence rate is

$$\hat{I}_T = \frac{n_R - \hat{\beta}_T n_+}{n_S \cdot \left(\widehat{\Omega}_T - \hat{\beta}_T T\right)}. \tag{A14}$$

While the true incidence in the study population is non-negative, the incidence estimator can take on any real value as a result of variability in input counts and estimated test characteristics, as well as potential biases in inputs (for example, that may occur if test characteristics are estimated in populations that are not representative of the study population).

The limits of the 95% confidence interval for incidence, $\left[\hat{I}_{T,l}, \hat{I}_{T,u}\right]$, are given by

$$\hat{I}_{T,l} = \hat{I}_T - z_{1-\alpha/2} \cdot \hat{c}_{\hat{I}_T} \cdot \hat{I}_T \tag{A15}$$

$$\hat{I}_{T,u} = \hat{I}_T + z_{1-\alpha/2} \cdot \hat{c}_{\hat{I}_T} \cdot \hat{I}_T, \tag{A16}$$

where $z_q = \Phi^{-1}(q)$ is the lower q^{th} percentile of a standard normal distribution, $\alpha = 0.05$, and $\hat{c}_{\hat{I}_T}$ is the estimated CoV of the incidence estimator. The CoV of the incidence estimator, $c_{\hat{I}_T}$, is approximated (see Section A.1) by

$$\begin{aligned} \hat{c}^2_{\hat{I}_T} &= \frac{1}{n\hat{P}_+} \cdot \left(\frac{1}{\hat{P}_S} + \frac{\hat{P}_R \hat{P}_{NR}}{\left(\hat{P}_R - \hat{\beta}_T \hat{P}_+\right)^2}\right) \\ &+ \hat{\sigma}^2_{\widehat{\Omega}_T} \cdot \left(\frac{1}{\widehat{\Omega}_T - \hat{\beta}_T T}\right)^2 \\ &+ \hat{\sigma}^2_{\hat{\beta}_T} \cdot \left(\frac{\hat{P}_+ \widehat{\Omega}_T - \hat{P}_R T}{\left(\hat{P}_R - \hat{\beta}_T \hat{P}_+\right)\left(\widehat{\Omega}_T - \hat{\beta}_T T\right)}\right)^2, \end{aligned} \tag{A17}$$

where

$$\hat{P}_S = 1 - \hat{P}_+ = \frac{n_S}{n} \tag{A18}$$

$$\hat{P}_+ = \frac{n_+}{n} \tag{A19}$$

$$\hat{P}_R = \frac{n_R}{n} \tag{A20}$$

$$\hat{P}_{NR} = \frac{n_{NR}}{n} \tag{A21}$$

are the estimated proportions of susceptible, infected, ‘recently’ infected and ‘non-recently’ infected individuals in the population.

The result for the uncertainty of the incidence estimator was obtained by expressing sample counts and estimated test characteristics in terms of independently distributed standard normal random variables α_i ($i = 1,2,3,4$) as follows (outlined in Section 2.2.2, but reproduced here for completeness):

$$n_S = n_s(\alpha_1) = nP_S + \sigma_S\alpha_1 \tag{A22}$$

$$n_R = n_R(\alpha_1, \alpha_2) = nP_R - \sigma_R\alpha_1 + \sigma_{R,NR}(\alpha_1)\alpha_2 \tag{A23}$$

$$n_{NR} = n_{NR}(\alpha_1, \alpha_2) = nP_{NR} - \sigma_{NR}\alpha_1 - \sigma_{R,NR}(\alpha_1)\alpha_2 \tag{A24}$$

$$\widehat{\Omega}_T = \widehat{\Omega}_T(\alpha_3) = \Omega_T + \sigma_{\widehat{\Omega}_T}\alpha_3 \tag{A25}$$

$$\hat{\beta}_T = \hat{\beta}_T(\alpha_4) = \beta_T + \sigma_{\hat{\beta}_T}\alpha_4, \tag{A26}$$

where P_S, P_R and P_{NR} are the proportions of susceptible, 'recently' infected and 'non-recently' infected individuals in the population respectively; $\sigma_{\widehat{\Omega}_T}$ and $\sigma_{\hat{\beta}_T}$ are the standard deviations of the (unbiased) estimators for Ω_T and β_T respectively; and

$$\sigma_S = \sqrt{nP_S \cdot (1 - P_s)} \tag{A27}$$

$$\sigma_R = \frac{P_R}{P_R + P_{NR}}\sigma_s \tag{A28}$$

$$\sigma_{NR} = \frac{P_{NR}}{P_R + P_{NR}}\sigma_s \tag{A29}$$

$$\sigma_{R,NR}(\alpha) = \frac{\sqrt{(n - nP_S - \sigma_S\alpha\,)P_RP_{NR}}}{P_R + P_{NR}}. \tag{A30}$$

The resulting approximation for the CoV of the incidence estimator is

$$\begin{aligned} c_{\hat{I}_T}^2 \;&\approx \frac{1}{nP_+} \cdot \left(\frac{1}{P_S} + \frac{P_RP_{NR}}{(P_R - \beta_TP_+)^2}\right) \\ &+ \sigma_{\widehat{\Omega}_T}^2 \cdot \left(\frac{1}{\Omega_T - \beta_TT}\right)^2 \\ &+ \sigma_{\hat{\beta}_T}^2 \cdot \left(\frac{P_+\Omega_T - P_RT}{(P_R - \beta_TP_+)(\Omega_T - \beta_TT)}\right)^2. \end{aligned} \tag{A31}$$

In Equations (A15) to (A17), the estimated CoV, $\hat{c}_{\hat{I}_T}$, is negative if the point estimate for incidence (used as a proxy for the mean estimate) is negative.

Annual risk of HIV infection, R_T

The annual risk of infection is the probability that a susceptible individual will get infected over the period of a year. The point estimate for this annual risk is

$$\hat{R}_T = 1 - \exp(-\hat{I}_T), \quad \text{(A32)}$$

based on a constant hazard of infection of $\hat{I}_T$ over the year.

The limits of the 95% confidence interval for the annual risk of infection, $[\hat{R}_{T,l}, \hat{R}_{T,u}]$, are obtained by applying an equivalent transformation to the confidence interval limits for the incidence rate:

$$\hat{R}_{T,l} = 1 - \exp(-\hat{I}_{T,l}) \quad \text{(A33)}$$

$$\hat{R}_{T,u} = 1 - \exp(-\hat{I}_{T,u}). \quad \text{(A34)}$$

A3 Sample Size Calculator

Objective

This tool calculates the sample size required to achieve a desired precision of incidence estimation in a surveillance survey from the input test characteristics, epidemiological context, and desired precision of the incidence estimator.

Inputs

- Post-infection time cut-off T
- Estimated (or assumed) test characteristics
  - Estimated MDRI $\widehat{\Omega}_T$
  - Estimated FRR $\hat{\beta}_T$
  - Estimated CoV of MDRI estimator $\hat{c}_{\widehat{\Omega}_T}$
  - Estimated CoV of FRR estimator $\hat{c}_{\hat{\beta}_T}$
- Assumed epidemiological context
  - HIV incidence I
  - HIV prevalence P
- Desired CoV of the incidence estimator $c_{\hat{I}_T}$

Sample size, n

The (estimated) sample size, required to achieve the desired CoV for the incidence estimator, is

$$\tilde{n} = \frac{1}{\tilde{P}_+} \cdot \left(\frac{1}{\tilde{P}_S} + \frac{\tilde{P}_R \tilde{P}_{NR}}{\left(\tilde{P}_R - \hat{\beta}_T \tilde{P}_+\right)^2} \right) \times$$

$$\left(c_{\hat{I}_T}^2 - \hat{\sigma}_{\widehat{\Omega}_T}^2 \cdot \left(\frac{1}{\Omega_T - \beta_T T} \right)^2 - \hat{\sigma}_{\hat{\beta}_T}^2 \cdot \left(\frac{\tilde{P}_+ \widehat{\Omega}_T - \tilde{P}_R T}{\left(\tilde{P}_R - \hat{\beta}_T \tilde{P}_+\right)\left(\widehat{\Omega}_T - \hat{\beta}_T T\right)} \right)^2 \right)^{-1} \quad \text{(A35)}$$

where

$$\tilde{P}_S = 1 - P \quad \text{(A36)}$$

$$\tilde{P}_+ = P \quad \text{(A37)}$$

$$\tilde{P}_R = I \cdot (1 - P)\left(\widehat{\Omega}_T - \hat{\beta}_T T\right) + \hat{\beta}_T P \quad \text{(A38)}$$

$$\tilde{P}_{NR} = P - \tilde{P}_R \quad \text{(A39)}$$

are the proportions of susceptible, infected, 'recently' infected and 'non-recently' infected individuals in the population for the input context (based on input test characteristic estimates). This result is obtained by solving for n in the CoV of the incidence estimator given in Equation (A31).

A4 Incidence Ratio Calculator

Objective

This calculator estimates the ratio between the incidence values in two study populations, from input sample counts observed in the cross-sectional survey performed in each of the two study populations and test characteristics. Test characteristics are assumed to be the same in both populations.

Inputs

- Estimated (or assumed) test characteristics
  - Estimated FRR $\hat{\beta}_T$
  - Estimated CoV of FRR estimator $\hat{c}_{\hat{\beta}_T}$
- Observed sample counts in the survey of n_k subjects in study population k ($k = 1,2$)

- Number of HIV-negative subjects $n_{S,k}$ versus HIV-positive subjects $n_{+,k}$, where $n_{S,k} + n_{+,k} = n_k$
- Number of 'recently' infected subjects $n_{R,k}$ versus 'non-recently' infected subjects $n_{NR,k}$, where $n_{R,k} + n_{NR,k} = n_{+,k}$

No knowledge of the time cut-off T or MDRI is required, as the term containing T and the MDRI appearing in the incidence estimator cancels out in the ratio of two incidence estimators (assuming equal values of T and test characteristics in the populations).

Incidence ratio, *IR*

The incidence ratio, the ratio of the incidence in study population 2 to the incidence in study population 1, is equal to

$$IR = \frac{N_{R,2} - \beta_T N_{+,2}}{N_{S,2} \cdot (\Omega_T - \beta_T T)} \div \frac{N_{R,1} - \beta_T N_{+,1}}{N_{S,1} \cdot (\Omega_T - \beta_T T)}$$

$$= \frac{N_{R,2} - \beta_T N_{+,2}}{N_{R,1} - \beta_T N_{+,1}} \cdot \frac{N_{S,1}}{N_{S,2}}, \tag{A40}$$

where $N_{S,k}$ and $N_{+,k}$ are the numbers of susceptible (or uninfected) and infected individuals respectively, in study population k ($k = 1,2$) of total size $N_k = N_{S,k} + N_{+,k}$, and $N_{R,k}$ is the number of 'recently' infected individuals in the population.

The point estimate for this parameter is

$$\widehat{IR} = \frac{n_{R,2} - \hat{\beta}_T n_{+,2}}{n_{R,1} - \hat{\beta}_T n_{+,1}} \cdot \frac{n_{S,1}}{n_{S,2}}. \tag{A41}$$

The limits of the 95% confidence interval for the incidence ratio, $\left[\widehat{IR}_l, \widehat{IR}_u\right]$, are given by

$$\widehat{IR}_l = \widehat{IR} - z_{1-\alpha/2} \cdot \hat{c}_{\widehat{IR}} \cdot \widehat{IR} \tag{A42}$$

$$\widehat{IR}_u = \widehat{IR} + z_{1-\alpha/2} \cdot \hat{c}_{\widehat{IR}} \cdot \widehat{IR}, \tag{A43}$$

where $z_q = \Phi^{-1}(q)$ is the lower q^{th} percentile of a standard normal distribution, $\alpha = 0.05$ and $\hat{c}_{\widehat{IR}}$ is the estimated CoV of the incidence ratio estimator.

The CoV of the incidence ratio estimator, $c_{\widehat{IR}}$, is estimated (see Section A.1) by:

$$\hat{c}^2_{\widehat{IR}} = \frac{1}{n\hat{P}_{+,1}} \cdot \left(\frac{1}{\hat{P}_{S,1}} + \frac{\hat{P}_{R,1}\hat{P}_{NR,1}}{\left(\hat{P}_{R,1} - \hat{\beta}_T \hat{P}_{+,1}\right)^2} \right) + \frac{1}{n\hat{P}_{+,2}} \cdot \left(\frac{1}{\hat{P}_{S,2}} + \frac{\hat{P}_{R,2}\hat{P}_{NR,2}}{\left(\hat{P}_{R,2} - \hat{\beta}_T \hat{P}_{+,2}\right)^2} \right) + \hat{\sigma}^2_{\hat{\beta}_T} \cdot \left(\frac{\hat{P}_{+,1}}{\hat{P}_{R,1} - \hat{\beta}_T \hat{P}_{+,1}} - \frac{\hat{P}_{+,2}}{\hat{P}_{R,2} - \hat{\beta}_T \hat{P}_{+,2}} \right)^2, \quad \text{(A44)}$$

where

$$\hat{P}_{S,k} = \frac{n_{S,k}}{n_k} \quad \text{(A45)}$$

$$\hat{P}_{+,k} = \frac{n_{+,k}}{n} \quad \text{(A46)}$$

$$\hat{P}_{R,k} = \frac{n_{R,k}}{n_k} \quad \text{(A47)}$$

$$\hat{P}_{NR,k} = \frac{n_{NR,k}}{n_k} \quad \text{(A48)}$$

are the estimated proportions of susceptible, infected, 'recently' infected and 'non-recently' infected individuals in study population k.

The CoV approximation was obtained by expressing the sample counts and estimated test characteristics as functions of independently distributed standard normal random variables α_i $(i = 1,2, \ldots 5)$ as follows:

$$n_{S,1} = n_{S,1}(\alpha_1) = n_1 P_{S,1} + \sigma_{S,1}\alpha_1 \quad \text{(A49)}$$

$$n_{R,1} = n_{R,1}(\alpha_1, \alpha_2) = n_1 P_{R,1} - \sigma_{R,1}\alpha_1 + \sigma_{R,NR,1}(\alpha_1)\alpha_2 \quad \text{(A50)}$$

$$n_{NR,1} = n_{NR,1}(\alpha_1, \alpha_2) = n_1 P_{NR,1} - \sigma_{NR,1}\alpha_1 - \sigma_{R,NR,1}(\alpha_1)\alpha_2 \quad \text{(A51)}$$

$$n_{S,2} = n_{S,2}(\alpha_3) = n_2 P_{S,2} + \sigma_{S,2}\alpha_3 \quad \text{(A52)}$$

$$n_{R,2} = n_{R,2}(\alpha_3, \alpha_4) = n_2 P_{R,2} - \sigma_{R,2}\alpha_3 + \sigma_{R,NR,2}(\alpha_3)\alpha_4 \quad \text{(A53)}$$

$$n_{NR,2} = n_{NR,2}(\alpha_3, \alpha_4) = n_2 P_{NR,2} - \sigma_{NR,2}\alpha_3 - \sigma_{R,NR,2}(\alpha_3)\alpha_4 \quad \text{(A54)}$$

$$\hat{\beta}_T = \hat{\beta}_T(\alpha_5) = \beta_T + \sigma_{\hat{\beta}_T}\alpha_5, \quad \text{(A55)}$$

where $P_{S,k}$ $P_{R,k}$, $P_{NR,k}$ are the proportions of susceptible, 'recently' infected and 'non-recently' infected individuals in population k respectively, $P_{S,k} + P_{R,k} + P_{NR,k} = 1$; $\sigma_{\widehat{\beta_T}}$ is the standard deviation of the (unbiased) estimator for β_T; and

$$\sigma_{S,k} = \sqrt{n_k P_{S,k} \cdot \left(1 - P_{S,k}\right)} \tag{A56}$$

$$\sigma_{R,k} = \frac{P_{R,k}}{P_{R,k} + P_{NR,k}} \sigma_{S,k} \tag{A57}$$

$$\sigma_{NR,k} = \frac{P_{NR,k}}{P_{R,k} + P_{NR,k}} \sigma_{S,k} \tag{A58}$$

$$\sigma_{R,NR,k}(\alpha) = \frac{\sqrt{\left(n_k - n_k P_{S,k} - \sigma_{S,k}\alpha\right) P_{R,k} P_{NR,k}}}{P_{R,k} + P_{NR,k}}. \tag{A59}$$

The approximation for the CoV of the incidence ratio estimator that is produced is

$$\begin{aligned} c^2_{\widehat{IR}} \approx\ & \frac{1}{nP_{+,1}} \cdot \left(\frac{1}{P_{S,1}} + \frac{P_{R,1} P_{NR,1}}{\left(P_{R,1} - \beta_T P_{+,1}\right)^2} \right) \\ & + \frac{1}{nP_{+,2}} \cdot \left(\frac{1}{P_{S,2}} + \frac{P_{R,2} P_{NR,2}}{\left(P_{R,2} - \beta_T P_{+,2}\right)^2} \right) \\ & + \sigma^2_{\widehat{\beta_T}} \cdot \left(\frac{P_{+,1}}{P_{R,1} - \beta_T P_{+,1}} - \frac{P_{+,2}}{P_{R,2} - \beta_T P_{+,2}} \right)^2. \end{aligned} \tag{A60}$$

In Equations (A42) to (A44), the estimated CoV, $\hat{c}_{\widehat{IR}}$, is negative if the point estimate for the incidence ratio (used as a proxy for the mean estimate) is negative.

A5 P-value for Difference Calculator

Objective

This tool calculates a p-value (probability of obtaining a result as extreme as that observed) for the difference between incidence estimates for two study populations, under a hypothesis of equal HIV incidence, from input sample counts observed in the cross-sectional survey performed in each study population and test characteristics. Test characteristics are the same in the two populations, and equal HIV prevalence is also assumed.

Inputs

- Estimated (or assumed) test characteristics
  - Estimated FRR $\hat{\beta}_T$
- Observed sample counts in the survey of n_k subjects in study population k ($k = 1,2$)
  - Number of HIV-negative subjects $n_{S,k}$ versus HIV-positive subjects $n_{+,k}$, where $n_{S,k} + n_{+,k} = n_k$
  - Number of 'recently' infected subjects $n_{R,k}$ versus 'non-recently' infected subjects $n_{NR,k}$, where $n_{R,k} + n_{NR,k} = n_{+,k}$

No knowledge of the time cut-off T, MDRI, or uncertainty in test characteristics is required, as these parameters appear in common factors in the incidence difference or fall away given the particular approximations used.

P-value for difference between incidence estimates, p

The parameter about which inference is of interest is the incidence difference, ID, which is given by

$$\begin{aligned} ID &= I_1 - I_2 \\ &\frac{N_{R,1} - \beta_T N_{+,1}}{N_{S,1} \cdot (\Omega_T - \beta_T T)} - \frac{N_{R,2} - \beta_T N_{+,2}}{N_{S,2} \cdot (\Omega_T - \beta_T T)}, \end{aligned} \tag{A61}$$

where $N_{S,k}$ and $N_{+,k}$ are the numbers of susceptible (or uninfected) and infected individuals respectively, in study population k ($k = 1,2$) of size $N_k = N_{S,k} + N_{+,k}$, and $N_{R,k}$ is the number of 'recently' infected individuals in the population. This parameter is estimated by

$$\begin{aligned} \widehat{ID} &= \frac{n_{R,1} - \hat{\beta}_T n_{+,1}}{n_{S,1} \cdot (\widehat{\Omega}_T - \hat{\beta}_T T)} - \frac{n_{R,2} - \hat{\beta}_T n_{+,2}}{n_{S,2} \cdot (\widehat{\Omega}_T - \hat{\beta}_T T)} \\ &= \frac{1}{\widehat{\Omega}_T - \hat{\beta}_T T} \cdot \left(\frac{n_{R,1} - \hat{\beta}_T n_{+,1}}{n_{S,1}} - \frac{n_{R,2} - \hat{\beta}_T n_{+,2}}{n_{S,2}} \right). \end{aligned} \tag{A62}$$

For $\widehat{ID} \sim N(ID, \sigma^2_{\widehat{ID}})$, the two sided p-value is

$$p = 2 \cdot \left(1 - \Phi \left(\left| \frac{\widehat{ID} - ID_o}{\sigma_{\widehat{ID},0}} \right| \right) \right), \tag{A63}$$

where $\Phi(.)$ is the cumulative distribution function of a standard normal random variable, and ID_0 and $\sigma^2_{\widehat{ID},0}$ are the expected value and variance of $\widehat{ID}$ under the null hypothesis. Here, the null hypothesis is that HIV incidence is the same in the two populations, and therefore $ID_0 = 0$. Assuming equal HIV prevalence and tests characteristics, under the null hypothesis the proportions of susceptible, infected, 'recently' infected and 'non-recently' infected individuals, $P_{S,k}$, $P_{+,k}$, $P_{R,k}$ and $P_{NR,k}$, respectively, are the same in the two populations. The variance of the incidence difference under the null hypothesis, $\sigma_{\widehat{ID},0}$, is estimated (see Section A.1) by:

$$\hat{\sigma}^2_{\widehat{ID},0} = \left(\frac{\hat{P}_{R,0} - \hat{\beta}_T \hat{P}_{+,0}}{\hat{P}_{S,0} \cdot (\widehat{\Omega}_T - \hat{\beta}_T T)}\right)^2 \frac{1}{\hat{P}_{+,0}} \cdot \left(\frac{1}{\hat{P}_{S,0}} + \frac{\hat{P}_{R,0}\hat{P}_{NR,0}}{\left(\hat{P}_{R,0} - \hat{\beta}_T \hat{P}_{+,0}\right)^2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \quad \text{(A64)}$$

where

$$\hat{P}_{J,0} = \frac{n_{J,1} + n_{J,2}}{n_1 + n_2} \quad \text{(A65)}$$

is the estimated proportion of individuals in state J, $J \in \{S, +, R, NR\}$. Substituting $ID_0 = 0$ and Equation (A64) into their corresponding terms in Equation (A63), the reported p-value is

$$\tilde{p} = 2 \cdot \left(1 - \Phi\left(\left|\frac{\widehat{ID}}{\hat{\sigma}_{\widehat{ID},0}}\right|\right)\right). \quad \text{(A66)}$$

The variance for the incidence difference was obtained by expressing sample counts and estimated test characteristics as functions of independently distributed standard normal random variables α_i ($i = 1,2 \ldots,6$) as follows:

$$n_{S,1} = n_{S,1}(\alpha_1) = n_1 P_{S,1} + \sigma_{S,1}\alpha_1 \quad \text{(A67)}$$

$$n_{R,1} = n_{R,1}(\alpha_1, \alpha_2) = n_1 P_{R,1} - \sigma_{R,1}\alpha_1 + \sigma_{R,NR,1}(\alpha_1)\alpha_2 \quad \text{(A68)}$$

$$n_{NR,1} = n_{NR,1}(\alpha_1, \alpha_2) = n_1 P_{NR,1} - \sigma_{NR,1}\alpha_1 - \sigma_{R,NR,1}(\alpha_1)\alpha_2 \quad \text{(A69)}$$

$$n_{S,2} = n_{S,2}(\alpha_3) = n_2 P_{S,2} + \sigma_{S,2}\alpha_3 \quad \text{(A70)}$$

$$n_{R,2} = n_{R,2}(\alpha_3, \alpha_4) = n_2 P_{R,2} - \sigma_{R,2}\alpha_3 + \sigma_{R,NR,2}(\alpha_3)\alpha_4 \quad \text{(A71)}$$

$$n_{NR,2} = n_{NR,2}(\alpha_3, \alpha_4) = n_2 P_{NR,2} - \sigma_{NR,2}\alpha_3 - \sigma_{R,NR,2}(\alpha_3)\alpha_4 \quad \text{(A72)}$$

$$\hat{\beta}_T = \hat{\beta}_T(\alpha_5) = \beta_T + \sigma_{\hat{\beta}_T}\alpha_5 \quad \text{(A73)}$$

$$\widehat{\Omega}_T = \widehat{\Omega}_T(\alpha_6) = \Omega_T + \sigma_{\widehat{\Omega}_T}\alpha_6, \quad \text{(A74)}$$

where $P_{S,k}$ $P_{R,k}$, $P_{NR,k}$ are the proportions of susceptible, 'recently' infected and 'non-recently' infected individuals in population k respectively, $P_{S,k} + P_{R,k} + P_{NR,k} = 1$; $\sigma_{\widehat{\beta}_T}$ and $\sigma_{\widehat{\Omega}_T}$ are the standard deviations of the (unbiased) estimators for β_T and Ω_T respectively; and

$$\sigma_{S,k} = \sqrt{n_k P_{S,k} \cdot \left(1 - P_{S,k}\right)} \tag{A75}$$

$$\sigma_{R,k} = \frac{P_{R,k}}{P_{R,k} + P_{NR,k}} \sigma_{S,k} \tag{A76}$$

$$\sigma_{NR,k} = \frac{P_{NR,k}}{P_{R,k} + P_{NR,k}} \sigma_{S,k} \tag{A77}$$

$$\sigma_{R,NR,k}(\alpha) = \frac{\sqrt{\left(n_k - n_k P_{S,k} - \sigma_{S,k}\alpha\right) P_{R,k} P_{NR,k}}}{P_{R,k} + P_{NR,k}}. \tag{A78}$$

The delta method then provides the following approximation for the *variance* of $\widehat{ID}$:

$$\begin{aligned}
\sigma^2_{\widehat{ID}} \approx & \left(\frac{P_{R,1} - \beta_T P_{+,1}}{P_{S,1} \cdot (\Omega_T - \beta_T T)}\right)^2 \frac{1}{n_1 P_{+,1}} \cdot \left(\frac{1}{P_{S,1}} + \frac{P_{R,1} P_{NR,1}}{\left(P_{R,1} - \beta_T P_{+,1}\right)^2}\right) \\
& + \left(\frac{P_{R,2} - \beta_T P_{+,2}}{P_{S,2} \cdot (\Omega_T - \beta_T T)}\right)^2 \frac{1}{n_2 P_{+,2}} \cdot \left(\frac{1}{P_{S,2}} + \frac{P_{R,2} P_{NR,2}}{\left(P_{R,2} - \beta_T P_{+,2}\right)^2}\right) \\
& + \sigma^2_{\widehat{\Omega}_T} \cdot \left(\frac{1}{\Omega_T - \beta_T T}\right)^4 \left(\frac{P_{R,1} - \beta_T P_{+,1}}{P_{S,1}} - \frac{P_{R,2} - \beta_T P_{+,2}}{P_{S,2}}\right)^2 \\
& + \sigma^2_{\widehat{\beta}_T} \cdot \left(\frac{1}{\Omega_T - \beta_T T}\right)^4 \left(\frac{P_{+,1}\Omega_T - P_{R,1} T}{P_{S,1}} - \frac{P_{+,2}\Omega_T - P_{R,2} T}{P_{S,2}}\right)^2.
\end{aligned} \tag{A79}$$

This tool aims to provide a simple interpretation of an observed difference between two incidence estimates. However, the statistic used to test for a difference in incidence should be optimised. Under the assumption of equal HIV prevalence and test characteristics, fluctuations in observed HIV prevalence or the estimated test characteristics create unnecessary variability in the incidence difference, and the difference between the proportions of 'recently' infected individuals *among those* individuals who are infected, which directly captures differences in only incidence, could instead be considered.

A6 Power to Detect Difference Calculator

Objective

This tool calculates the 'power' to correctly reject the null hypothesis of equal incidence in two study populations, when considering the difference between two incidence estimates, from input test characteristics, epidemiological context, sample sizes for the incidence surveys that would be performed in the study populations, and the significance level used for testing for a difference in incidence. Test characteristics are the same in the two populations, and equal HIV prevalence is also assumed.

Inputs

- Post-infection time cut-off T
- Estimated (or assumed) test characteristics[45]
  - Estimated MDRI $\widehat{\Omega}_T$
  - Estimated FRR $\hat{\beta}_T$
  - Estimated CoV of MDRI estimator $\hat{c}_{\widehat{\Omega}_T}$
  - Estimated CoV of FRR estimator $\hat{c}_{\hat{\beta}_T}$
- Assumed epidemiological context
  - HIV incidence I_k in study population k $(k = 1,2)$, where $I_1 > I_2$
  - HIV prevalence P (the same in both populations)
- The survey size n_k in study population k $(k = 1,2)$
- Significance level α

'Power' to reject null hypothesis of equal incidence, ϑ

This calculator is closely related to the *P-value for Difference Calculator* described in Section A5. The parameter about which inference is of interest is again the incidence difference, provided in Equation (A61), and the estimator for the parameter and its uncertainty are provided in Equations (A62) and (A79) respectively. While a p-value

[45]The current version of the calculator assumes $\hat{c}_{\widehat{\Omega}_T} = 0$ and $\hat{c}_{\hat{\beta}_T} = 0$, but these uncertainties are explicitly accounted for here, for completeness.

(probability of obtaining a difference as extreme as that observed, under a null hypothesis of equal incidence) was previously calculated, here an inferential 'power' (probability of obtaining a small p-value in an assumed context where the null hypothesis is in fact false) is provided. This tool aims to calculate the probability of *correct* inference, or, more specifically, the probability of *both*

(i) rejecting the null hypothesis of equal incidence in a two-tailed hypothesis test based on the incidence difference (that is, observing a two-sided p-value below the specified significance level); *and*

(ii) observing an incidence difference in the correct direction (that is, obtaining a higher incidence estimate for population 1 than for population 2).

The 'power' defined here would be smaller than the power conventionally calculated in statistics, which captures the probability of (i) above alone. The decrease in 'power' is expected to be negligible, unless the difference between incidence values in the two study populations is particularly small.

For $\widehat{ID} \sim N(ID, \sigma_{\widehat{ID}})$, the 'power' is given by:

$$\begin{aligned} \vartheta &= P\left(\frac{\widehat{ID}-ID_0}{\sigma_{\widehat{ID},0}} > z_{1-\alpha/2} \middle| \widehat{ID} \sim N\left(ID_1, \sigma^2_{\widehat{ID},1}\right)\right) \\ &= \Phi\left(\frac{z_{1-\alpha/2}\sigma_{\widehat{ID},0} + ID_0 - ID_1}{\sigma_{\widehat{ID},1}}\right), \end{aligned} \tag{A80}$$

where $\Phi(.)$ is the cumulative distribution function of a standard normal random variable, $z_q = \Phi^{-1}(q)$ is the lower q^{th} percentile of the standard normal distribution, $ID_0 = 0$ and $\sigma^2_{\widehat{ID},0}$ are the (presumed) expected value and variance of the incidence difference estimator under the null hypothesis of equal incidence, and ID_1 and $\sigma^2_{\widehat{ID},1}$ are the expected value and variance of the incidence difference estimator in reality.

For the input epidemiological context, in reality, the proportions of uninfected, infected, 'recently' infected and 'non-recently' infected individuals in population k ($k = 1,2$) are

$$\tilde{P}_{S,k} = 1 - P \tag{A81}$$

$$\tilde{P}_{+,k} = P \tag{A82}$$

$$\tilde{P}_{R,k} = I_k \cdot (1-P)\left(\hat{\Omega}_T - \hat{\beta}_T T\right) + \hat{\beta}_T P \tag{A83}$$

$$\tilde{P}_{NR,k} = P - \tilde{P}_{R,k}. \tag{A84}$$

The variance of the incidence difference estimator in reality, $\sigma^2_{\widehat{ID},1}$, is therefore estimated by

$$\begin{aligned}\tilde{\sigma}^2_{\widehat{ID},1} &= \left(\frac{\tilde{P}_{R,1} - \hat{\beta}_T\tilde{P}_{+,1}}{\tilde{P}_{S,1}\cdot(\widehat{\Omega}_T - \hat{\beta}_T T)}\right)^2 \frac{1}{n_1\tilde{P}_{+,1}}\cdot\left(\frac{1}{\tilde{P}_{S,1}} + \frac{\tilde{P}_{R,1}\tilde{P}_{NR,1}}{\left(\tilde{P}_{R,1} - \hat{\beta}_T\tilde{P}_{+,1}\right)^2}\right)\\ &+ \left(\frac{\tilde{P}_{R,2} - \hat{\beta}_T\tilde{P}_{+,2}}{\tilde{P}_{S,2}\cdot(\widehat{\Omega}_T - \hat{\beta}_T T)}\right)^2 \frac{1}{n_2\tilde{P}_{+,2}}\cdot\left(\frac{1}{\tilde{P}_{S,2}} + \frac{\tilde{P}_{R,2}\tilde{P}_{NR,2}}{\left(\tilde{P}_{R,2} - \hat{\beta}_T\tilde{P}_{+,2}\right)^2}\right)\\ &+ \hat{\sigma}^2_{\widehat{\Omega}_T}\cdot\left(\frac{1}{\widehat{\Omega}_T - \hat{\beta}_T T}\right)^4\left(\frac{\tilde{P}_{R,1} - \hat{\beta}_T\tilde{P}_{+,1}}{\tilde{P}_{S,1}} - \frac{\tilde{P}_{R,2} - \hat{\beta}_T\tilde{P}_{+,2}}{\tilde{P}_{S,2}}\right)^2\\ &+ \hat{\sigma}^2_{\hat{\beta}_T}\cdot\left(\frac{1}{\widehat{\Omega}_T - \hat{\beta}_T T}\right)^4\left(\frac{\tilde{P}_{+,1}\widehat{\Omega}_T - \tilde{P}_{R,1}T}{\tilde{P}_{S,1}} - \frac{\tilde{P}_{+,2}\widehat{\Omega}_T - \tilde{P}_{R,2}T}{\tilde{P}_{S,2}}\right)^2. \qquad \text{(A85)}\end{aligned}$$

By substituting Equations (A81) to (A84) into Equation (A85), the variance could instead be written directly in terms of input parameters:

$$\begin{aligned}\tilde{\sigma}^2_{\widehat{ID},1} &= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{P\hat{\beta}_T\cdot\left(1-\hat{\beta}_T\right)}{\left(\widehat{\Omega}_T - \hat{\beta}_T\right)^2(1-P)^2}\right)\\ &+ \left(\frac{I_1^2}{n_1} + \frac{I_2^2}{n_2}\right)\left(\frac{1}{1-P}\right)\\ &+ \left(\frac{I_1}{n_1} + \frac{I_2}{n_2}\right)\left(\frac{1-2\hat{\beta}_T}{\left(\widehat{\Omega}_T - \hat{\beta}_T T\right)(1-P)}\right)\\ &+ \hat{\sigma}^2_{\widehat{\Omega}_T}\cdot\left(\frac{1}{\widehat{\Omega}_T - \hat{\beta}_T T}\right)^2(I_1 - I_2)^2\\ &+ \hat{\sigma}^2_{\hat{\beta}_T}\cdot\left(\frac{T}{\widehat{\Omega}_T - \hat{\beta}_T T}\right)^2(I_1 - I_2)^2. \qquad \text{(A86)}\end{aligned}$$

Under the null hypothesis of equal HIV incidence, assuming equal HIV prevalence and test characteristics, the proportions of susceptible and infected individuals, $P_{S,k}$ and $P_{+,k}$, and 'recently' and 'non-recently' infected individuals, $P_{R,k}$ and $P_{NR,k}$, would be the same in the two populations. Therefore, the 'expected' estimate of $\sigma^2_{\widehat{ID},0}$ is

$$\tilde{\sigma}^2_{\widehat{ID},0} = \left(\frac{\tilde{P}_{R,0} - \hat{\beta}_T\tilde{P}_{+,0}}{\tilde{P}_{S,0}\cdot\left(\widehat{\Omega}_T - \hat{\beta}_T T\right)}\right)^2 \frac{1}{\tilde{P}_{+,0}}\cdot\left(\frac{1}{\tilde{P}_{S,0}} + \frac{\tilde{P}_{R,0}\tilde{P}_{NR,0}}{\left(\tilde{P}_{R,0} - \hat{\beta}_T\tilde{P}_{+,0}\right)^2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \qquad \text{(A87)}$$

where

$$\tilde{P}_{J,0} = \frac{n_1\tilde{P}_{J,1} + n_2\tilde{P}_{J,2}}{n_1 + n_2} \quad \text{(A88)}$$

is the expected estimate of $P_{J,0}$, which is understood to equal both $P_{J,1}$ and $P_{J,2}$ under the null hypothesis, for each state J, $J \in \{S, +, R, NR\}$.

Substituting $ID_0 = 0$, $ID_1 = I_1 - I_2$, and Equations (A85) and (A87) into their corresponding terms in Equation (A80), the reported 'power' is therefore

$$\tilde{\vartheta} = \Phi\left(\frac{z_{1-\alpha/2}\tilde{\sigma}_{\widehat{ID},0} - (I_1 - I_2)}{\tilde{\sigma}_{\widehat{ID},1}}\right). \quad \text{(A89)}$$

While this tool aims to provide straightforward guidance on the interpretation of a difference between two incidence estimates, further refinement of the analysis tool should include the optimisation of the statistic used to test for a difference between two incidence values.

A7 Test Performance Calculator

Objective

This tool calculates the performance of a test for recent infection, from input test characteristics, epidemiological context and the incidence survey size. The reported metric of performance is the CoV of the incidence estimator.

Inputs

- Post-infection time cut-off T
- Assumed test characteristics
  - MDRI Ω_T
  - FRR β_T
  - CoV of MDRI estimator $c_{\widehat{\Omega}_T}$
  - CoV of FRR estimator $c_{\widehat{\beta}_T}$
- Assumed epidemiological context
  - HIV incidence I
  - HIV prevalence P
- The survey size n

Coefficient of variation of the incidence estimator, $c_{\hat{I}_T}$

The CoV of the incidence estimator is reported as

$$\begin{aligned}\tilde{c}_{\hat{I}_T}^2 &= \frac{1}{n\tilde{P}_+}\cdot\left(\frac{1}{\tilde{P}_S}+\frac{\tilde{P}_R\tilde{P}_{NR}}{\left(\tilde{P}_R-\beta_T\tilde{P}_+\right)^2}\right)\\ &+\sigma_{\hat{\Omega}_T}^2\cdot\left(\frac{1}{\Omega_T-\beta_T T}\right)^2\\ &+\sigma_{\hat{\beta}_T}^2\cdot\left(\frac{\tilde{P}_+\Omega_T-\tilde{P}_R T}{\left(\tilde{P}_R-\beta_T\tilde{P}_+\right)(\Omega_T-\beta_T T)}\right)^2,\end{aligned} \tag{A90}$$

where

$$\tilde{P}_S = 1-P \tag{A91}$$

$$\tilde{P}_+ = P \tag{A92}$$

$$\tilde{P}_R = I\cdot(1-P)(\Omega_T-\beta_T T)+\beta_T P \tag{A93}$$

$$\tilde{P}_{NR} = P-\tilde{P}_R \tag{A94}$$

are the proportions of susceptible, infected, 'recently' infected and 'non-recently' infected individuals in the population for the input context (using input test characteristics). This result is based on Equation (A31).

It is recognised that there are some regimes in which the above closed-form approximation for the CoV would be poor. For example, for tests with particularly small FRRs, the FRR estimator's distribution is likely to deviate substantially from the assumed normal distribution, even when sample sizes are large for FRR estimation. Later versions of the tool could therefore utilise more general methods for estimating the CoV of the incidence estimator, such as bootstrap resampling [174]. Also, the current version of the calculator treats all inputs as known, hypothetical parameter values, and therefore does not report any uncertainty around the calculated CoV. In practice, when assessing the performance of an available test in a real-world population of interest, the inputs to the CoV calculation would be uncertain, and all such uncertainties should be propagated through the calculation (for example, to produce confidence intervals for the CoV of the incidence estimator). Approaches that could be used to account for the variability include those summarised in Section A1 (if using closed-form approximations for the CoV) or general, computationally-demanding approaches (such as bootstrap resampling).

Appendix B

Selected Matlab Code

Code blocks to reproduce some of the analyses presented in the main body of text are provided below. Programming was performed in Matlab (R2013b, 8.2.0.701, 64-bit), and the Statistic and Optimisation toolboxes were utilised.

The code provided in Section B1 explores bias in incidence estimation for modelled scenarios, and was used to produce Figure 2.2 in Section 2.2.1.

Code to assess the accuracy of the delta method approximation for the coefficient of variation of the incidence estimator, arising from variability in realised sample counts, is provided in Section B2. This code produces analysis outputs that support the discussion in Section 2.2.2.

In Section B3, a function for estimating the MDRI from longitudinal data is provided, and was used in the benchmarking exercise described in Section 4.1. In particular, code for Methods 1 and 2 ('Interp_Linear_SE' and 'Interp_Linear_ME') of Figure 4.1 is presented, together with example input data and outputs.

B1 Bias of the Incidence Estimator

The function *fn_biascalc* calculates the true (weighting of recent) incidence that one aims to measure, the expected value of the incidence estimator, and the error terms γ_1, γ_2 and γ_3, in specified scenarios. The script *figure2p2*, which calls the function, produces Figure 2.2 (see Section 2.2.1 for more details of the analysis).

Function 'fn_biascalc.m'

```
function [I_T_true, I_T_hat, gamma1, gamma2, gamma3] = ...
   fn_biascalc (T,...
   H0, r_vec,...
   scenario_inc, I_vec, I_ref, ...
   musurv_vec, covsurv, ...
   T0, T1, frr_constant)

%
% Calculates (i) true incidence, (ii) expected value of the
%   incidence estimator, (iii) gamma 1, gamma 2 and gamma 3 error terms,
%   as a function of (a) population growth rate, (b) value to which
%   incidence increases or from which it decreases, (c) mean survival time
%
% Inputs
%   T: post- infection time cut-off in definition of incidence estimator
%   H0: susceptible population size at time t = 0 (time of survey)
%   r_vec: historical annual growth rate of population (a)
%   scenario_inc: 1 for increasing incidence / 2 for decreasing incidence
%   I_vec: incidence at t = 0 / incidence up until t = -T (b)
%   I_ref: incidence up until t = -T / incidence at t = 0
%   musurv_vec: mean survival time (c) (Weibull distribution)
%   covsurv: coefficient of variation of survival time (extreme values create instabilities)
%   T0: probability of being 'recent'|alive remains at 1 until time T0 after infection...
%   T1: ...then linearly decreases until time T1...
%   frr_constant: ...to a constant value of frr_constant, for all times larger than T1
% Inputs (a),(b),(c) are vectors (to capture multiple input values),
%   other inputs are scalars, time is in years
%
% Outputs
%   I_T_true: the true (weighted) incidence to be measured
%   I_T_hat: the expected value of the incidence estimator
%   gamma1, gamma2, gamma3: the first three (typically largest) error terms
% Outputs are three dimensional, dimension 1 captures (the different values
%   in) r_vec, 2 captures I_vec, and 3 captures musurv_vec
%

% Measure input dimensions

n_r = length(r_vec);
n_I = length(I_vec);
n_musurv = length(musurv_vec);

% Solve for shape and scale parameters of Weibull survival function
```

```
Solve_b = @(b) covsurv^2 - (gamma(1+2/b)-(gamma(1+1/b))^2)/(gamma(1+1/b))^2;
[b_est,~,~] = fzero(Solve_b, 4);
a_estvec = musurv_vec/gamma(1+1/b_est);

% Initialise results variables

gamma1 = NaN(n_r,n_I,n_musurv);      % gamma 1 varies with all dimensions
gamma2 = NaN(n_r,n_I,n_musurv);      % gamma 2 varies with dimension 1
gamma3 = NaN(n_r,n_I,n_musurv);      % gamma 3 varies with dimension 3
Omega_T = NaN(n_r,n_I,n_musurv);     % MDRI varies with dimension 3
beta_T = NaN(n_r,n_I,n_musurv);      % FRR varies with all dimensions
                                     % (in principle, although in this contrived example it is constant)
I_T_true = NaN(n_r,n_I,n_musurv);    % true weighted incidence varies with all dimensions
P_R = NaN(n_r,n_I,n_musurv);         % proportion of population 'recently' infected at t = 0
                                     % varies with all dimensions
P_NR = NaN(n_r,n_I,n_musurv);        % proportion of population 'non-recently' infected at t = 0
                                     % varies with all dimensions

% Perform calculations, looping over dimension 1, then 3, then 2

PRonA_ = @(x) 1-((x>T0)&(x<T1)).*(x-T0)*(1-frr_constant)/(T1-T0)-(x>=T1)*(1-frr_constant);
   % probability 'recent'|alive function
for d1 = 1:n_r
   r = r_vec(d1);
   NS_ = @(x) H0*(1+r).^x; % susceptible population size function
   delta_NS_ = @(x) NS_(x)/NS_(0)-1; % susceptible population size deviation function
   gamma2(d1,:,:) = integral(@(x) delta_NS_(x),-T,0)/T; % gamma 2 value
   for d3 = 1:n_musurv
      a_est = a_estvec(d3);
      PA_ = @(x) 1-wblcdf(x,a_est,b_est); % probability of being alive function
      PRA_ = @(x) PRonA_(x).*PA_(x); % probability of being 'recently' infected and alive function
      delta_PA_ = @(x) PA_(x)-1; % probability of being alive deviation function
      if d1==1
            gamma3(:,:,d3) = integral(@(x) delta_PA_(-x),-T,0)/T ; % gamma 3 value
            Omega_T(:,:,d3) = integral(@(x) PRA_(-x),-T,0); % MDRI value
      end
      max_surv = ceil(wblinv(0.99999,a_est,b_est));
      for d2 = 1:n_I
         if scenario_inc == 1
            I0 = I_ref; I1 = I_vec(d2);
         elseif scenario_inc == 2
            I0 = I_vec(d2); I1 = I_ref;
         end
         Inc_ = @(x) I0+((x<0)&(x>-T)).*(x+T)/T*(I1-I0)+(x>=0)*(I1-I0); % incidence function
         I_T_true(d1,d2,d3) = ...
            integral(@(x) Inc_(x).*NS_(x).*PRA_(-x),-T,0)/...
            integral(@(x) NS_(x).*PRA_(-x),-T,0); % true incidence value
         delta_I_ = @(x) Inc_(x)/I_T_true(d1, d2, d3)-1; % incidence deviation function
         gamma1(d1,d2,d3)  = integral(@(x) delta_I_(x),-T,0)/T; % gamma 1 value
         beta_T(d1,d2,d3) = ...
            integral(@(x) Inc_(x).*NS_(x).*PRA_(-x),-max_surv,-T) /...
            integral(@(x) Inc_(x).*NS_(x).*PA_(-x),-max_surv,-T); % FRR value
         N_R  = integral(@(x) Inc_(x).*NS_(x).*PRA_(-x),-max_surv,0); % Number 'recent' at t = 0
         N_NR = integral(@(x) Inc_(x).*NS_(x).*PA_(-x),-max_surv,0)-N_R;  % Number 'non-recent' at t = 0
         N_S = NS_(0); % Number susceptible at t = 0
         N = N_R+N_NR+N_S;
         P_R(d1,d2,d3) = N_R/N; % Proportion 'recent' at t = 0
         P_NR(d1,d2,d3) = N_NR/N; % Proportion 'non-recent' at t = 0
```

```
      end
   end
end
I_T_hat = (P_R-beta_T.*(P_NR+P_R))./((Omega_T-T*beta_T).*(1-P_NR-P_R));
   % Expected value of incidence estimator

end
```

Script 'figure2p2.m'

```
%
% Script to produce Figure 2.2,
%   calls on function fn_biascalc
%

% Inputs common to scenarios 1A, 1B, 2A, 2B

T = 1;
H0 = 1e6;
r_vec = -0.1:0.01:0.1;
I_vec = [0.01:0.0025:0.05];
I_ref = 0.01;
musurv_vec = 8;
covsurv = 0.5;
T0 = 0.25;
T1 = 0.75;

% Create figure window and plot relative bias for each scenario in turn

scrsz = get(0,'ScreenSize');
figure('OuterPosition',[1 1 scrsz(3)/2 scrsz(4)])

% Scenario 1A

subplot(2,2,1)
scenario_inc = 1;
frr_constant = 0.01;
[I_T_true,I_T_hat,~,~,~] = ...
   fn_biascalc(T,H0,r_vec,scenario_inc,I_vec,I_ref,musurv_vec,covsurv,T0,T1,frr_constant);
[C,h] = contour(100*(I_vec-I_ref)./I_ref,100*r_vec,100*(I_T_hat./I_T_true-1),100*(-0.1:0.005:0.1));
set(h,'LineWidth',3)
title(['Scenario 1A: Increasing incidence, {\it\beta_T}=' num2str(100*frr_constant) '%'])
ylabel('Annual growth of population size (%)')
clabel(C,h,'LabelSpacing',72*3)
set(gca,'FontName','Times','FontSize', 12)
set(gca,'YTick',-10:5:10)

% Scenario 1B

subplot(2,2,2)
scenario_inc = 1;
frr_constant = 0.05;
[I_T_true,I_T_hat,~,~,~] = ...
   fn_biascalc(T,H0,r_vec,scenario_inc,I_vec,I_ref,musurv_vec,covsurv,T0,T1,frr_constant);
[C,h] = contour(100*(I_vec-I_ref)./I_ref,100*r_vec,100*(I_T_hat./I_T_true-1),100*(-0.1:0.005:0.1));
set(h,'LineWidth',3)
title(['Scenario 1B: Increasing incidence, {\it\beta_T} =' num2str(100*frr_constant) '%'])
```

```
clabel(C,h,'LabelSpacing',72*3)
set(gca,'FontName','Times','FontSize', 12)
set(gca,'YTick',-10:5:10)
text(-50, -13, {['Percentage increase in incidence over the last {\itT} = ' num2str(T) ' year period,'];...
    ['from an equilibrium of ' num2str(I_ref*100) '% per annum']}, 'HorizontalAlignment','center')

 % Scenario 2A

subplot(2,2,3)
scenario_inc = 2;
frr_constant = 0.01;
[I_T_true,I_T_hat,~,~,~] = ...
   fn_biascalc(T,H0,r_vec,scenario_inc,I_vec,I_ref,musurv_vec,covsurv,T0,T1,frr_constant);
[C,h] = contour(-100*(I_ref-I_vec)./I_vec,100*r_vec,100*(I_T_hat./I_T_true-1),100*(-0.1:0.005:0.1));
set(h,'LineWidth',3)
title(['Scenario 2A: Decreasing incidence, {\it\beta_T} =' num2str(100*frr_constant) '%'])
ylabel('Annual growth of population size (%)')
clabel(C,h,'LabelSpacing',72*3)
set(gca,'FontName','Times','FontSize', 12)
set(gca,'YTick',-10:5:10)

% Scenario 2B

subplot(2,2,4)
scenario_inc = 2;
frr_constant = 0.05;
[I_T_true,I_T_hat,~,~,~] = ...
   fn_biascalc(T,H0,r_vec,scenario_inc,I_vec,I_ref,musurv_vec,covsurv,T0,T1,frr_constant);
[C,h] = contour(-100*(I_ref-I_vec)./I_vec,100*r_vec,100*(I_T_hat./I_T_true-1),100*(-0.1:0.005:0.1));
set(h,'LineWidth',3)
title(['Scenario 2B: Decreasing incidence, {\it\beta_T} =' num2str(100*frr_constant) '%'])
clabel(C,h,'LabelSpacing',72*3)
set(gca,'FontName','Times','FontSize', 12)
set(gca,'YTick',-10:5:10)
text(-10, -13, {['Percentage decrease in incidence over the last {\itT} = ' num2str(T) ' year period,'];...
    ['from an equilibrium to ' num2str(I_ref*100) '% per annum']},'HorizontalAlignment','center')

c = findall(gcf,'Type','text');
set(c, 'FontName', 'Times', 'Fontsize', 12)
```

B2 Accuracy of the Delta Method Uncertainty for the Incidence Estimator

The function *fn_cov* calculates the coefficient of variation (CoV) of the incidence estimator, arising from variability in realised sample counts of HIV-negative, 'recently' infected and 'non-recently' infected subjects in an incidence survey. The CoV is calculated (i) using the delta method approximation (based on Gaussian uncertainty of counts), and (ii) by directly enumerating all possible counts (based on a trinomial

distribution). The script *comparecov_section2p2p2* compares the two output CoVs for a selection of scenarios (see Section 2.2.2).

Function 'fn_cov.m'

```
function [CoVDelta, CoVTrinomial] = ...
   fn_cov (T, Inc_vec, PrevInc_vec, MDRI_vec, FRR_vec, N)

%
% Calculates the CoV of the incidence estimator,
%   resulting from uncertainty in sample counts
%   (i) using the delta method,
%   (ii) by directly enumerating all possible trinomial counts
%   as a function of (a) incidence, (b) prevalence to incidence ratio,
%   (c) MDRI, (d) FRR
%
% Inputs
%   T: post-infection time cut-off in definition of incidence estimator
%   Inc_vec: Incidence value (a)
%   PrevInc_vec: Prevalence to incidence ratio (b)
%   MDRI_vec: mean duration of recent infection (c)
%   FRR_vec: false-recent rate (d)
%   N: cross-sectional incidence survey size
% Inputs (a)-(d) are vectors (to capture multiple input values),
%   other inputs are scalars, time is in years
%
% Outputs
%   CoVDelta: CoV of incidence estimator, approximated using the delta method
%   CoVTrinomial: CoV of incidence estimator, enumerating all possible
%       trinomial draws
% Outputs are four-dimensional, dimension 1 captures (the different values
%   in) Inc_vec, 2 captures PrevInc_vec, 3 captures MDRI_vec, and
%   4 captures FRR_vec
%

% Measure input dimensions

n_Inc = length(Inc_vec);
n_PrevInc = length(PrevInc_vec);
n_MDRI = length(MDRI_vec);
n_FRR = length(FRR_vec);
n_total = n_Inc*n_PrevInc*n_MDRI*n_FRR;

% Initialise results variables

CoVDelta = NaN(n_Inc,n_PrevInc,n_MDRI,n_FRR);
CoVTrinomial = NaN(n_Inc,n_PrevInc,n_MDRI,n_FRR);

% Perform calculations, looping over dimensions 4 to 1

count = 0;
for d4 = 1:n_FRR
   FRR = FRR_vec(d4);
   for d3 = 1:n_MDRI
      MDRI = MDRI_vec(d3);
      for d2 = 1:n_PrevInc
```

```
        PrevInc = PrevInc_vec(d2);
        for d1 = 1:n_Inc
           Inc = Inc_vec(d1);
           Prev = Inc*PrevInc;
           P_S = 1-Prev; % proportion of population susceptible
           P_R = Inc*(1-Prev)*(MDRI-FRR*T)+FRR*Prev; % proportion of population 'recently' infected
           P_NR = Prev-P_R; % proportion of population 'non-recently' infected
           [N_S, N_R] = meshgrid(1:N,0:N); % exclude 0 susceptible counts (improbable)
           N_S = N_S(:); N_R = N_R(:);
           inclIndex = N_S+N_R<=N; N_S = N_S(inclIndex); N_R = N_R(inclIndex); N_NR = N-N_S-N_R;
           prob_sample = mnpdf([N_S N_R N_NR],[P_S P_R P_NR]); % probability of each draw of counts
           inc_sample = (N_R-FRR*(N_R+N_NR))./N_S./(MDRI-FRR*T);
              % incidence estimate for each draw of counts
           mean_inc_sample = sum(inc_sample.*prob_sample); % mean incidence estimate
           mean_inc_sample2 = sum((inc_sample.^2).*prob_sample); % mean squared incidence estimate
           CoVTrinomial(d1,d2,d3,d4) =  ...
              (mean_inc_sample2-mean_inc_sample^2)^0.5/mean_inc_sample; % true CoV
           CoVDelta(d1,d2,d3,d4) =  ...
              (1/N*1/(P_R+P_NR)*(1/P_S+P_R*P_NR/(P_R-FRR*(P_R+P_NR))^2))^0.5; % delta method CoV
           count = count+1;
           if mod(count,500)==0, disp(['Loop ' num2str(count) ' of '  num2str(n_total) ' complete']), end
        end
     end
  end
end
disp('Complete')

end
```

Script 'deltacov_section2p2p2.m'

```
%
% Compares the delta method CoV of the incidence estimator
%   to the true CoV,
%   not accounting for uncertainty in test characteristics
%

% Inputs

T = 1;
Inc_vec = (0.1:0.1:3)/100;
PrevInc_vec = 2:1:10;
MDRI_vec = (100:50:300)/365.25;
FRR_vec = 0:0.025:0.15;
N = 100;

% Perform calculations (function call)

[CoVDelta, CoVTrinomial] = ...
   fn_cov(T,Inc_vec,PrevInc_vec,MDRI_vec,FRR_vec,N);

% Summarise outputs by limits of differences

[min(CoVDelta(:)-CoVTrinomial(:)),...
   max(CoVDelta(:)-CoVTrinomial(:))]
```

B3 MDRI Estimation by Linear Interpolation

The function *fn_mdri_linearint* estimates the MDRI by linearly interpolating between data points (see Section 4.1.2 for more details). By setting the input 'se' to 1, single continuous sojourns in the 'recent' infection state are enforced (corresponding to Method 1, 'Interp_Linear_SE'), and by setting 'se' to 0, multiple transitions between the 'recent' and 'non-recent' infection states are allowed (corresponding to Method 2, 'Interp_Linear_ME'). The script *eg_mdri_linearint* presents an example application of the function. The outputs produced by the script are also summarised.

Function 'fn_mdri_linearint.m'

```
function [mdrihat, lowerci, upperci, covmdrihat, stopmsg] = ...
   fn_mdri_linearint (datatimes, datavals, dataints, ...
   thresholds, T, ...
   fig_ind, n_intervals,...
   starting_value, se, ...
   ci_ind, alphaci, n_bootstrap)

%
% Estimates MDRI for a biomarker for recent infection,
%   by linear interpolating between data points
%   as a function of (a) threshold (below which a measurement
%   indicates 'recent' infection)
%
% Inputs
%   datatimes: matrix of (increasing) visits times, one row per subject,
%       time of first HIV-positive visit is reference time 0 per subject,
%       NaNs for missing values / where data is absent
%   datavals: matrix of corresponding biomarker readings
%   dataints: vector capturing the interval from last HIV-negative visit
%       to first HIV-positive visit per subject
%   thresholds: readings below the threshold indicate 'recent' infections (a)
%       (multiple values can be provided in vector)
%   T: post-infection time cut-off contained in definition of MDRI
%   alphaci: confidence interval (CI) should provide (1- alphaci)*100% coverage
%       (multiple values can be provided in vector)
%   ci_ind: 1 to produce CIs (by bootstrap resampling, percentile CIs)
%   n_bootstrap: number of bootstrap samples for CI calculation
%   fig_ind: 1 to show estimated P_R(t) as a function of t
%   n_intervals: number of intervals into which [0,T] is divided when
%       integrating estimated P_R(t) using the (composite) trapezoidal rule
%   starting_value: assumed biomarker value at infection time (unless
%       a value is already provided in the data)
%   se: 1 for enforcing single (earliest) exits out of the 'recent' state
%
% Outputs
%   mdrihat: estimated MDRI, entry (i) for thresholds(i)
```

```
%   lowerci: lower CI limit (NaN if ci_ind is not 1),
%      entry (i,j) relates to thresholds(i) and alphaci(j)
%   upperci: upper CI limit (NaN if ci_ind is not 1),
%      entry (i,j) relates to thresholds(i) and alphaci(j)
%   covmdrihat: coefficient of variation of bootstrap MDRI estimate replicates
%      (NaN if ci_ind is not 1), entry (i) for thresholds(i)
%   stopmsg: 1 if data is insufficient (mdrihat, lowerci, upperci, covmdrihat then
%      contain NaN), 0 otherwise
%

% Data cleaning (remove empty panels and panels with missing HIV-negative to -positive visit gaps)

dataints = dataints(:);
inclindex = (sum(not(isnan(datavals)|isnan(datatimes)),2)>0)&not(isnan(dataints));
datatimes = datatimes(inclindex,:); datavals = datavals(inclindex,:); dataints = dataints(inclindex);

% Reshape and manipulate inputs, and measure their dimensions

alphaci = alphaci(:)';
n_alpha = length(alphaci);
n_times = n_intervals + 1;
time_boundaries = linspace(0,T,n_times); % times at which to evaluate P_R(t) (*)
thresholds = thresholds(:);
n_thresholds = length(thresholds);
thresholdmatrix = repmat(thresholds,1,n_times);
n_subjects = size(datatimes,1);
n_obs = size(datatimes,2);
datatimes_mp = datatimes + repmat(dataints/2,1,n_obs); % midpoints used as infection times

% For each threshold, time point in (*), and subject,
%   store whether the biomarker reading produces a 'recent' result (1)
%   or 'non-recent' result (0), or is unknown (NaN)

threshold_time_subject_class = NaN(n_thresholds, n_times, n_subjects);
for subjectcount = 1:n_subjects
   datatimes_ind = datatimes_mp(subjectcount,:); datavals_ind = datavals(subjectcount,:);
   inclindex = not(isnan(datatimes_ind)|isnan(datavals_ind));
   n_incl = sum(inclindex);
   datatimes_ind = datatimes_ind(inclindex); datavals_ind = datavals_ind(inclindex);
   if datatimes_ind(1) ~= 0
      datatimes_ind = [0 datatimes_ind]; datavals_ind = [starting_value datavals_ind];
      n_incl = n_incl + 1;
   end
   if n_incl > 1
      f = @(t) interp1q(datatimes_ind', datavals_ind', t')';
      threshold_time_subject_class(:,:,subjectcount) = ...
         repmat(f(time_boundaries),n_thresholds,1) <= thresholdmatrix;
      threshold_time_subject_class(:,isnan(f(time_boundaries)),subjectcount) = NaN;
   end
end

% If single exits, all classifications after each subject's first 'non-recent' result become 'non-recent'

if se == 1
   afternr = cumsum(threshold_time_subject_class==0,2)>0;
   threshold_time_subject_class(afternr) = 0;
end
```

```
% Measure the proportion of 'recently' infected subjects, among subjects not yet lost to follow-up,
%   at each time point in (*) and each threshold

n_recent = sum(threshold_time_subject_class==1,3);
n_total = sum(threshold_time_subject_class>=0,3);
P_recent = n_recent./n_total;

% Check that at least one subject provides data until T after (estimated) infection

if n_total(1,end) == 0       % Exit function due to insufficient data

   stopmsg = 1;
   mdrihat = NaN(n_thresholds,1);
   lowerci = NaN(n_thresholds,n_alpha);
   upperci = NaN(n_thresholds,n_alpha);
   covmdrihat = NaN(n_thresholds,1);

else                         % Continue with estimation

   stopmsg = 0;

   % Estimate MDRI using the composite trapezoidal rule

   mdrihat = (2*sum(P_recent,2)-P_recent(:,1)-P_recent(:,end))/2*(T/n_intervals);

   % Plot estimated P_R(t)

   if fig_ind == 1
      figure
      plot(time_boundaries', P_recent','LineWidth',2)
      legend(num2str(thresholds), 'location', 'eastoutside')
      xlabel('Time since infection','FontName','Times','FontSize',11)
      ylabel('Probability of testing "recent"','FontName','Times','FontSize',11)
      grid on
      set(gca,'FontName','Times','FontSize',11)
      set(gca,'YLim',[0 1])
      set(gca,'XLim',[0 T*1.05])
      hold on, plot([T T],[0 1],'k--','LineWidth',2), hold off
   end

   % Obtain confidence interval limits from percentiles of bootstrap MDRI estimate replicates

   if ci_ind == 1
      bootstrapmdri = NaN(n_thresholds,n_bootstrap);
      for bootstrapcount = 1:n_bootstrap
         bootstrapindex = randsample(1:n_subjects,n_subjects,true);
         bootstrap_class = threshold_time_subject_class(:,:,bootstrapindex);
         n_recent = sum(bootstrap_class==1,3);
         n_total = sum(bootstrap_class>=0,3);
         P_recent = n_recent./n_total;
         if n_total(1,end) ~= 0
            bootstrapmdri(:,bootstrapcount) = …
              (2*sum(P_recent,2)-P_recent(:,1)-P_recent(:,end))/2*(T/n_intervals);
         end
      end
      if n_thresholds > 1
         bootstrapmdri = bootstrapmdri(:,not(isnan(bootstrapmdri(1,:))));
      else
```

```
        bootstrapmdri = bootstrapmdri(not(isnan(bootstrapmdri)));
    end
    lowerci = prctile(bootstrapmdri,alphaci/2*100,2);
    upperci = prctile(bootstrapmdri,100-alphaci/2*100,2);
    covmdrihat = std(bootstrapmdri,[],2)./mean(bootstrapmdri,2);
  else
    lowerci = NaN(n_thresholds,n_alpha);
    upperci = NaN(n_thresholds,n_alpha);
    covmdrihat = NaN(n_thresholds,1);
  end

end

end
```

Script 'eg_mdri_linearint.m'

```
%
% Demonstrates application of fn_mdri_linearint
%

 % Dataset

datatimes = [
     0.2675    0.4277    0.6396    0.9542       NaN       NaN       NaN       NaN       NaN
     0.4241     0.615    0.8653    1.0678    1.4144    1.6966     1.976       NaN       NaN
     0.2171     0.447    0.7267       NaN       NaN       NaN       NaN       NaN       NaN
     0.3031     0.529     0.756    1.0553      1.31    1.5768     1.806       NaN       NaN
     0.1826    0.4799    0.6827    0.9863    1.2467    1.4974    1.7532    2.0181       NaN
     0.3068    0.5752    0.8277    0.9795      1.25    1.5843       NaN       NaN       NaN
     0.2287    0.4841    0.7531    0.9521    1.2125       NaN       NaN       NaN       NaN
     0.1636    0.4299       NaN       NaN       NaN       NaN       NaN       NaN       NaN
     0.2174     0.457    0.7753    1.0156    1.2352    1.4146    1.6271    1.8641    2.1053
     0.2427    0.4439    0.7162     1.007    1.2991       NaN       NaN       NaN       NaN
     0.2185    0.4832    0.7504     0.967    1.2367    1.4919    1.8043       NaN       NaN
     0.3116    0.5438    0.7895      1.07    1.4269    1.6547    1.8606       NaN       NaN
      0.313    0.5456    0.8284     1.054    1.3397    1.5123    1.7787       NaN       NaN
     0.2878    0.5382    0.7905    1.1004    1.3466       NaN       NaN       NaN       NaN
     0.2257    0.4099    0.6935    0.9032    1.1561    1.9865    2.1556       NaN       NaN
  ];
datavals = [
    10.6283    0.0125   20.0914   63.7114       NaN       NaN       NaN       NaN       NaN
     0.9405   12.4182   45.4135   62.6159    77.891   68.1201   79.9291       NaN       NaN
     1.8642   42.5928   55.0881       NaN       NaN       NaN       NaN       NaN       NaN
     3.0956   13.0108   37.8741   48.0773   63.1614   70.7904   77.8083       NaN       NaN
     0.9861       NaN   39.5436    63.925   68.7903   62.3159   70.5174   78.6165       NaN
     4.8589   21.5184   44.7078   32.5498   50.5489   70.3637       NaN       NaN       NaN
     2.1248   13.7792   50.1477   73.6649   69.6588       NaN       NaN       NaN       NaN
     0.8965    0.5121       NaN       NaN       NaN       NaN       NaN       NaN       NaN
     1.5689    3.2719   20.1856   44.3737   35.2546   61.3277    63.072   81.7462   77.4984
     0.2678    0.7413   25.6545   41.1602   75.0971       NaN       NaN       NaN       NaN
```

```
      4.818   18.2466    36.031   53.9034   57.0645   63.6801   69.9678       NaN       NaN
     9.7489   32.7693   52.5553   68.2775   72.4863   73.3452   70.2274       NaN       NaN
     2.5468   25.4565   10.6181   48.1057   69.4896    34.548   70.9096       NaN       NaN
     5.9428   20.1445    51.351   70.5699   89.5246       NaN       NaN       NaN       NaN
     2.0306   26.8103   65.1453   80.0954   72.5498       NaN       NaN       NaN       NaN
  ];
dataints = [
     0.1756
     0.2004
     0.2458
     0.1986
     0.1255
     0.2259
     0.2175
     0.2158
      0.163
     0.1872
     0.1438
     0.2203
     0.2211
        NaN
     0.2221
  ];

% Inputs for fn_mdri_linearint

thresholds = 30:10:50;
T = 1.5;
figind = 1;
n_intervals = 20000;
starting_value = 0;
ci_ind = 1;
alphaci = [0.05 0.2];
n_bootstrap = 10000;

% Single-exit estimation

se = 1;
[mdrihat,lowerci,upperci,covmdrihat,stopmsg] = ...
   fn_mdri_linearint(datatimes,datavals,dataints, ...
   thresholds,T, ...
   figind,n_intervals,...
   starting_value,se,...
   ci_ind,alphaci,n_bootstrap);
[ [mdrihat lowerci(:,1) upperci(:,1) lowerci(:,2) upperci(:,2)]*365.25 covmdrihat*100] % Results

% Multiple-transition estimation

se = 0;
[mdrihat,lowerci,upperci,covmdrihat,stopmsg] = ...
   fn_mdri_linearint (datatimes,datavals,dataints, ...
   thresholds,T, ...
   figind,n_intervals,...
```

```
    starting_value,se,...
    ci_ind,alphaci,n_bootstrap);
[ [mdrihat lowerci(:,1) upperci(:,1) lowerci(:,2) upperci(:,2)]*365.25 covmdrihat*100] % Results
```

Outputs from 'eg_mdri_linearint.m'

Outputs produced by the script file *eg_mdri_linearint.m* are summarised below. The results obtained when enforcing a single continuous sojourn in the 'recent' state are provided in Figure B1 and Table B1, and those obtained when allowing multiple transitions between states are provided in Figure B2 and Table B2.

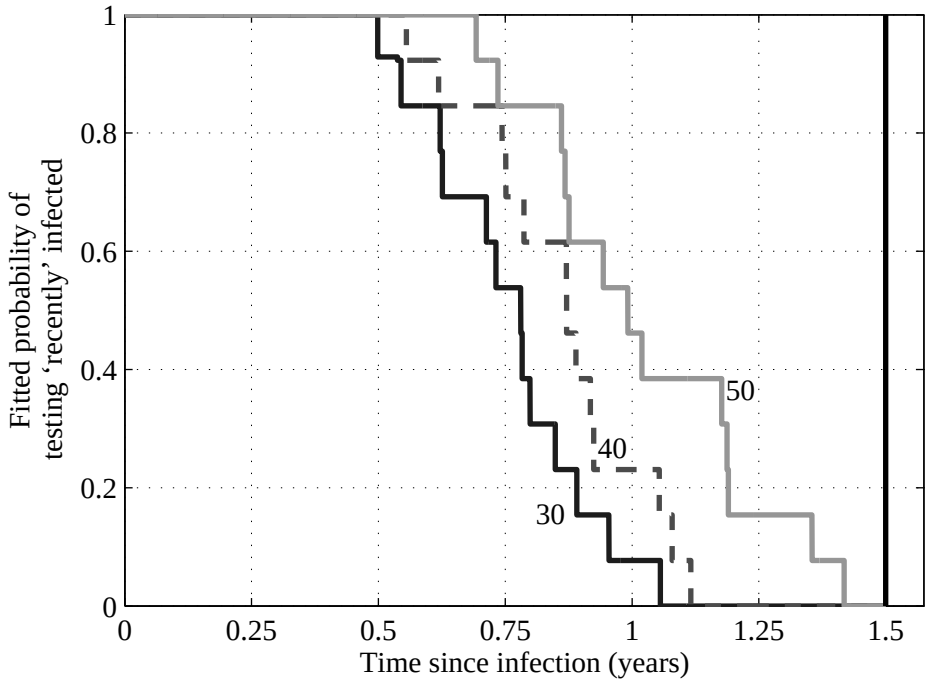


Figure B1: Example of a fitted probability of being 'recently' infected over time since infection, by threshold, using linear interpolation and enforcing single exits from the 'recent' state

The fitted probability of testing 'recently' infected is shown as a function of time since infection, and was obtained by applying the function *fn_mdri_linearint* to an example dataset, as outlined in the script *eg_mdri_linearint*. The probability is shown for a threshold of 30, 40 and 50 in turn (a biomarker reading below the threshold indicates 'recent' infection). Biomarker readings are linearly interpolated, and single exits from the 'recent' state are enforced.

Threshold	Estimated MDRI (days)	95% CI for MDRI (days)	80% CI for MDRI (days)	CoV of MDRI estimator (%)
30	277	246-308	257-298	5.7
40	314	281-346	293-336	5.3
50	374	331-419	346-403	6.0

Table B1: Example of MDRI estimation outputs, by threshold, using linear interpolation and enforcing single exits from the 'recent' state

Estimates of the MDRI (days) are shown, and were obtained by applying the function *fn_mdri_linearint* to an example dataset, as captured in the script *eg_mdri_linearint*. Confidence intervals (CIs) for the MDRI (days) and estimated coefficients of variation (CoVs) of the MDRI estimator (%) are also provided. The MDRI was estimated for a threshold of 30, 40 and 50 in turn (a biomarker reading below the threshold indicates 'recent' infection), and $T = 1.5$ years. Biomarker readings were linearly interpolated, and single exits from the 'recent' state were enforced.

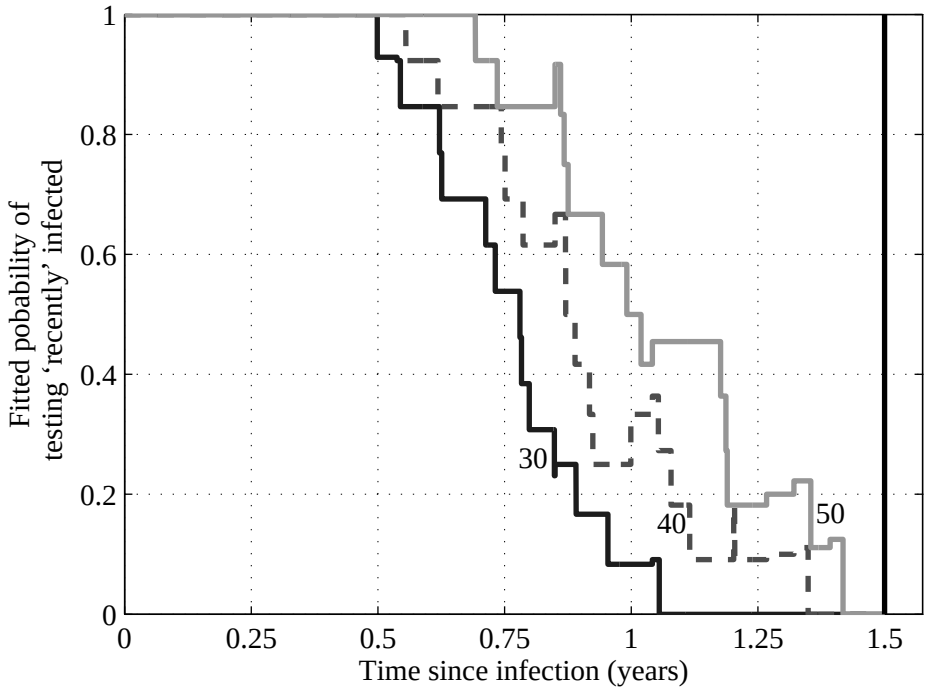


Figure B2: Example of a fitted probability of being 'recently' infected over time since infection, by threshold, using linear interpolation and allowing multiple transitions between 'recent' and 'non-recent' states

The fitted probability of testing 'recently' infected is shown as a function of time since infection, and was obtained by applying the function *fn_mdri_linearint* to an example dataset, as outlined in the script file *eg_mdri_linearint*. The probability is shown for a threshold of 30, 40 and 50 in turn (a biomarker reading below the threshold indicates 'recent' infection). Biomarker readings are linearly interpolated, and multiple transitions between the 'recent' and 'non-recent' states are allowed.

Threshold	Estimated MDRI (days)	95% CI for MDRI (days)	80% CI for MDRI (days)	CoV of MDRI estimator (%)
30	278	247-308	258-298	5.6
40	328	290-365	303-353	5.9
50	385	338-429	353-414	6.1

Table B2: Example of MDRI estimation outputs, by threshold, using linear interpolation and allowing multiple transitions between states

Estimates of the MDRI (days) are shown, and were obtained by applying the function *fn_mdri_linearint* to an example dataset, as captured in the script *eg_mdri_linearint*. Confidence intervals (CIs) for the MDRI (days) and estimated coefficients of variation (CoVs) of the MDRI estimator (%) are also provided. The MDRI was estimated for a threshold of 30, 40 and 50 in turn (a biomarker reading below the threshold indicates 'recent' infection), and $T = 1.5$ years. Biomarker readings were linearly interpolated, and multiple transitions between the 'recent' and 'non-recent' states were allowed.

References

1. Gallo RC. Historical essay: The early years of HIV/AIDS. *Science* 2002, 298(5599):1728-1730.
2. Montagnier L. Historical essay: A history of HIV discovery. *Science* 2002, 298(5599):1727-1728.
3. Gallo RC, Montagnier L. The discovery of HIV as the cause of AIDS. *N Engl J Med* 2003, 349(24):2283-2285.
4. Joint United Nations Programme on HIV/AIDS (UNAIDS). Global Report: UNAIDS Report on the Global AIDS Epidemic 2013. 2013.
5. The South African Department of Health. The 2011 National Antenatal Sentinel HIV and Syphilis Prevalence Survey. 2012.
6. Rehle T, Lazzari S, Dallabetta G, Asamoah-Odei E. Second-generation HIV surveillance: Better data for decision-making. *Bull World Health Organ* 2004, 82(2):121-127.
7. Brookmeyer R, Quinn TC. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *Am J Epidemiol* 1995, 141(2):166-172.
8. Le Vu S, Pillonel J, Semaille C, Bernillon P, Le Strat Y, Meyer L, Desenclos JC. Principles and uses of HIV incidence estimation from recent infection testing – a review. *Euro Surveill* 2008, 13(36):11-16.
9. Murphy G, Parry JV. Assays for the detection of recent infections with human immunodeficiency virus type 1. *Euro Surveill* 2008, 13(36):4-10.
10. Busch MP, Pilcher CD, Mastro TD, Kaldor J, Vercauteren G, Rodriguez W, Rousseau C, Rehle TM, Welte A, Averill MD, Garcia Calleja JM. Beyond detuning: 10 years of progress and new challenges in the development and application of assays for HIV incidence estimation. *AIDS* 2010, 24(18):2763-2771.
11. Welte A, McWalter TA, Laeyendecker O, Hallett TB. Using tests for recent infection to estimate incidence: Problems and prospects for HIV. *Euro Surveill* 2010, 15(24):pii=19589.
12. Mastro TD, Kim AA, Hallett T, Rehle T, Welte A, Laeyendecker O, Oluoch T, Garcia-Calleja JM. Estimating HIV incidence in populations using tests for recent infection: Issues, challenges and the way forward. *J HIV AIDS Surveill Epidemiol* 2010, 2(1):1-14.
13. UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance. When and How to Use Assays for Recent Infection to Estimate HIV Incidence at a Population Level. http://www.who.int/diagnostics_laboratory/hiv_incidence_may13_final.pdf. 2011.

14. Incidence Assay Critical Path Working Group. More and better information to tackle HIV epidemics: Towards improved HIV incidence assays. *PLOS Med* 2011, 8(6):e1001045.
15. Kaplan EH, Brookmeyer R. Snapshot estimators of recent HIV incidence rates. *Operations Research* 1999, 47:29-37.
16. McDougal JS, Parekh BS, Peterson ML, Branson BM, Dobbs T, Ackers M, Gurwith M. Comparison of HIV type 1 incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the BED capture enzyme immunoassay. *AIDS Res Hum Retroviruses* 2006, 22(10):945-952.
17. Hargrove JW, Humphrey JH, Mutasa K, Parekh BS, McDougal JS, Ntozini R, Chidawanyika H, Moulton LH, Ward B, Nathoo K, Iliff PJ, Kopp E. Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. *AIDS* 2008, 22(4):511-518.
18. Brookmeyer R. Should biomarker estimates of HIV incidence be adjusted? *AIDS* 2009, 23(4):485-491.
19. Hargrove JW. BED estimates of HIV incidence must be adjusted. *AIDS* 2009, 23(15):2061-2062; author reply 2066-2068.
20. McDougal JS. BED estimates of HIV incidence must be adjusted. *AIDS* 2009, 23(15):2064-2065; author reply 2066-2068.
21. Welte A, McWalter TA, Bärnighausen T. Reply to 'Should biomarker estimates of HIV incidence be adjusted?'. *AIDS* 2009, 23(15):2062-2063; author reply 2066-2068.
22. Brookmeyer R. Response to correspondence on 'Should biomarker estimates of HIV incidence be adjusted?'. *AIDS* 2009, 23(15):2066-2068.
23. McWalter TA, Welte A. A comparison of biomarker based incidence estimators. *PLOS One* 2009, 4(10):e7368.
24. Wang R, Lagakos SW. On the use of adjusted cross-sectional estimators of HIV incidence. *JAIDS* 2009, 52(5):538-547.
25. McWalter TA, Welte A. Relating recent infection prevalence to incidence with a sub-population of assay non-progressors. *J Math Biol* 2010, 60(5):687-710.
26. Bärnighausen T, McWalter TA, Rosner Z, Newell ML, Welte A. HIV incidence estimation using the BED capture enzyme immunoassay: Systematic review and sensitivity analysis. *Epidemiology* 2010, 21(5):685-697.
27. Brookmeyer R. On the statistical accuracy of biomarker assays for HIV incidence. *JAIDS* 2010, 54(4):406-414.
28. UNAIDS Reference Group on Estimates, Modelling and Projections. Statement on the use of the BED assay for the estimation of HIV-1 incidence for surveillance or epidemic monitoring. *Wkly Epidemiol Rec* 2006, 81(4):40.
29. Kassanjee R, McWalter TA, Bärnighausen T, Welte A. A new general biomarker-based incidence estimator. *Epidemiology* 2012, 23(5):721-728.
30. Kassanjee R, McWalter TA, Welte A. Defining optimality of a test for recent infection for HIV incidence surveillance. *AIDS Res Hum Retroviruses* 2014, 30(1):45-49.
31. Kassanjee R, Welte A, McWalter TA, Keating SM, Vermeulen M, Stramer SL, Busch MP. Seroconverting blood donors as a resource for characterising and optimising recent infection testing algorithms for incidence estimation. *PLOS One* 2011, 6(6):e20027.

32. Kassanjee R, Pilcher CD, Keating SM, Facente SN, McKinney E, Price MA, Martin JN, Little S, Hecht FM, Kallas EG, Welte A, Busch MP, Murphy G: Independent assessment of candidate HIV incidence assays on specimens in the CEPHIA repository. *AIDS* 2014, 28(16):2439-2449.
33. Kassanjee R, Welte A, McWalter TA, Viljoen J, Bärnighausen T, Newell ML, Fatti LP. Calibration of BED assay for use in incidence estimation. E-poster CDB018 at the *5th IAS Conference on HIV Pathogenesis, Treatment and Prevention*, 19-22 July 2009, South Africa.
34. Kassanjee R, Hargrove J, Marinda E, Humphrey J, McWalter TA, Welte A. New criteria for defining biomarker-derived 'recent HIV infection' for the purposes of incidence estimation. E-poster CDC0474 at the *XVIII International AIDS Conference*,18-23 July 2010, Austria.
35. Kassanjee R, Welte A, Jehuda-Cohen T. SMARTube^{TM} as a test for recent infection. Poster 41 and presentation at the *2010 HIV Diagnostics Conference*, 24 -26 March 2010, Florida.
36. Welte A, Kassanjee R. Defining and measuring 'recent infection' and 'incidence'. Presentation at the *52nd Annual Conference of the South African Statistical Association*, 8-12 November 2010, South Africa.
37. Pilcher CD, Busch MP, Facente S, Kassanjee R, Keating S, Hecht FM, Deeks S, McKinney E, Murphy G, Kamali A, Amornkul P, Welte A, Martin J, Little S. CEPHIA: A specimen and data repository for independent evaluation of HIV incidence assays. Poster at the *2012 HIV Diagnostics Conference*, 12-14 December 2012, Atlanta, GA.
38. Pilcher CD, Busch MP, Welte A, Facente S, Kassanjee R, Keating SM, McKinney E, Amornkul P, Kamali A, Hecht FM, Martin J, Deeks S, Little S, Murphy G, on behalf of the Consortium for the Evaluation and Performance of HIV Incidence Assays (CEPHIA). A novel specimen and data repository for evaluation of HIV incidence assays. Poster and abstract TUPE192 at the *XIX International AIDS Conference*, July 22-27 2012, Washington, DC.
39. Pilcher CD, Kassanjee R, Motta L, Facente S, Keating SM, McKinney E, Gunasekera D, Hecht FM, Sperhacke RD, Santos B, Busch MP, Murphy G, Esfandiari J, Welte A. A novel quantitative, multi-analyte rapid test for recent HIV infection and HIV incidence estimation. Poster and abstract TUPE194 at the *XIX International AIDS Conference*, July 22-27 2012, Washington, DC.
40. Pilcher CD, Kassanjee R, Motta L, Facente S, Keating S, McKinney E, Hecht FM, Sperhacke RD, Busch MP, Murphy G, Esfandiari J, Welte A. DPP supplemental rapid test data can discriminate recent from longstanding HIV infection. Presentation at the *2012 HIV Diagnostics Conference*, 12-14 December 2012, Atlanta, GA.
41. Kassanjee R, Murphy G, Busch MP, Pilcher CD, McKinney E, Keating S, Facente S, MacArthur J, Welte A. The performance of candidate assays to detect recent HIV infection for cross-sectional incidence estimation: An independent, comparative evaluation. Poster 1056 / abstract X-168 at the *20th Conference on Retroviruses and Opportunistic Infections*, 3-6 March 2013, Atlanta, GA.
42. Kassanjee R, Murphy G, Busch MP, Pilcher CD, McKinney E, Keating S, Facente S, Welte A, on behalf of the Consortium for the Evaluation and Performance of HIV Incidence Assays (CEPHIA). Evaluation panel analysis

results. Presentation (satellite session) at the *20th Conference for Retroviruses and Opportunistic Infections*, 3-6 March 2013, Atlanta, GA.

43. Kassanjee R, Pilcher CD, Keating SM, Facente SN, Price MA, Martin JN, Little S, Welte A, Busch MP, Murphy G, on behalf of the Consortium for the Evaluation and Performance of HIV Incidence Assays (CEPHIA). Independent evaluation of predicate incidence assays for HIV surveillance. Poster 1005 and themed discussion presentation at the *Conference for Retroviruses and Opportunistic Infections 2014*, 3-6 March 2014, Boston, MA.
44. Keating SM, Murphy G, Welte A, Kassanjee R, Matten D, Facente SN, Lebedeva L, Busch MP, Pilcher CD, on behalf of the Consortium for the Evaluation and Performance of HIV Incidence Assays (CEPHIA). Application of the Geenius™ HIV1/2 Supplemental Assay for detection of recent HIV infections. Poster 615 at the *Conference for Retroviruses and Opportunistic Infections 2014*, 3-6 March 2014, Boston, MA.
45. Kassanjee R. Testing for recent infection to estimate HIV incidence from single cross-sectional surveys. *SACEMA Quarterly* Nov 2011.
46. Kassanjee R. 20th Conference for Retroviruses and Opportunistic Infections: Highlights, and developments in HIV incidence estimation. *SACEMA Quarterly* June 2013.
47. The Consortium for the Evaluation and Performance of HIV Incidence Assays (CEPHIA). http://www.incidence-estimation.com/page/cephia. Accessed 1 July 2014.
48. Incidence-Estimation, SACEMA's online resource for incidence estimation. http://www.incidence-estimation.com. Accessed 1 July 2014.
49. World Health Organization. Meeting report: Annual meeting of the WHO technical working group on HIV incidence assays, 26-27 September 2012, Geneva, Switzerland. http://apps.who.int/iris/bitstream/10665/86397/1/9789241506069_eng.pdf. 2013.
50. SACEMA Holds Training Course on Estimating HIV Cases. https://blogs.sun.ac.za/news/2012/05/23/sacema-holds-training-course-on-estimating-hiv-cases. Accessed 1 July 2014.
51. Centers for Disease Control and Prevention. http://www.cdc.gov. Accessed 1 July 2014.
52. WHO Technical Working Group on HIV Incidence Assays. http://www.who.int/diagnostics_laboratory/links/hiv_incidence_assay/en. Accessed 1 July 2014.
53. HIV Modelling Consortium Work Package on Characterisation of Tests for Recent Infection. http://www.hivmodelling.org/projects/incidence-estimation. Accessed 1 July 2014.
54. Brookmeyer R. Accounting for follow-up bias in estimation of human immunodeficiency virus incidence rates. *J R Stat Soc Ser A* 1997, 160(1):127-140.
55. Brookmeyer R. Measuring the HIV/AIDS epidemic: Approaches and challenges. *Epidemiol Rev* 2010, 32(1):26-37.
56. Podgor MJ, Leske MC. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Stat Med* 1986, 5(6):573-578.

57. Brunet RC, Struchiner CJ. A non-parametric method for the reconstruction of age- and time-dependent incidence from the prevalence data of irreversible diseases with differential mortality. *Theor Popul Biol* 1999, 56(1):76-90.
58. Williams B, Gouws E, Wilkinson D, Karim SA. Estimating HIV incidence rates from age prevalence data in epidemic situations. *Stat Med* 2001, 20(13):2003-2016.
59. Hallett TB, Zaba B, Todd J, Lopman B, Mwita W, Biraro S, Gregson S, Boerma JT. Estimating incidence from prevalence in generalised HIV epidemics: Methods and validation. *PLOS Med* 2008, 5(4):e80.
60. Mahiane GS, Ouifki R, Brand H, Delva W, Welte A. A general HIV incidence inference scheme based on likelihood of individual level data and a population renewal equation. *PLOS One* 2012, 7(9):e44377.
61. Shisana O, Rehle T, Simbayi LC, Zuma K, Jooste S, Zungu NP, Labadarios D, Onoya D, Davids A, Ramlagan S, Mbelle N, van Zyl J, Onoya D, Wabiri N. South African National HIV Prevalence, Incidence and Behaviour Survey, 2012. HSRC Press, 2014.
62. Bao L, Salomon JA, Brown T, Raftery AE, Hogan DR. Modelling national HIV/AIDS epidemics: Revised approach in the UNAIDS Estimation and Projection Package 2011. *Sex Transm Infect* 2012, 88 Suppl 2:i3-10.
63. Stover J, Brown T, Marston M. Updates to the Spectrum/Estimation and Projection Package (EPP) model to estimate HIV trends for adults and children. *Sex Transm Infect* 2012, 88 Suppl 2:i11-16.
64. United Nations AIDS (UNAIDS) Spectrum/EPP 2014. http://www.unaids.org/en/dataanalysis/datatools/spectrumepp2013. Accessed 1 July 2014.
65. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: Application to AIDS prediction in England and Wales. *Stat Med* 1997, 16(19):2191-2210.
66. Bellocco R, Marschner IC. Joint analysis of HIV and AIDS surveillance data in back-calculation. *Stat Med* 2000, 19(3):297-311.
67. Cui J, Becker NG. Estimating HIV incidence using dates of both HIV and AIDS diagnoses. *Stat Med* 2000, 19(9):1165-1177.
68. Posner SJ, Myers L, Hassig SE, Rice JC, Kissinger P, Farley TA. Estimating HIV incidence and detection rates from surveillance data. *Epidemiology* 2004, 15(2):164-172.
69. Case KK, Ghys PD, Gouws E, Eaton JW, Borquez A, Stover J, Cuchi P, Abu-Raddad LJ, Garnett GP, Hallett TB. Understanding the modes of transmission model of new HIV infection and its use in prevention planning. *Bull World Health Organ* 2012, 90(11):831-838A.
70. United Nations AIDS (UNAIDS) Incidence by Modes of Transmission. http://www.unaids.org/en/dataanalysis/datatools/incidencebymodesoftransmission. Accessed 1 July 2014.
71. Freeman J, Hutchison GB. Prevalence, incidence and duration. *Am J Epidemiol* 1980, 112(5):707-723.
72. Killian MS, Norris PJ, Rawal BD, Lebedeva M, Hecht FM, Levy JA, Busch MP. The effects of early antiretroviral therapy and its discontinuation on the HIV-specific antibody response. *AIDS Res Hum Retroviruses* 2006, 22(7):640-647.

73. Hayashida T, Gatanaga H, Tanuma J, Oka S. Effects of low HIV type 1 load and antiretroviral treatment on IgG-capture BED-enzyme immunoassay. *AIDS Res Hum Retroviruses* 2008, 24(3):495-498.
74. Laeyendecker O, Rothman RE, Henson C, Horne BJ, Ketlogetswe KS, Kraus CK, Shahan J, Kelen GD, Quinn TC. The effect of viral suppression on cross-sectional incidence testing in the Johns Hopkins Hospital emergency department. *JAIDS* 2008, 48(2):211-215.
75. Marinda ET, Hargrove J, Preiser W, Slabbert H, van Zyl G, Levin J, Moulton LH, Welte A, Humphrey J. Significantly diminished long-term specificity of the BED capture enzyme immunoassay among patients with HIV-1 with very low CD4 counts and those on antiretroviral therapy. *JAIDS* 2010, 53(4):496-499.
76. Hladik W, Olara D, Mermin J, Moore D, Were W, Alexander L, Downing R. Effect of CD4(+) T cell count and antiretroviral treatment on two serological HIV incidence assays. *AIDS Res Hum Retroviruses* 2012, 28(1):95-99.
77. Laeyendecker O, Brookmeyer R, Oliver AE, Mullis CE, Eaton KP, Mueller AC, Jacobson LP, Margolick JB, Brown J, Rinaldo CR, Quinn TC, Eshleman SH, for The Multicenter AIDS Cohort Study (MACS). Factors associated with incorrect identification of recent HIV infection using the BED capture immunoassay. *AIDS Res Hum Retroviruses* 2012, 28(8):816-822.
78. Wendel SK, Mullis CE, Eshleman SH, Blankson JN, Moore RD, Keruly JC, Brookmeyer R, Quinn TC, Laeyendecker O. Effect of natural and ARV-induced viral suppression and viral breakthrough on anti-HIV antibody proportion and avidity in patients with HIV-1 subtype B infection. *PLOS One* 2013, 8(2):e55525.
79. Longosz AF, Mehta SH, Kirk GD, Margolick JB, Brown J, Quinn TC, Eshleman SH, Laeyendecker O. Incorrect identification of recent HIV infection in adults in the United States using a limiting-antigen avidity assay. *AIDS* 2014, 28(8):1227-1232.
80. Longosz AF, Morrison CS, Chen PL, Arts E, Nankya I, Salata RA, Franco V, Quinn TC, Eshleman SH, Laeyendecker O. Immune responses in Ugandan women infected with Subtypes A and D HIV using the BED Capture Immunoassay and an antibody avidity assay. *JAIDS* 2014, 65(4):390-396.
81. Longosz AF, Serwadda D, Nalugoda F, Kigozi G, Franco V, Gray RH, Quinn TC, Eshleman SH, Laeyendecker O. Impact of HIV subtype on performance of the Limiting Antigen-Avidity Enzyme Immunoassay, the Bio-Rad Avidity Assay, and the BED Capture Immunoassay in Rakai, Uganda. *AIDS Res Hum Retroviruses* 2014, 30(4):339-344.
82. Welte A, McWalter TA, Bärnighausen T. A simplified formula for inferring HIV incidence from cross-sectional surveys using a test for recent infection. *AIDS Res Hum Retroviruses* 2009, 25(1):125-126.
83. Hubert JB, Burgard M, Dussaix E, Tamalet C, Deveau C, Le Chenadec J, Chaix ML, Marchadier E, Vilde JL, Delfraissy JF, Meyer L, Rouzioux. Natural history of serum HIV-1 RNA levels in 330 patients with a known date of infection – The SEROCO Study Group. *AIDS* 2000, 14(2):123-131.
84. Porter K, Babiker A, Bhaskaran K, Darbyshire J, Pezzotti P, Walker AS. Determinants of survival following HIV-1 seroconversion after the introduction of HAART. *Lancet* 2003, 362(9392):1267-1274.

85. Kitahata MM, Gange SJ, Abraham AG, Merriman B, Saag MS, Justice AC, Hogg RS, Deeks SG, Eron JJ, Brooks JT, Rourke SB, Gill MJ, Bosch RJ, Martin JN, Klein MB, Jacobson LP, Rodriguez B, Sterling TR, Kirk GD, Napravnik S, Rachlis AR, Calzavara LM, Horberg MA, Silverberg MJ, Gebo KA, Goedert JJ, Benson CA, Collier AC, Van Rompaey SE, Crane HM, McKaig RG, Lau B, Freeman AM, Moore RD. Effect of early versus deferred antiretroviral therapy for HIV on survival. *N Engl J Med* 2009, 360(18):1815-1826.
86. Curtis K, Kennedy S, Charurat ME, Nasidi A, Delaney K, Spira T, Owen M. Development and characterization of a bead-based, multiplex assay for estimation of recent HIV-1 infection. *AIDS Res Hum Retroviruses* 2012, 28(2):188-197.
87. McKeown K, Jewell NP. Current status observation of a three-state counting process with application to simultaneous accurate and diluted HIV test data. *Can J Stat* 2011, 39(3):475–487.
88. Laeyendecker O, Brookmeyer R, Cousins MM, Mullis CE, Konikoff J, Donnell D, Celum C, Buchbinder SP, Seage GR, Kirk GD, Mehta SH, Astemborski J, Jacobson LP, Margolick JB, Brown J, Quinn TC, Eshleman SH. HIV incidence determination in the United States: A multiassay approach. *J Infect Dis* 2013, 207(2):232-239.
89. Laeyendecker O, Piwowar-Manning E, Fiamma A, Kulich M, Donnell D, Bassuk D, Mullis CE, Chin C, Swanson P, Hackett Jr. J, Clarke W, Marzinke M, Szekeres G, Gray G, Richter L, Alexandre MW, Chariyalertsak S, Chingono A, Celentano DD, Morin SF, Sweat M, Coates T, Eshleman SH. Estimation of HIV incidence in a large, community-based, randomized clinical trial: NIMH Project Accept (HIV Prevention Trials Network 043). *PLOS One* 2013, 8(7):e68349.
90. Brookmeyer R, Laeyendecker O, Donnell D, Eshleman SH. Cross-sectional HIV incidence estimation in HIV prevention research. *JAIDS* 2013, 63 Suppl 2:S233-239.
91. Janssen RS, Satten GA, Stramer SL, Rawal BD, O'Brien TR, Weiblen BJ, Hecht FM, Jack N, Cleghorn FR, Kahn JO, Chesney MA, Busch MP. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. *JAMA* 1998, 280(1):42-48.
92. Parekh BS, Kennedy MS, Dobbs T, Pau CP, Byers R, Green T, Hu DJ, Vanichseni S, Young NL, Choopanya K, Mastro TD, McDougal JS. Quantitative detection of increasing HIV type 1 antibodies after seroconversion: A simple assay for detecting recent HIV infection and estimating incidence. *AIDS Res Hum Retroviruses* 2002, 18(4):295-307.
93. Wang R, Lagakos SW. Augmented cross-sectional prevalence testing for estimating HIV incidence. *Biometrics* 2009, 66(3):864-874.
94. Sweeting MJ, De Angelis D, Parry J, Suligoi B. Estimating the distribution of the window period for recent HIV infections: A comparison of statistical methods. *Stat Med* 2010, 29(30):3194-3202.
95. Braunstein SL, Nash D, Kim AA, Ford K, Mwambarangwe L, Ingabire CM, Vyankandondera J, van de Wijgert JH. Dual testing algorithm of BED-CEIA and AxSYM Avidity Index assays performs best in identifying recent HIV infection in a sample of Rwandan sex workers. *PLOS One* 2011, 6(4):e18402.
96. Parekh BS, Hanson DL, Hargrove J, Branson B, Green T, Dobbs T, Constantine N, Overbaugh J, McDougal JS. Determination of mean recency period for

estimation of HIV type 1 Incidence with the BED-capture EIA in persons infected with diverse subtypes. *AIDS Res Hum Retroviruses* 2011, 27(3):265-273.

97. Sommen C, Commenges D, Le Vu S, Meyer L, Alioum A. Estimation of the distribution of infection times using longitudinal serological markers of HIV: Implications for the estimation of HIV incidence. *Biometrics* 2011, 67(2):467-75.
98. Duong YT, Qiu M, De AK, Jackson K, Dobbs T, Kim AA, Nkengasong JN, Parekh BS. Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: Potential for HIV-1 incidence estimates and avidity maturation studies. *PLOS One* 2012, 7(3):e33328.
99. Hargrove J, Eastwood H, Mahiane G, van Schalkwyk C. How should we best estimate the mean recency duration for the BED method? *PLOS One* 2012, 7(11):e49661.
100. Keating SM, Hanson D, Lebedeva M, Laeyendecker O, Ali-Napo NL, Owen SM, Stramer SL, Moore RD, Norris PJ, Busch MP. Lower-sensitivity and avidity modifications of the Vitros anti-HIV 1+2 assay for detection of recent HIV infections and incidence estimation. *J Clin Microbiol* 2012, 50(12):3968-3976.
101. Brookmeyer R, Konikoff J, Laeyendecker O, Eshleman SH. Estimation of HIV incidence using multiple biomarkers. *Am J Epidemiol* 2013, 177(3):264-272.
102. Curtis KA, Hanson DL, Kennedy MS, Owen SM. Evaluation of a multiplex assay for estimation of HIV-1 incidence. *PLOS One* 2013, 8(5):e64201.
103. Mahiane SG, Fiamma A, Auvert B. Mixture models for calibrating the BED for HIV incidence testing. *Stat Med* 2014, 33(10):1767-1783.
104. Rawal BD, Degula A, Lebedeva L, Janssen RS, Hecht FM, Sheppard HW, Busch MP. Development of a new less-sensitive enzyme immunoassay for detection of early HIV-1 infection. *JAIDS* 2003, 33(3):349-355.
105. Kothe D, Byers RH, Caudill SP, Satten GA, Janssen RS, Hannon WH, Mei JV. Performance characteristics of a new less sensitive HIV-1 enzyme immunoassay for use in estimating HIV seroincidence. *JAIDS* 2003, 33(5):625-634.
106. Li H, Ketema F, Sill AM, Kreisel KM, Cleghorn FR, Constantine NT. A simple and inexpensive particle agglutination test to distinguish recent from established HIV-1 infection. *Int J Infect Dis* 2007, 11(5):459-465.
107. Parekh BS, McDougal JS. New approaches for detecting recent HIV-1 infection. *AIDS Review* 2001, 3:183-193.
108. Parekh BS, McDougal JS. Application of laboratory methods for estimation of HIV-1 incidence. *Indian J Med Res* 2005, 121(4):510-518.
109. Sakarovitch C, Rouet F, Murphy G, Minga AK, Alioum A, Dabis F, Costagliola D, Salamon R, Parry JV, Barin F. Do tests devised to detect recent HIV-1 infection provide reliable estimates of incidence in Africa? *JAIDS* 2007, 45(1):115-122.
110. Mullis CE, Munshaw S, Grabowski MK, Eshleman SH, Serwadda D, Brookmeyer R, Nalugoda F, Kigozi G, Kagaayi J, Tobian AA, Wawer M, Gray RH, Quinn TC, Laeyendecker O. Differential specificity of HIV incidence assays in HIV subtypes A and D-infected individuals from Rakai, Uganda. *AIDS Res Hum Retroviruses* 2013, 29(8):1146-1150.
111. Karita E, Price M, Hunter E, Chomba E, Allen S, Fei L, Kamali A, Sanders EJ, Anzala O, Katende M, Ketter N. Investigating the utility of the HIV-1 BED capture enzyme immunoassay using cross-sectional and longitudinal seroconverter specimens from Africa. *AIDS* 2007, 21(4):403-408.

112. Guy R, Gold J, Calleja JM, Kim AA, Parekh B, Busch M, Rehle T, Hargrove J, Remis RS, Kaldor JM. Accuracy of serological assays for detection of recent infection with HIV and estimation of population incidence: A systematic review. *Lancet Infect Dis* 2009, 9(12):747-759.
113. Wei X, Liu X, Dobbs T, Kuehl D, Nkengasong JN, Hu DJ, Parekh BS. Development of two avidity-based assays to detect recent HIV type 1 seroconversion using a multisubtype gp41 recombinant protein. *AIDS Res Hum Retroviruses* 2010, 26(1):61-71.
114. Konikoff J, Brookmeyer R, Longosz AF, Cousins MM, Celum C, Buchbinder SP, Seage GR, Kirk GD, Moore RD, Mehta SH, Margolick JB, Brown J, Mayer KH, Koblin BA, Justman JE, Hodder SL, Quinn TC, Eshleman SH, Laeyendecker O. Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of HIV incidence in the United States. *PLOS One* 2013, 8(12):e82772.
115. Suligoi B, Galli C, Massi M, Di Sora F, Sciandra M, Pezzotti P, Recchia O, Montella F, Sinicco A, Rezza G. Precision and accuracy of a procedure for detecting recent human immunodeficiency virus infections by calculating the antibody avidity index by an automated immunoassay-based method. *J Clin Microbiol* 2002, 40(11):4015-4020.
116. Suligoi B, Massi M, Galli C, Sciandra M, Di Sora F, Pezzotti P, Recchia O, Montella F, Sinicco A, Rezza G. Identifying recent HIV infections using the avidity index and an automated enzyme immunoassay. *JAIDS* 2003, 32(4):424-428.
117. Wilson KM, Johnson EI, Croom HA, Richards KM, Doughty L, Cunningham PH, Kemp BE, Branson BM, Dax EM. Incidence immunoassay for distinguishing recent from established HIV-1 infection in therapy-naive populations. *AIDS* 2004, 18(17):2253-2259.
118. Barin F, Meyer L, Lancar R, Deveau C, Gharib M, Laporte A, Desenclos JC, Costagliola D. Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use on dried serum spots. *J Clin Microbiol* 2005, 43(9):4441-4447.
119. Chawla A, Murphy G, Donnelly C, Booth CL, Johnson M, Parry JV, Phillips A, Geretti AM. Human immunodeficiency virus (HIV) antibody avidity testing to identify recent infection in newly diagnosed HIV type 1 (HIV-1)-seropositive persons infected with diverse HIV-1 subtypes. *J Clin Microbiol* 2007, 45(2):415-420.
120. Schupbach J, Gebhardt MD, Tomasik Z, Niederhauser C, Yerly S, Burgisser P, Matter L, Gorgievski M, Dubs R, Schultze D, Steffen I, Andreutti C, Martinetti G, Guntert B, Staub R, Daneel S, Vernazza P. Assessment of recent HIV-1 infection by a line immunoassay for HIV-1/2 confirmation. *PLOS Med* 2007, 4(12):e343.
121. Suligoi B, Rodella A, Raimondo M, Regine V, Terlenghi L, Manca N, Casari S, Camoni L, Salfa MC, Galli C. Avidity Index for anti-HIV antibodies: comparison between third- and fourth-generation automated immunoassays. *J Clin Microbiol* 2011, 49(7):2610-2613.
122. Andersson E, Shao W, Bontell I, Cham F, Cuong do D, Wondwossen A, Morris L, Hunt G, Sonnerborg A, Bertagnolio S, Maldarelli F, Jordan MR. Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1

infection in transmitted drug resistance surveys. *Infect Genet Evol* 2013, 18:125-131.

123. Cousins MM, Laeyendecker O, Beauchamp G, Brookmeyer R, Towler WI, Hudelson SE, Khaki L, Koblin B, Chesney M, Moore RD, Kelen GD, Coates T, Celum C, Buchbinder SP, Seage GR, Quinn TC, Donnell D, Eshleman SH. Use of a high resolution melting (HRM) assay to compare gag, pol, and env diversity in adults with different stages of HIV infection. *PLOS One* 2011, 6(11):e27211.
124. Ragonnet-Cronin M, Aris-Brosou S, Joanisse I, Merks H, Vallee D, Caminiti K, Rekart M, Krajden M, Cook D, Kim J, Malloch L, Sandstrom P, Brooks J. Genetic diversity as a marker for timing infection in HIV-infected patients: evaluation of a 6-month window and comparison with BED. *J Infect Dis* 2012, 206(5):756-764.
125. Park SY, Love TM, Nelson J, Thurston SW, Perelson AS, Lee HY. Designing a genome-based HIV incidence assay with high sensitivity and specificity. *AIDS* 2011, 25(16):F13-19.
126. Puren A, Takuva S. HIV incidence estimates for South Africa: Update on laboratory methods and post-test algorithms. *Communicable Diseases Surveillance Bulletin* 2013, 11(4):115-121.
127. Soroka SD, Granade TC, Candal D, Parekh BS. Modification of rapid human immunodeficiency virus (HIV) antibody assay protocols for detecting recent HIV seroconversion. *Clin Diagn Lab Immunol* 2005, 12(8):918-921.
128. Kshatriya R, Cachafeiro AA, Kerr RJ, Nelson JA, Fiscus SA. Comparison of two rapid human immunodeficiency virus (HIV) assays, Determine HIV-1/2 and OraQuick Advance Rapid HIV-1/2, for detection of recent HIV seroconversion. *J Clin Microbiol* 2008, 46(10):3482-3483.
129. Granade TC, Nguyen S, Kuehl DS, Parekh BS. Development of a novel rapid HIV test for simultaneous detection of recent or long-term HIV type 1 infection using a single testing device. *AIDS Res Hum Retroviruses* 2013, 29(1):61-67.
130. World Health Organization. Meeting report: Annual meeting of the WHO technical working group on HIV incidence assays, 18-24 August 2011, Atlanta, Georgia. http://www.who.int/diagnostics_laboratory/links/who_incidence_meeting_draftreport_december2011.pdf. 2011.
131. Bärnighausen T, Wallrauch C, Welte A, McWalter TA, Mbizana N, Viljoen J, Graham N, Tanser F, Puren A, Newell ML. HIV incidence in rural South Africa: Comparison of estimates from longitudinal surveillance and cross-sectional cBED assay testing. *PLOS One* 2008, 3(11):e3640.
132. Kim AA, Hallett T, Stover J, Gouws E, Musinguzi J, Mureithi PK, Bunnell R, Hargrove J, Mermin J, Kaiser RK, Barsigo A, Ghys PD. Estimating HIV incidence among adults in Kenya and Uganda: A systematic comparison of multiple methods. *PLOS One* 2011, 6(3):e17535.
133. McNicholl JM, McDougal JS, Wasinrapee P, Branson BM, Martin M, Tappero JW, Mock PA, Green TA, Hu DJ, Parekh B. Assessment of BED HIV-1 incidence assay in seroconverter cohorts: Effect of individuals with long-term infection and importance of stable incidence. *PLOS One* 2011, 6(3):e14748.
134. Lee LM, McKenna MT. Monitoring the incidence of HIV infection in the United States. *Public Health Rep* 2007, 122 Suppl 1:72-79.

135. Hall HI, Song R, Rhodes P, Prejean J, An Q, Lee LM, Karon J, Brookmeyer R, Kaplan EH, McKenna MT, Janssen RS. Estimation of HIV incidence in the United States. *JAMA* 2008, 300(5):520-529.
136. Karon JM, Song R, Brookmeyer R, Kaplan EH, Hall HI. Estimating HIV incidence in the United States from HIV/AIDS surveillance data and biomarker HIV test results. *Stat Med* 2008, 27(23):4617-4633.
137. Le Vu S, Le Strat Y, Barin F, Pillonel J, Cazein F, Bousquet V, Brunet S, Thierry D, Semaille C, Meyer L, Desenclos JC. Population-based HIV-1 incidence in France, 2003-08: A modelling analysis. *Lancet Infect Dis* 2010, 10(10):682-687.
138. Prejean J, Song R, Hernandez A, Ziebell R, Green T, Walker F, Lin LS, An Q, Mermin J, Lansky A, Hall HI. Estimated HIV incidence in the United States, 2006-2009. *PLOS One* 2011, 6(8):e17502.
139. Garrett N, Lattimore S, Gilbart V, Aghaizu A, Mensah G, Tosswill J, Murphy G, Delpech V. The Recent Infection Testing Algorithm (RITA) in clinical practice: A survey of HIV clinicians in England and Northern Ireland. *HIV Med* 2012, 13(7):444-447.
140. European Centre for Disease Prevention and Control. Technical Report: Monitoring Recently Acquired HIV Infections in the European Context. http://www.ecdc.europa.eu/en/publications/Publications/monitoring-recently-acquired-HIV-infections-european-context.pdf. 2013.
141. Aghaizu A, Murphy G, Tosswill J, DeAngelis D, Charlett A, Gill ON, Ward H, Lattimore S, Simmons R, Delpech V. Recent infection testing algorithm (RITA) applied to new HIV diagnoses in England, Wales and Northern Ireland, 2009 to 2011. *Euro Surveill* 2014, 19(2):pii: 20673.
142. Schwarcz S, McFarland W, Delgado V, Dilley J, Rinaldi J, Adler B, Withum D. Partner notification for persons recently infected with HIV: Experience in San Francisco. *JAIDS* 2001, 28(4):403-404.
143. Welte A. Relating incidence to 'recent infection' prevalence: Application to HIV. *S Afr J Sci* 2008, 104:199-202.
144. HIV Modelling Consortium. http://www.hivmodelling.org. Accessed 1 July 2014.
145. Balasubramanian R, Lagakos SW. Estimating HIV incidence based on combined prevalence testing. *Biometrics* 2010, 66(1):1-10.
146. McWalter TA, Kassanjee R, Welte A. Incidence from cross-sectional surveys: Improved characterization of tests for recent infection. E-poster CDC0473 at the *XVIII International AIDS Conference*, 18-23 July 2010, Austria.
147. Hallett TB. Estimating the HIV incidence rate: Recent and future developments. *Curr Opin HIV AIDS* 2011, 6(2):102-107.
148. Agresti A. *Categorical data analysis*. Second Edition. New York: Wiley, 2002.
149. Claggett B, Lagakos SW, Wang R. Augmented cross-sectional studies with abbreviated follow-up for estimating HIV incidence. *Biometrics* 2012, 68(1):62-74.
150. Rice JA. *Mathematical statistics and data analysis*. Third Edition. Belmont, CA: Duxbury Press, 2007.
151. David M, Masson S, Zeh C, Mukui I, Kiubi B, Szumilin E, Welte A, Riche B, Etard JF. High incidence despite increasing ART and circumcision coverage in Western Kenya: The Ndhiwa HIV impact in population study. Abstract PB176 at

the *17th International Conference on AIDS and Sexually Transmitted Infections in Africa (ICASA*), 7- 11 December 2013, South Africa.

152. The R Project for Statistical Computing. http://www.r-project.org. Accessed 1 July 2014.
153. Keating SM, Hanson D, Lebedeva L, Laeyendecker O, Ali-Napo L, Contestable P, Edwards S, Ethridge S, Branson B, Norris PJ, Busch MP. Optimization and calibration of less sensitive and avidity modified protocols for the Vitros immunodiagnostic products anti-HIV1+2 assay for detection of early HIV infections and incidence estimation. Presentation at the *2010 HIV Diagnostics Conference*, 24-26 March 2010, Orlando, FL.
154. Avioq, Inc. Avioq HIV-1 Microelisa System: Package Insert. 2009.
155. Remis RS, Palmer RW. Testing bias in calculating HIV incidence from the Serologic Testing Algorithm for Recent HIV Seroconversion. *AIDS* 2009, 23(4):493-503.
156. Schreiber GB, Glynn SA, Satten GA, Kong F, Wright D, Busch MP, Tu Y, Kleinman SH. HIV seroconverting donors delay their return: Screening test implications. *Transfusion* 2002, 42(4):414-421.
157. Cox DR. *Principles of Statistical Inference*. New York: Cambridge University Press, 2006.
158. Hart JD. *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag, 1997.
159. Schwarcz S, Weinstock H, Louie B, Kellogg T, Douglas J, Lalota M, Dickinson G, Torian L, Wendell D, Paul S, Goza G, Ruiz J, Boyett B, McCormick L, Bennett D. Characteristics of persons with recently acquired HIV infection: Application of the serologic testing algorithm for recent HIV seroconversion in 10 US cities. *JAIDS* 2007, 44(1):112-115.
160. Busch MP, Stramer SL, Vermeulen M, Goncales T, Keating SM, Remis RS. Derivation of HIV incidence assay 'window periods' from seroconverting blood donors in countries with diverse HIV clades. Poster 995a at the *16th Conference on Retroviruses and Opportunistic Infections*, 8-11 February 2009, Canada.
161. Young CL, Hu DJ, Byers R, Vanichseni S, Young NL, Nelson R, Mock PA, Choopanya K, Janssen R, Mastro TD, Mei JV. Evaluation of a sensitive/less sensitive testing algorithm using the bioMerieux Vironostika-LS assay for detecting recent HIV-1 subtype B' or E infection in Thailand. *AIDS Res Hum Retroviruses* 2003, 19(6):481-486.
162. Fang CT, Field SP, Busch MP, Heyns Adu P. Human immunodeficiency virus-1 and hepatitis C virus RNA among South African blood donors: Estimation of residual transfusion risk and yield of nucleic acid testing. *Vox Sang* 2003, 85(1):9-19.
163. Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. *JAIDS* 2002, 29(2):184-190.
164. Brennan CA, Stramer SL, Holzmayer V, Yamaguchi J, Foster GA, Notari Iv EP, Schochetman G, Devare SG. Identification of human immunodeficiency virus type 1 non-B subtypes and antiretroviral drug-resistant strains in United States blood donors. *Transfusion* 2009, 49(1):125-133.

165. Delwart EL, Orton S, Parekh B, Dobbs T, Clark K, Busch MP. Two percent of HIV-positive U.S. blood donors are infected with non-subtype B strains. *AIDS Res Hum Retroviruses* 2003, 19(12):1065-1070.
166. Calypte Biomedical Corporation. Aware™ BED™ EIA HIV-1 Incidence Test (IgG-Capture HIV EIA): Enzyme Immunoassay for Population Estimates of HIV-1 Incidence, Cat. No. 98003. 2008.
167. The South African Department of Health. 2008 National Antenatal Sentinel HIV & Syphilis Prevalence Survey. www.gov.za/documents/download.php?f=109007. 2009.
168. Statistics South Africa. Mid-year Population Estimates 2009, Statistical Release P0302. http://www.statssa.gov.za/publications/P0302/P03022009.pdf. 2009.
169. Smart Biotech Ltd. http://www.smartube-bio.com. Accessed 1 July 2014.
170. Novikov I, Jehuda-Cohen T. HIV type 1 infection among Ethiopian immigrants to Israel: Enhanced in vitro antibody stimulation for estimating the length of the window period. *AIDS Res Hum Retroviruses* 2009, 25(2):165-174.
171. Mumo J, Vansover A, Jehuda-Cohen T. Detecting seronegative-early HIV infections among adult versus student Kenyan blood donors, by using Stimmunology. *Exp Biol Med* 2009, 234(8):931-939.
172. Africa Centre for Health and Population Studies. http://www.africacentre.ac.za. Accessed 1 July 2014.
173. Bärnighausen T, Tanser F, Newell ML. Lack of a decline in HIV incidence in a rural community with high HIV prevalence in South Africa, 2003-2007. *AIDS Res Hum Retroviruses* 2009, 25(4):405-409.
174. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap* (*Monographs on Statistics and Applied Probability 57)*. New York: Chapman & Hall / CRC, 1993.
175. Burchell AN, Calzavara L, Ramuscak N, Myers T, Major C, Rachlis A, Gough K, Raboud J, Remis RS. Symptomatic primary HIV infection or risk experiences? Circumstances surrounding HIV testing and diagnosis among recent seroconverters. *Int J STD AIDS* 2003, 14(9):601-608.
176. Centers for Disease Control and Prevention. HIV Testing Survey, 2002. http://www.cdc.gov/hiv/pdf/statistics_2004_HIV_Surveillance_special_no5.pdf. 2004.
177. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958, 53(282):457-481.
178. Klein J, Moeschnerger M. *Survival Analysis Techniques for Censored and Truncated Data*. Second Edition. New York: Springer-Verlag, 2003.
179. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc Ser B Method* 1976, 38(3):290-295.
180. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, Heldebrant C, Smith R, Conrad A, Kleinman SH, Busch MP. Dynamics of HIV viremia and antibody seroconversion in plasma donors: Implications for diagnosis and staging of primary HIV infection. *AIDS* 2003, 17(13):1871-1879.
181. Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, Salazar-Gonzalez JF, Pham KT, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Hahn BH, Shaw GM, Korber BT, Bhattacharya T, Perelson AS. Modeling sequence evolution in acute HIV-1 infection. *J Theor Biol* 2009, 261(2):341-360.
182. Iweala OI. HIV diagnostic tests: an overview. *Contraception* 2004, 70(2):141-147.

183. Ly TD, Ebel A, Faucher V, Fihman V, Laperche S. Could the new HIV combined p24 antigen and antibody assays replace p24 antigen specific assays? *J Virol Methods* 2007, 143(1):86-94.
184. Owen SM, Yang C, Spira T, Ou CY, Pau CP, Parekh BS, Candal D, Kuehl D, Kennedy MS, Rudolph D, Luo W, Delatorre N, Masciotra S, Kalish ML, Cowart F, Barnett T, Lal R, McDougal JS. Alternative algorithms for human immunodeficiency virus infection diagnosis using tests that are licensed in the United States. *J Clin Microbiol* 2008, 46(5):1588-1595.
185. Masciotra S, McDougal JS, Feldman J, Sprinkle P, Wesolowski L, Owen SM. Evaluation of an alternative HIV diagnostic algorithm using specimens from seroconversion panels and persons with established HIV infections. *J Clin Virol* 2011, 52 Suppl 1:S17-22.
186. Hargrove J, van Schalkwyk C, Eastwood H. BED estimates of HIV incidence: resolving the differences, making things simpler. *PLOS One* 2012, 7(1):e29736.
187. Masciotra S, Dobbs T, Candal D, Hanson D, Delaney K, Rudolph D, Charurat M, Harrigan R, McDougal S, Owen M. Antibody avidity-based assay for identifying recent HIV-1 infections based on Genetic Systems™ 1/2 Plus O EIA. Abstract 937 at the *17th Conference on Retroviruses and Opportunistic Infections*, 16-19 February 2010, San Francisco, CA.
188. Bill & Melinda Gates Foundation Letters of Inquiry (LOI): New Biomarkers for HIV Incidence Measurement. https://docs.gatesfoundation.org/Documents/hiv-incidence-rules-and-guidelines.pdf. Accessed 1 July 2014.
189. Hallett TB, Ghys P, Bärnighausen T, Yan P, Garnett GP. Errors in 'BED'-derived estimates of HIV incidence will vary by place, time and age. *PLOS One* 2009, 4(5):e5720.
190. American Red Cross. http://www.redcrossblood.org. Accessed 1 July 2014.
191. Blood Centers of the Pacific. http://www.bloodcenters.org. Accessed 1 July 2014.
192. South African National Blood Service. http://www.sanbs.org.za. Accessed 1 July 2014.
193. Fundacao Pro-Sangue Hemocentro de Sao Paulo http://www.prosangue.sp.gov.br/hemocentros/Default.aspx. Accessed 1 July 2014.
194. Jain V, Liegler T, Vittinghoff E, Hartogensis W, Bacchetti P, Poole L, Loeb L, Pilcher CD, Grant RM, Deeks SG, Hecht FM. Transmitted drug resistance in persons with acute/early HIV-1 in San Francisco, 2002-2009. *PLOS One* 2010, 5(12):e15510.
195. Winkelstein Jr. W, Wiley JA, Padian NS, Samuel M, Shiboski S, Ascher MS, Levy JA. The San Francisco Men's Health Study: Continued decline in HIV seroconversion rates among homosexual/bisexual men. *Am J Public Health* 1988, 78(11):1472-1474.
196. Morris SR, Little SJ, Cunningham T, Garfein RS, Richman DD, Smith DM. Evaluation of an HIV nucleic acid testing program with automated internet and voicemail systems to deliver results. *Ann Intern Med* 2010, 152(12):778-785.
197. Costa PR, Hartogensis W, Diaz RS, Bacchetti P, Hecht R, Santos BR, AMPLIAR study team, Kallas EG, Pilcher C. Subjects infected with Clade C-Containing HIV exhibit higher CD4+ T cell activation. Poster 289 at the *2014 Conference on Retroviruses and Opportunistic Infections*, March 3-6 2014, Boston, MA.

198. Price MA, Wallis CL, Lakhi S, Karita E, Kamali A, Anzala O, Sanders EJ, Bekker LG, Twesigye R, Hunter E, Kaleebu P, Kayitenkore K, Allen S, Ruzagira E, Mwangome M, Mutua G, Amornkul PN, Stevens G, Pond SL, Schaefer M, Papathanasopoulos MA, Stevens W, Gilmour J. Transmitted HIV type 1 drug resistance among individuals with recent HIV infection in East and Southern Africa. *AIDS Res Hum Retroviruses* 2011, 27(1):5-12.
199. Hunt PW, Brenchley J, Sinclair E, McCune JM, Roland M, Page-Shafer K, Hsue P, Emu B, Krone M, Lampiris H, Douek D, Martin JN, Deeks SG. Relationship between T cell activation and CD4+ T cell count in HIV-seropositive individuals with undetectable plasma HIV RNA levels in the absence of therapy. *J Infect Dis* 2008, 197(1):126-133.
200. Sedia Biosciences Corporation. Sedia™ HIV-1 LAg-Avidity EIA: Single Well Avidity Enzyme Immunoassay for Detection of Recent HIV-1 Infection Using Liquid Serum or Plasma, Cat. No. 1002. 2013.
201. Sedia Biosciences Corporation. Sedia™ HIV-1 LAg-Avidity EIA: Single Well Avidity Enzyme Immunoassay for Detection of Recent Infection, Cat. No. 1002. 2012.
202. Bollinger RC, Brookmeyer RS, Mehendale SM, Paranjape RS, Shepherd ME, Gadkari DA, Quinn TC. Risk factors and clinical presentation of acute primary HIV infection in India. *JAMA* 1997, 278(23):2085-2089.
203. Daar ES, Little S, Pitt J, Santangelo J, Ho P, Harawa N, Kerndt P, Glorgi JV, Bai J, Gaut P, Richman DD, Mandel S, Nichols S. Diagnosis of primary HIV-1 infection – Los Angeles County Primary HIV Infection Recruitment Network. *Ann Intern Med* 2001, 134(1):25-29.
204. Hecht FM, Busch MP, Rawal B, Webb M, Rosenberg E, Swanson M, Chesney M, Anderson J, Levy J, Kahn JO. Use of laboratory tests and clinical symptoms for identification of primary HIV infection. *AIDS* 2002, 16(8): 1119-1129.
205. Powers KA, Miller WC, Pilcher CD, Mapanje C, Martinson FE, Fiscus SA, Chilongozi DA, Namakhwa D, Price MA, Galvin SR, Hoffman IF, Cohen MS. Improved detection of acute HIV-1 infection in sub-Saharan Africa: Development of a risk score algorithm. *AIDS* 2007, 21(16):2237-2242.
206. Schacker T, Collier AC, Hughes J, Shea T, Corey L. Clinical and epidemiologic features of primary HIV infection. *Ann Intern Med* 1996, 125(4):257-264.
207. Borrow P, Lewicki H, Wei X, Horwitz MS, Peffer N, Meyers H, Nelson JA, Gairin JE, Hahn BH, Oldstone MB, Shaw GM. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat Med* 1997, 3(2): 205-211.
208. Lindback S, Thorstensson R, Karlsson AC, von Sydow M, Flamholc L, Blaxhult A, Sonnerborg A, Biberfeld G, Gaines H. Diagnosis of primary HIV-1 infection and duration of follow-up after HIV exposure – Karolinska Institute Primary HIV Infection Study Group. *AIDS* 2000, 14(15):2333-2339.
209. Pilcher CD, Eron JJ, Jr., Vemazza PL, Battegay M, Harr T, Yerly S, Vom S, Perrin L. Sexual transmission during the incubation period of primary HIV infection. *JAMA* 2001, 286(14):1713-1714.
210. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *Am Stat* 1998, 52(2):119-126.

211. Mehrotra DV, Chan IS, Berger RL. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* 2003, 59(2):441-450.
212. Lydersen S, Fagerland MW, Laake P. Recommended tests for association in 2 × 2 tables. *Stat Med* 2009, 28(7):1159-1175.
213. Chaillon A, Le Vu S, Brunet S, Gras G, Bastides F, Bernard L, Meyer L, Barin F. Decreased specificity of an assay for recent infection in HIV-1-infected patients on highly active antiretroviral treatment: Implications for incidence estimates. *Clin Vaccine Immunol* 2012, 19(8):1248-1253.
214. Department of Health and Human Services. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. http://aidsinfo.nih.gov/contentfiles/lvguidelines/AdultandAdolescentGL.pdf. Accessed 5 March 2014.
215. South African Department of Health. The South African Antiretroviral Treatment Guidelines,Version 14, March 2013. http://www.sahivsoc.org/upload/documents/2013%20ART%20Guidelines-Short%20Combined%20FINAL%20draft%20guidelines%2014%20March%202013.pdf. Accessed 5 March 2014.
216. Le Vu S, Velter A, Meyer L, Peytavin G, Guinard J, Pillonel J, Barin F, Semaille C. Biomarker-based HIV incidence in a community sample of men who have sex with men in Paris, France. *PLOS One* 2012, 7(6):e39872.
217. Sharma UK, Schito M, Welte A, Rousseau C, Fitzgibbon J, Keele B, Shapiro S, McMichael A, Burns DN. Workshop summary: Novel biomarkers for HIV incidence assay development. *AIDS Res Hum Retroviruses* 2012, 28(6):532-539.
218. National Institutes of Health: Funding opportunity announcement for 'HIV Incidence Assays with Improved Specificity'. http://grants.nih.gov/grants/guide/pa-files/PA-12-012.html. Accessed 1 July 2014.
219. Braunstein SL, Nash D, Ingabire C, Mwamarangwe L, van de Wijgert JH. Performance of BED-CEIA and Avidity Index assays in a sample of ART-naïve, female sex workers in Kigali, Rwanda. Abstract 939 at the *17th Conference on Retroviruses and Opportunistic Infection*, 16-19 February 2010, San Francisco, CA.
220. Cousins MM, Donnell D, Eshleman SH. Impact of mutation type and amplicon characteristics on genetic diversity measures generated using a high-resolution melting diversity assay. *J Mol Diagn* 2013, 15(1):130-137.
221. Grant opportunity: Population-based HIV Impact Assessments in Resource-Constrained Settings under the President's Emergency Plan for AIDS Relief (PEPFAR). http://www.grants.gov/web/grants/view-opportunity.html?oppId=252788. Accessed 1 July 2014.
222. Greene WH. *Econometric Analysis*. Seventh Edition. Upper Saddle River, New Jersey: Prentice Hall, 2011.