OPTIMISATION AND BENCHMARKING OF ANALYTICAL APPROACHES TO ESTIMATION OF POPULATION LEVEL HIV INCIDENCE FROM SURVEY DATA

by

Laurette Mhlanga

Thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosohpy in Epidemiology in the Faculty of Medicine and Health Sciences at Stellenbosch University

Department of Global Health,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Prof. Alex Welte & Co - Supervisor: Dr. Eduard Grebe

April 2022

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: .... April 2022 ....

Abstract

OPTIMISATION AND BENCHMARKING OF ANALYTICAL APPROACHES TO ESTIMATION OF POPULATION LEVEL HIV INCIDENCE FROM SURVEY DATA

L. Mhlanga

Department of Global Health,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: PhD. (Epidemiology)

April 2022

Disease prevalence (the proportion of a population with a condition of interest) is conceptually and procedurally much more straightforward to estimate than disease incidence (the rate of occurrence of new cases - for example, infections). For long-lasting conditions, incidence is fundamentally more difficult to estimate than prevalence, but also more interesting, as it sheds light on current epidemiological trends such as the emerging burden on health systems and the impact of recent policy interventions. Progress towards reducing reliance on questionable assumptions in the analysis of large population based surveys (for the estimation of HIV incidence) has been slow. The work of Kassanjee et al and the work of Mahiane et al, in particular, provide rigorous ways of estimating incidence by using 1) markers of 'recent infection', 2) the 'gradient' of prevalence, and 3) 'excess mortality' associated with HIV infection, without the need for simplifying assumptions to the effect that any particular parameters are constant over ranges of time and/or age. To date, the use of these methods has largely ignored 1) the rich details of the age and time structure of survey data, and 2) the opportunities for combining the two methods.

The primary objective of this work was to find stable approaches to applying the Mahiane and Kassanjee methods to large age/time structured population survey data sets which include HIV status, and optionally, 'recent infection' status. In order to evaluate proposed methods, a sophisticated simulation platform was created to simulate HIV epidemics and generate survey data sets that are structured like real population survey data, with the underlying incidence, prevalence, and mortality explicitly known.

The first non-trivial step in the analysis of survey data amounts essentially to performing a smoothing procedure from which the (age/time specific) prevalence of HIV infection, the prevalence of 'recent infection', and the gradient of prevalence of infection can be inferred without recourse to 'epidemiological' assumptions. The second step involves the correct accounting for uncertainty in a context-specific weighted mean of the Mahiane and Kassanjee estimators. These two steps are approached incrementally, as there are numerous details which have not previously been systematically elucidated.

The investigation culminates in a proposed generic 'once size fits most' algorithm based on: 1) fitting survey data to generalised linear models defined by simple link functions and high order polynomials in age and time; 2) the use of a 'moving window' rule for data inclusion into a separate analysis for each age/time point for which incidence is to be estimated; 3) a 'variance optimal' weighting scheme for the combination of the Mahiane and Kassanjee estimators (when both are applicable); 4) flexible use of a delta method expansion or bootstrapping to estimate confidence intervals and p values. We find it is relatively easy to obtain estimates with practically negligible bias, but sample-sizes/sampling-density requirements are always considerable. We also make numerous observations on survey design and the inherent challenges faced by all attempts to estimate HIV incidence using surveys of reasonable size.

Keywords: incidence, prevalence, cross-sectional surveys, population-level surveys, HIV incidence estimation.

Opsomming

OPTIMERING EN VERGELYKING VAN ANALITIESE BENADERINGS TOT DIE BERAMING VAN BEVOLKINGSVLAK MIV-INSIDENSIE UIT OPNAME DATA

L. Mhlanga

Departement Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: PhD. (Epidemiologie)

April 2022

Die prevalensie van siektes (die proporsie van ′n bevolking met ′n sekere siekte) is konseptueel en prosedureel baie eenvoudiger om te beraam as die insidensie van siektes (die voorkoms van nuwe gevalle - byvoorbeeld infeksies). Vir langdurige toestande is die insidensie fundamenteel moeiliker om te beraam as die prevalensie, maar ook interessanter, aangesien dit lig werp op die huidige epidemiologiese tendense, soos die opkomende las op gesondheidstelsels en die impak van onlangse beleidsintervensies. Twyfelagtige aannames word gemaak gedurende die ontleding van groot bevolkingsopnames om die insidensie van MIV te beraam, en tog word daar gesteun op hierdie studies. Die werk van Kassanjee et al, en veral die werk van Mahiane et al, bied deeglike metodes om insidensie te beraam deur 1) merkers van ′onlangse infeksie′, 2) die ′gradiënt′ van prevalensie en 3) ′oortollige sterftes′ wat verband hou met MIV -infeksie te gebruik. Hierdie metodes maak nie die aannames dat sekere parameters konstant is oor tydsperiodes en/of ouderdomme nie. Tot op datum het die gebruik van hierdie metodes grootliks 1) die ryk besonderhede van die ouderdom en tydstruktuur van opname-data, en 2) die geleenthede om die twee metodes te kombineer, geïgnoreer.

Die primêre doel van hierdie werk was om stabiele benaderings te vind vir die toepassing van die Mahiane- en Kassanjee-metodes op groot ouderdom-/tyd-gestruktureerde opname datastelle, wat MIV-status, en soms die status van 'onlangse infeksie' insluit. Om voorgestelde metodes te evalueer, is 'n gesofistikeerde simulasieplatform geskep om MIV-epidemies te simuleer en opname datastelle te genereer wat soos werklike bevolkingsopname data is, met die onderliggende insidensie, prevalensie en sterftes uitdruklik bekend.

Die eerste nie-triviale stap in die analise van opname-data kom in wese neer op die uitvoering van 'n afstrykingsprosedure waaruit die (ouderdom/tydspesifieke) prevalensie van MIV-infeksie, die prevalensie van 'onlangse infeksie' en die gradiënt van prevalensie van infeksie afgelei kan word sonder om van 'epidemiologiese' aannames gebruik te maak. Die tweede stap behels die korrekte kwantifisering van onsekerheid in 'n konteks-spesifieke geweegde gemiddelde van die Mahiane en Kassanjee beramings. Hierdie twee stappe word inkrementeel benader, aangesien daar 'n groot aantal besonderhede is wat nie voorheen stelselmatig ondersoek is nie.

Die ondersoek loop uit op 'n voorgestelde generiese 'once size fits most' algoritme gebaseer op: 1) die pas van opname data tot veralgemeende lineêre modelle gedefinieer deur eenvoudige skakelfunksies en hoë orde polinome in ouderdom en tyd; 2) die gebruik van 'n 'bewegende venster' -reël vir die insluiting van data in 'n aparte analise vir elke ouderdom/tydspunt waarvoor die insidensie beraam moet word; 3) 'n 'variansie-optimale' wegings-skema vir die kombinasie van die Mahiane- en Kassanjee -beramers (wanneer beide van toepassing is); 4) buigsame gebruik van 'n delta-metode uitbreiding of bootstrapping om vertrouensintervalle en p-waardes te skat. Ons vind dit relatief maklik om beramings te verkry met onbeduidende sydigheid, maar die vereistes vir steekproefgroottes/steekproefdigtheid is altyd aansienlik. Ons maak ook talle opmerkings oor die ontwerp van opnames en die inherente uitdagings waarmee alle pogings om die insidensie van MIV uit opname data te beraam, gekonfronteer word.

Sleutelwoorde: prevalensie, insidensie, deursnee-opnames, bevolkingsopname, beraming van insidensie van MIV.

Acknowledgements

First, I would like to express my heartfelt gratitude to Professor Alex Welte, my supervisor (mentor). The journey was not easy, but he was present at every step of the way. I am humbled by his patience with me as he gently and diligently guided me through the discourse of HIV surveillance. I valued most his erudition, thoroughgoing criticism, scholarly counsel, and suggestions to this work. I am and will be eternally grateful for having worked with Professor Alex Welte words will never be enough to express the depth of my gratitude. I will also like to thank my co-supervisor Eduard Grebe, for his input, this body of work benefited immensely from his suggestions and keen eye.

I would like to express my utmost gratitude to SACEMA for funding my studies, being my support system, and bestowing me with the honour to become part of the family. When I arrived in South Africa, the SACEMA family took me under their wing and made me feel at home. I have had a phenomenal experience and I will forever cherish the great moments shared with each of the SACEMA family members. The smiles, motivation and conversations exchanged may seem minor but had a huge impact on my well being and ability to continue with my dissertation. I take this moment to especially thank Prof. Juliet Pulliam (Director of SACEMA) and Prof. Gavin Hitchcock for constantly reaching out, checking on me and motivating me.

I would like to thank and appreciate my collaborators from SANBS and WCBS, for affording me the chance to be part of the team and work with them on the COVID 19 seroprevalence study. I enjoyed the work, and it was a pleasure to be part of the meetings and brainstorming sessions.

I would like to thank and attribute my sound sanity to my friends (anchor points) Gamuchirai Mamhende, Zinhle Mthombothi, Admire Phiri, Keboneilwe Toolo, and Kuda. They each provided me with the unique human contact I needed through this phase of my life. I appreciate the time they spent with me and their willingness to listen to me ramble on about my life and PhD journey.

Last but not least, I owe a debt of gratitude, appreciation, and love to my family (parents, siblings and partners, nieces and nephews) for having my best interests at heart. I would never have come this far without their gentle push, support and love. I have never regretted the decisions I have made based on your guidance. A special thank you to Ngoni (aka Dr Mhlanga) for pointing me to SACEMA.

Dedications

This body of work is dedicated to my parents (Alice and John Mapati Nyambe Mhlanga), siblings and their partners, lovely nieces and nephews, for their love and support throughout the years.
To the Lord all mighty for in the fullness of time he makes it all perfect.

Publications

1. The following preprints are presented verbatim as Chapter 5, 6, and 7 in this body of work and are currently being prepared for submission to peer reviewed journals;

   - Mhlanga L, Eduard G, Welte A. Optimal accounting for age and time structure of HIV incidence estimates based on cross-sectional survey data with ascertainment of 'recent infection'. **DOI:** 10.21203/rs.3.rs-871044/v2
   - Mhlanga L, Eduard G, Welte A. Smoothing age/time structure of HIV prevalence, for optimal use in synthetic cohort based incidence estimation.**DOI:** 10.21203/rs.3.rs-959136/v2
   - Mhlanga L, Grebe E, Welte A. The added value of recent-infection testing in population-based HIV surveys. **DOI:** 10.21203/rs.3.rs-996585/v2

2. Chapter 8 is essentially comprised of two preprints, based on research conducted for the most part by the South African National and Western Cape Blood Services (SANBS / WCBS) for which I contributed the core primary analysis, and which are being combined into a unified submission to a journal. These preprints are

   - Vermeulen M, Mhlanga L, Sykes W, Coleman C, Pietersen N, Cable R, Swanevelder R, Glatt TN, Grebe E, Welte A, van den Berg K. Prevalence of anti-SARS-CoV-2 antibodies among blood donors in South Africa during the period January-May 2021. **DOI:** 10.21203/rs.3.rs-690372/v2
   - Mhlanga L, Vermeulen M, Grebe E, Welte A. SARS CoV 2 Infection Fatality Rate Estimates for South Africa. **DOI:** 10.21203/rs.3.rs-707813/v2

   The contributions of the authors are summarised below. The 'extent of contribution' estimates, which add up to 100%, are intended to cover the totality of the two

Name/ Signature	**Email**	**Nature of Contribution**	**Extent of Contribution**
Laurette Mhlanga	Laurette@sun.ac.za	Core primary analysis	7
Marion Vermeulen	Marion.vermeulen@sanbs.org.za	Principal Investigator All aspects of study	20
Alex Welte	Alexwelte@sun.ac.za	Study/analysis design	5
Karin van den Berg	Karin.vandenberg@sanbs.org.za	All aspects of study	15
Ronel Swanevelder	Ronel.swanevelder@sanbs.org.za	Data Manager	10
Charl Coleman	Charl.coleman@sanbs.org.za	Head of testing for 'central' provinces	10
Wendy Sykes	Wendy.Sykes@sanbs.org.za	Head of testing for 'coastal' provinces	10
Tanya Glatt	Tanya.Glatt@sanbs.org.za	Field work at SANBS	10
Nadia Petersen	nadiap@wcbs.org.za	Testing at WCBS	5
Russel Cable	russell@wcbs.org.za	Study Coordinator - WCBS	5
Eduard Grebe	egrebe@vitalant.org	Analytical support	3
ALL		Collective management Interpretation of results Writing of manuscript	

preprints, though not all authors contributed to both. This is naturally not objectively ascertainable, but expresses a heuristic estimate to which the authors have agreed by signature.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background to Incidence Estimation

Chronic diseases are highly prevalent in the world, for example, in 2020 approximately 39 million people were living with HIV [1] with an estimated 1.5 million being new infections. It is important to track the epidemiological trajectory of diseases i.e., the main purpose of disease surveillance, in particular HIV surveillance, is to provide measures of trends and absolute state of the HIV epidemic [2, 3, 4, 5], which encompass providing an understanding of the epidemic, identifying the sources and drivers of new HIV infections, designing and evaluating the intervention programs, and informing policy.

Of the two key epidemiological measures in HIV surveillance i.e., prevalence (the proportion of the population with the condition of interest) and incidence (the rate of occurrence of new infections in a population), incidence is the most informative. Incidence sheds light to the questions at hand, for example, a constant/drop in the prevalence (in a region or country) does not necessarily mean that there have been no/few infections but showcases the complex interaction of incidence, migration, and mortality in the population under consideration [4]. Equally, an increase in (or greater) prevalence does not imply an increase in the (or greater) force of infection (rate at which susceptible individuals in a population acquire an infectious disease in that population, per unit time). Most importantly the survival times of individuals where Anti-Retro-viral Therapy (ART) is readily available, is likely to be longer than in countries where ART is (only) available to the minority. Consequently, countries with better ART programmes, or which accessed ART earlier may have high HIV prevalence.

Comparing the incidence estimation of non-transient (for example, HIV) to transient conditions (for example, influenza), estimating the incidence for transient conditions, less complex. Because there is a straightforward relationship between incidence and prevalence, such that one is utilised to infer the other, assuming prevalence is at an endemic equilibrium and the duration of the infection is fixed. More precisely:

$$\lambda = \frac{N_I}{N_S \cdot \mu_D} = \frac{P}{1-P} \cdot \frac{1}{\mu_D} \tag{1.1.1}$$

Where λ is the incidence normalised to the susceptible population (and $\lambda' = \frac{P}{\mu_D}$ is the incidence normalised to the total population), N_S is the number of susceptible people in the population, N_I is the number of infected people in the population, μ_D is the mean duration of the infection, and P is the prevalence.

The Joint United Nations Programme on HIV/AIDS (UNAIDS Reference Group on Estimates, Modelling, and Projections. UNAIDS Reference Group meetings) working group recommends tracking the temporal trajectory of the HIV epidemic through Spectrum/Estimation and Projection Package (Spectrum/EPP) and the 'recency' framework. Over 160, countries in the world use EPP/Spectrum, to create the HIV epidemic estimates [6]. Spectrum/EPP provides comparable epidemiological indicators and proxies to age-specific incidence, but there is a need to improve the age-structured HIV incidence estimates. This is a recurring theme in the bi-annual UNAIDS meetings i.e., the specific recommendation being to investigate the importance and feasibility of accommodating the age structure in the EPP/Spectrum model [6].

Typical age-structured incidence estimates are in 5 year age bins [7, 8], which implies a complex incidence average over the ages within the age bin. But incidence and prevalence are age-specific in important ways, and may differ considerably even within a given age bin, for example, the 15 - 19 year age bin. Eaton et al. [7] also explored the age structure and introduced the Estimation and Projection Package Age-Sex Model and the r-hybrid model, but still the HIV incidence estimates are not age specific. Evidently, with the epidemic having matured, it is important to yield age-specific incidence estimates as there may be no substantial changes in HIV prevalence and incidence in the time direction [9, 10].

1.2 Rationale and Purpose of the Study

Unfortunately, progress towards reducing reliance on questionable assumptions in HIV incidence estimation has been slow, independent of the abundance of data (population-level HIV surveillance data [11, 12, 13, 14]) which is key in improving the existing methods and relaxing some assumptions.

This body of work explores two approaches of incidence estimation that have been shown to work reasonable well, namely the Mahiane et al. [3] and Kassanjee et al. [4] framework.

1. The Mahiane et al estimator is an incidence estimator that naturally emerges from the population renewal equations themselves and it does not make unreasonable epidemiological/demographic assumptions but allows incidence to emerge naturally from the dynamical equation. However, the uptake of this approach has been slow, as it required population-level survey data and reliable information on excess mortality. Additionally, not much guidance was given as to how one can summarise population-level survey data into the input parameters required - prevalence, and gradient of prevalence.

   Fortunately, this framework is more applicable now, as data from routine cross-sectional surveys conducted by organisations, *Demographic Health Surveys* (DHS) [11], *Population Based HIV Impact Assessment* (PHIA) [12]. Additionally, there has been great improvement in the excess mortality estimates including the reduced rate of disease-induced death (due to improved treatment coverage). This abundance of routine cross-sectional survey and excess mortality data is motivation enough to revisit and investigate approaches to optimise Mahiane method , explore its applicability and limitations in yielding precise age specific incidence, and incidence trends, estimates.

2. The Kassanjee et al framework is well embraced and widely used to estimate incidence but normally to yield population-level survey estimates as the sample size is not sufficiently statistically robust (lacks statistical power) to yield informative age-specific incidence and incidence trends estimates. The Kassanjee et al. [4] framework also falls short of giving precise guidance on how the population

level survey data may be summarised into prevalence and recency estimates.

It is essential to investigate ways to improve the statistical power of incidence estimates derived from the Kassanjee method and whether/not the continuous (regression) methods yield more precise age-specific prevalence and recency estimates required by the incidence estimator versus the prevalence and recency estimates from the naïve binning approaches.

Hence we revisit these methods with the intent of shedding light on their applicability. The proposed investigations (testing, validating, and benchmarking) cannot be applied to real data, but on simulated data where the analyst has control of the experiment and knows the real answers. The proposed approach data ensures that both the random errors and statistical errors are quantifiable. Hence there is a need for a platform that simulates population-level survey data such as those of DHS [11], and PHIA [12] (including recency ascertainment). Unfortunately, such a specific simulation platform does not exist and hence we designed a simulation platform for the proposed investigations.

In summary, there is a need to investigate how survey data may be summarised into prevalence, gradient of prevalence and recency estimates to yield optimal age-specific incidence and incidence trend estimates and before embarking on the investigation an age-structured simulator is required.

1.3 Objectives

1. To create a simulation platform that simulates population dynamics (HIV epidemic) and carries out surveys for an age/time structured population
2. Use the platform to investigate alternative approaches optimal smoothing of prevalence data for surveys aimed at HIV incidence estimation.
3. Distil the lessons learned into distributable practical tools and guidance on how to use survey data.
4. Use the lower level tools (which implement analysis of real or simulated surveys) to build higher-level tools (which explore various conditions under which all this happens) to support context-appropriate survey design.

1.4 Overview of the Thesis

The bulk of this thesis is focused on testing, validating, and benchmarking HIV surveillance methods, particularly incidence. The thesis outline is as follows:

Chapter 2 - Literature Review

We revisits the progress made by various organisations in developing HIV surveillance methods from data collection up incidence. The chapter's main emphasis is on recapping details of HIV surveillance methods, their pros and cons, including data sources, estimation methods and tools available to countries and organisations that seek to estimate population-level incidence.

Chapter 3 - Time Structure but no Explicit Age Structure

This chapter demonstrates the key limitations of incidence estimation using a simplified and less complex scenario - of an age cohort. These people have the same birth date, and have the same age at any specific time and experience the same demographic rates - incidence and mortality. The chapter is a prelude to more complex investigations presented in the subsequent chapters.

Chapter 4 - Computational Platform for Scenario Simulation and Analysis Benchmarking

Primarily, we describe the age/time structured population simulation focusing on the analytical approaches and computational decisions made. The platform simulates an age- and time-structured population based on a standard Susceptible Iinfec model. A section of this chapter gives a detailed account of designing and simulating cross-sectional surveys and further summarises the methodological approaches implemented to summarise population survey data into a prevalence, prevalence of recency among HIV positive and gradient of prevalence and tools used to determine optimal inclusion distance and polynomial order permutations.

Chapter 5 - Optimal Accounting of Age/Time Structure in Cross-sectional Surveys with 'Recency' Data

After the creation of the simulation platform we embark on the investigation of optimal ways to smooth population survey data with recency ascertainment from a single cross-sectional survey into a prevalence and prevalence of recency. The main focus is investing optimal polynomial order(s) and inclusion distance to yield an accurate and informative incidence estimate. We further compare the incidence estimates from arbitrary pooling of the data versus fitting a regression model in the case of estimating age-specific incidence.

Chapter 6 - Smoothing Survey Data for Mahiane Incidence Estimator.

We explored the missed opportunities in the seminal work of Mahiane et al. [3] highlighting how best population survey data may potentially be summarised into a prevalence and rate of change of prevalence. The key aspect is how best to account for the age/time structure in major population-level survey data. We discuss the optimal choice of polynomial order and inclusion distance permutation and focus briefly on link functions. Given the 'one size fits most' solution we revisit key attributes associated with synthetic cohort approaches including the time between surveys (inter-survey interval), sample sizes, and effects of unknown disease-associated mortality (excess mortality).

Chapter 7 - Value of Recency Data in Surveys

We proceed to use the methodological developments and recommendations in previous chapters (on Kassanjee and Mahiane frameworks) to augment the two methods, using the optimally weighted incidence estimator. The chapter explores the sensitivity of the standard error to the normalised weights, weighs in on the value of recency to HIV surveillance and considers estimating incidence trends from two or more cross-sectional surveys.

Chapter 8 - COVID 19 analysis Among Blood Donors.

An additional chapter that that describes work carried out on COVID-19. The chapter focuses on the work done in collaboration with South African National Blood Services (SANBS) and Western Cape Blood Services (WCBS) to infer the seroprevalence among the South African blood donors and the associated infection fatality rate based on the readily available data from South African Medical Research Council (SAMRC).

Chapter 9 Concluding Remarks

We discuss the lessons learnt from our investigations focusing on methodological improvements, the benefits and giving advice/recommendations on deriving age-specific HIV incidence estimates based on Mahiane [3] and Kassanjee [4] approaches from population level surveys. We highlight the advantages and disadvantages of augmenting these two methods, and the shortcomings of each of the methods.

1.5 Contribution and Originality of the Study

The study explores several nuanced aspects of HIV incidence estimation and is focused on developing algorithms to facilitate method improvement and ultimately provide guidance and recommendation. The key features that add to the existing body of knowledge and offer uniqueness to the study are as follows:

1. New methods to acquire better performance from existing data formats.
2. Simulation/benchmarking techniques are not systematically used to evaluate/-compare HIV surveillance methods, especially for methods with different frameworks. The thesis presents a generic simulation platform that enables method comparison i.e, 'synthetic cohort' vs 'recency'.
3. A non mechanistic modelling framework which makes it possible to simulate an age/time structured population, and complex surveys without excessive complexity beyond the structured sampling design.
4. An explicit performance benchmarking of a wide range of methods and production of tools that can eventually be slotted into routine use/provide an alternative to existing methods.

Chapter 2

Review of HIV Incidence estimation methods and tools

2.1 Introduction

We provide an unstructured review of population level HIV surveillance, concentrating on;

1. Population level survey data sources
2. Broad categories of the HIV incidence estimation methods
3. Current tools meant for HIV surveillance and specifically HIV incidence estimation

2.2 Data sources for HIV Surveillance

2.2.1 Evolution of Data Collection

The methods of collecting HIV surveillance data have evolved, since the first HIV cases in the early 1990s. Initial surveillance was mainly in the form of AIDS case reporting (still applicable in some contexts), which was seen to be very limited in generalised epidemics. Subsequently, the anonymous HIV testing of pregnant women, STI patients in clinics, and sex workers at sentinel sites was introduced [15].

2.2.2 Sentinel Sites

This approach gives the number of people infected among those who frequent the antenatal clinics, but it cannot provide prevalence estimates of the general population [2, 15, 16, 17] though they may provide a meaningful indicator. Additionally, the data may be subject to site selection bias, most of the sites being government clinics, frequented mainly by a particular sub-population defined by social and economic status. Pisani et al. [15], Magnani et al. [16], Zaba and Gregson [18] argued for better sampling strategies to reduce the bias from antenatal data. Specifically, it is worth noting that the improved representativeness of ANC SS sites advocated by [15, 16, 18] were implemented across most countries in the early 2000s, following revised guidelines in 2001 by CDC and WHO for improved ANC SS.

2.2.3 The rise of Population Level Surveys

Sentinel data is usually augmented, in the generation of population level estimates, by data from other sources such as Demographic Health Surveys (DHS) and Population-Based HIV Impact Assessments (PHIA). Representativeness in HIV surveillance is key, and various survey methods/designs exist which are meant to estimate the population-level HIV epidemiological metrics [11, 12, 13, 14]. Cross-sectional surveys are snapshots of the population at a given time. These surveys entail visiting (possibly at intervals of 3 - 5 years) a subset/sample of the population (usually independently selected for each 'round') during a specified period of a given time (usually 5 - 8 months, though often conceived of as happening instantaneously) [11, 12, 13, 14].

Information on age, geographical area, serostatus, and (sometimes) 'recency status' (defined as testing negative/positive to newly acquired infection as classified by biological assay) of infected individuals, is collected from all consenting individuals. Some of the programmes that conduct/implement these national household-based surveys are, DHS, PHIA, (Kenyan, and other) AIDS Indicators Survey (KAIS), and various AIDS Impact Surveys (AIS). South Africa collects its own population-level survey data independently under the South African National HIV Prevalence HIV Incidence, Behaviour and Communication Survey (SABSSM), primarily via a division of the Human Sciences Research Council (HSRC) (HIV/AIDS, STIs, TB department (HAST)) [13]. To ensure representativeness, most of these use complex (multi-staged) surveys, which is an umbrella term used for surveys that aggregate two or more common sampling strategies (simple random (SRS), systematic, stratified, and cluster sampling) and are each implemented at different stages of sampling the population.

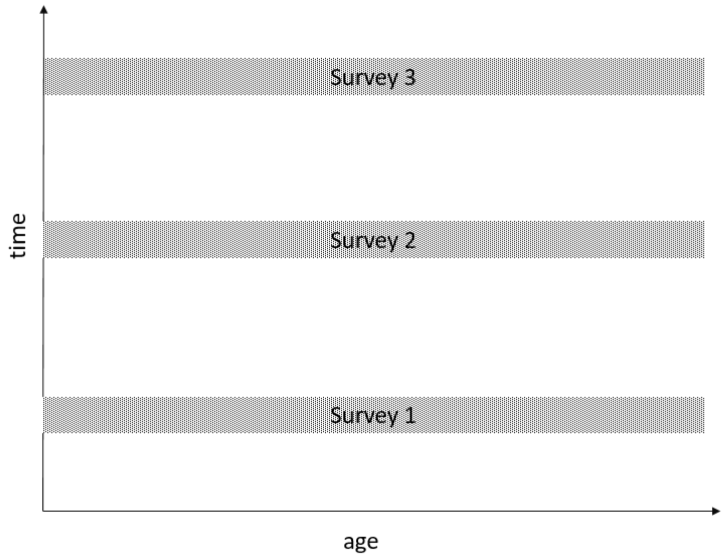

Figure 2.1: Depiction of 3 cross-sectional surveys equally spaced. The picture depicts how cross sectional surveys meant for HIV surveillance are distributed in time. A single speckled band represents the status of the population at that given time and a speckle represents an individual who was part of the survey.

South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey (SABSSM)

The SA HSRC has conducted five surveys since 2002. Approximately within the last decades HSRC introduced laboratory methods (HIV diagnostic testing) which facilitated direct HIV incidence estimation and exposure to ART [13]. The HSRC seeks to disseminate information on HIV prevalence, incidence, and crucially, HIV relevant behavioural factors.

All individuals residing in South Africa, excluding those staying in designated institutions, are part of the target population. The master sample (probability sample of the 2001 census enumerator areas (EAs) representing the provincial, settlement, and racial diversity of South Africa) is used to decide EAs that are considered as the Primary Sampling Units (PSU) [19]. The secondary sampling units are the households, and ultimately the sampling units are individuals eligible to participate in the survey (spent the night in the dwelling unit) [13, 19]. Each sampled unit is assigned a weight to support population representativeness. Households were randomly selected from all communities in South Africa [13].

In the most recent survey, 11000 households agreed to participate, 38000 individuals consented to complete the individual questionnaires and 29000 agreed to provide blood samples for HIV testing [13]. The collected data give insight into the observed trends of HIV incidence [19] and the data has facilitated numerous studies [19, 20, 21].

Demographic Health Surveys - DHS

Nationally representative DHS surveys are carried out in approximately ninety countries. However, only a subset of these include HIV testing, which are presumably the only surveys relevant to this research HIV[11]. The data collected are used to calculate various population indicators, including incidence [22]. DHS conducts two kinds of surveys; the standard DHS survey (repeated every 3 - 5 years) and interim surveys (in-between standard DHS surveys and reported annually) [22]. The surveys are voluntary and anonymous. Individuals aged 15-49 are eligible to participate and standard DHS sampling frames range from 5000- 10000 households [22].

The study design is a stratified two-stage cluster sampling. The first stage of the survey involves sampling from the enumeration areas (EA) and followed by sampling households from each EA [22]. The residential levels in the survey include both the rural and urban areas [22] and on average the surveys take 18 - 20 months. This time scale includes survey design, visits, final result analysis and communication [22].

Population-Based HIV Impact Assessments - PHIA

PHIA is a project led by the Centre for Disease Control (CDC) International Center for AIDS Care and Treatment Programs (ICAP) at Columbia University [12]. Surveys are conducted in approximately 15 countries supported by the President's Emergency Plan for AIDS Relief (PEPFAR), which is a US initiative aimed at addressing the HIV/AIDS epidemic. PHIA's strength is derived from the ICAP which has experience with Swaziland HIV Incidence Measurement Surveys (SHIMS). The studies are supported by the

United States Center for Disease Control and Prevention (CDC) and the Ministry of Health in the hosting countries [12].

Their main objective is to measure the reach and impact of the HIV interventions [12]. The data collection encompasses household interviews, individual surveys and diagnostic tests. Consenting individuals from the pre-selected households are interviewed and tested, for HIV and other related diseases [12].

Swaziland HIV Incidence Measurement Surveys - SHIMS

SHIMS1 (2011) was initially aimed at evaluating the HIV epidemic in Swaziland before an intervention (Male Circumcision Accelerated Saturation Initiative project (MC ASI) and after the MC ASI [23]. SHIMS2 was conducted in August 2016 to March 2017 under the PHIA guidelines and as part of the then 13 countries that PHIA supported. Note SHIMS1 and SHIMS2 had differing sampling strategies. The SHIMS study also involved a substantial cohort component [23]. The short term cohorts served to reduce the cohort effects e.g. logistics cost and the loss to follow up. The study was a household-based and nationally representative longitudinal cohort [24] survey and had a high cooperation (participation) rate and the samples were adjusted appropriately for the non-responses.

A two-staged sampling study design was implemented, with 575 EAs. The households in each EA were numbered using the geographical position system and then 26 households were randomly selected from each EA [23]. The selection strategy involved a systematic sampling of the households [23]. The probability of selecting a household in each EA was proportional to the population size and within each household, every eligible candidate was requested to participate [23].

Cohort A1 was recruited 6 months before the MC ASI and followed for another 6 months after the MC ASI [23]. Cohort A2 was recruited after MC ASI and followed for 12 months. Cohort B was created from a subset of cohort A2 and they are followed for an additionally 6 to 12 months after the initial 12 months [23].

The study design concentrated on the MC ASI. Other intervention programs were not accommodated, limiting the possibility of investigating the most effective intervention [23, 24].

Kenyan AIDS Indicators Survey - KAIS

Kenyan AIDS Indicators Survey (KAIS) is a Kenyan based organisation in collaboration with the Kenyan Ministry of Health. At the time of inception, the main aim was to evaluate the country's response to the HIV epidemic [14] and two surveys were conducted in 2007 and 2012.

The study design implemented was a 2 staged cluster survey [14]. 372 clusters were systematically sampled from 5360 clusters (from 47 counties, stratified into rural and urban areas - minus the Northern Eastern part of Kenya) of the National Sample Survey and Evaluation Programme V Frame (NASSEP V) [25]. The second phase used equal probability systematic sampling, to sample 25 households in each cluster [25] and the collected data was on households, demographics, recent infections and behavioural aspects.

The key improvements made to the KAIS 2012 was the introduction of children in the sampling frame. The children's sample (18months - 14 years) was randomly sampled from the rest of the households [14]. HIV prevalence was estimated for the fraction of the population aged between 18 months and 64 years [26].

2.3 HIV Incidence Estimation Methods

Incidence, rather than prevalence, is the more potent epidemiological metric to assess the impact of interventions. In this section we explore the evolution of HIV incidence estimation methods and their main categories.

Seminal papers with differing schools of thoughts have been published and all are aimed at one goal - to provide accurate and precise estimates of incidence. These incidence estimation methodologies can be classified into three different types approach; 1. The cohorts approach, 2. The dynamic approach and 3. The test for "recency" of HIV infection.

1. Repeated study-subject interactions (cohorts), involving multiple HIV status ascertainment, to accumulate paired status conversion event counts and exposure time.

2. Estimates derived from largely unrepeated study-subject combinations, but spread over some sufficient range of ages and/or times to allow estimation of gradients of prevalence. These measures may need to be augmented by particular contextually valid mortality estimates (cross - sectional survey data), and

3. Narrowly time-distributed ascertainment of both a) HIV infection and b) categorical recent/ non-recent infection according to a well-characterised case definition.

Cohorts

The gold standard for estimating HIV incidence is traditionally to follow up a cohort that is initially uninfected [27].

The cohorts are useful especially when addressing a concentrated epidemic [2] or at the beginning of an epidemic when we seek to; track temporal priority [28], understand the specific characteristics and progression of the epidemic [24]. A good example of a cohort survey that was robust in HIV incidence estimation was the SHIMS study, that highlighted the impact of male circumcision in reducing HIV incidence [24]. Examples of the cohort studies that are not generalisable are; Tanser et al. [29], Feldblum et al. [30], Xu et al. [31], and Feldblum et al. [32]

Unfortunately, the incidence estimate is limited to the cohort in question and is seldom representative of large population. This arises from selection criteria, sample sizes, repeated testing, and counselling which lead to behaviour change (Hawthorne effect) Grebe et al. [9], Brookmeyer et al. [17], Hall et al. [33]. Additionally, they are expensive, time-consuming, tedious, and are exposed to non-random loss to follow up. Large sample sizes are required for the results to be statistically robust and to achieve a useful measure representative of the total population [2, 34, 35].

It is worth pointing out that cohorts have been a success in surveillance of some chronic conditions, for example, heart diseases [34, 36]. In the case of HIV, cohorts offered a unique opportunity to study the interplay of infection/opportunistic infections, dynamics of the disease and provided a direct measure of the differential mortality, association with fertility and migration.

Dynamical Approach

The dynamical approach has been used to estimate incidence in several countries for example, South Africa [20], Dominican Republic, Mali, Tanzania and Zambia [37]. The dynamical methods [2, 3, 35, 38, 39, 40, 41, 42] depend on data from cross-sectional surveys. The survey data is made available by several organisations [11, 12, 13].

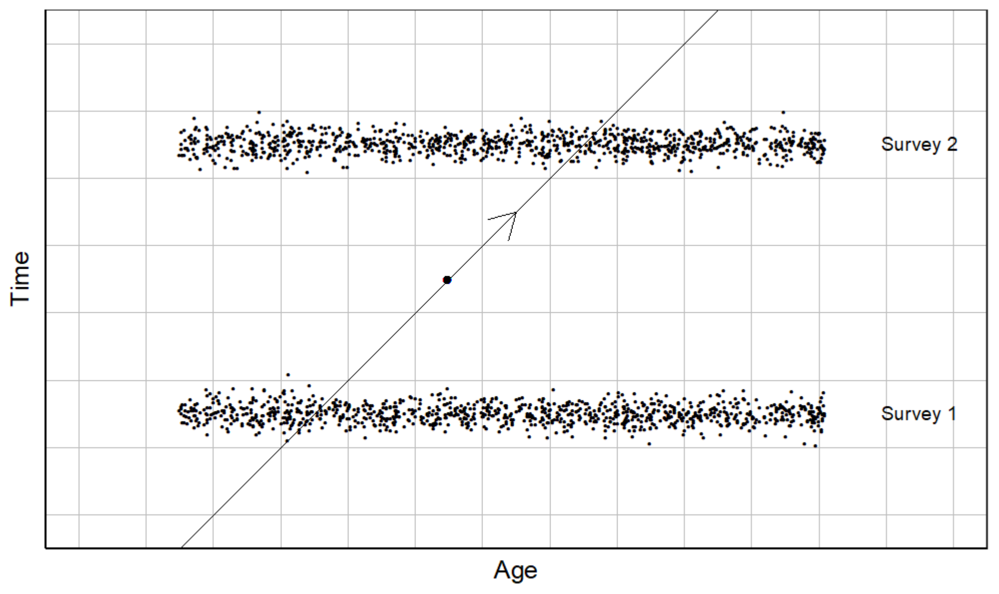

Figure 2.2: *Time/age distribution structure of cross sectional surveys.* Age/time plane indicating relationship between survey-subject interactions and values of (age,time) when incidence estimation can be considered based on those contacts. The many small dots indicate survey-subject contacts generated by two similarly structured population-level surveys. In order to estimate incidence at a point in the inter-survey interval, such as indicated by the bold black dot, one would smooth the HIV status information to provide an interpolated estimate of both prevalence and the rate of change of prevalence, at the (age,time) point of interest, moving along the indicated diagonal line, i.e. as experienced from the point of view of a given birth cohort.

The methods [2, 3, 35] rely on data from two/more cross-sectional surveys to estimate the prevalence and gradient of the prevalence, and in conjunction, with some external information on survival after infection, excess mortality, and/or migration estimates

the incidence. For example, in Figure 2.2 the data is from the two cross sectional surveys shown is smoothed into prevalence and its gradient.
Comparing the Mahiane et al. [3] to the related estimators Hallett et al. [2], Brookmeyer and Konikoff [35], Brookmeyer and Quinn [38], Podgor and Leske [39], Brunet and Struchiner [40, 41], the Mahiane et al. [3] incidence estimator is more informative and accurate, as it measures the instantaneous incidence with no epidemiological or demographic assumptions about any indicators or parameters being constant for ranges of time or having any particular stratification, and no fitting via underlying mechanistic assumptions.

Unfortunately, dynamical methods require accurate information on disease induced (excess) mortality/survival after infection which is complex to estimate, and hence the methods rely on proxies which may introduce bias and uncertainty in the incidence estimates. Improved data sources on survival rates/differential mortality will effectively improve the incidence estimates from the dynamical models [35]. Furthermore, the time between surveys is also a crucial metric in dynamical methods as a reasonable time between surveys is required to derive accurate and informative changes in prevalence i.e., if the time between surveys is long the estimates get biased and if surveys are too close together the incidence estimates are uninformative.

Most dynamical models produce 5 year age aggregated/binned HIV incidence estimates Hallett et al. [2], Brookmeyer and Konikoff [35], Brookmeyer and Quinn [38], Podgor and Leske [39], Brunet and Struchiner [40, 41], of which Grebe et al. [9], has shown that much of what drives the HIV epidemic is in the age structure and this raises the question on how much binning is acceptable without over aggregating the incidence estimates. An advantage of Mahiane et al. [3], is that it facilitates arbitrarily fine grained age-specific incidence estimates - though it is unlikely that there is much information in anything finer than integer age estimates.

Dynamical models directly use large samples sizes being used in any case for major cross sectional surveys being conducted for the purpose of prevalence estimation. Additionally with the current methodological and technological developments data is now readily available and well documented at population level which may improve the robustness of dynamical methods.

Back Calculation

Back calculation methods are similar to the aforementioned dynamical approach, but were mostly used in the 1990s [43], when the only available technology and methodology was the identification of incidence of clinical AIDS cases, and estimates of the distribution of the incubation period from HIV infection to first diagnosis of AIDS [33]. For many countries, where the data are well documented and dates back to the onset of the epidemic, a time series approach based on the AIDS incidence is adopted.

The shortcomings of this approach were cited by Rosenberg [44], Bellocco and Marschner [45]. Brookmeyer and Gail [46] also raised the sensitivity of the back calculation approach to the incubation period. With more understanding of the epidemic, the approach lost most of its attention due to the introduction of anti-retro viral treatment which distorted the assumed distribution of the incubation period, consequently making the AIDS data unreliable [43]. This gave birth to an evolution of this method termed the "extended back calculation" [33, 43], additionally [47] Other causes of concern is the uncertainty of the estimates, challenges of estimating the incubation period and difficulty of modelling the non-stationarity between diagnoses and reporting [48].

Test for 'Recent Infection'

Biomarkers/assays can be used to determine the recency state of an HIV (or, potentially, other) infection. Once the proportion of recently infected individuals among the HIV positive is determined, an incidence estimator that closely resembles that of transient conditions is obtained. To facilitate estimating HIV incidence population level surveys now sometimes include an ascertainment of 'recency' test. The methods of Kassanjee et al. [4], Brookmeyer et al. [17], Brookmeyer and Quinn [38], Janssen et al. [49], Hargrove et al. [50] rely on data with 'recency' ascertainment.

The method provides key advantages that include good local and national HIV incidence estimates, and provides up to date population level incidence estimates vs the retrospective estimation of HIV from the back calculation and dynamical approach. Other advantages involve being able to determine incidence from a single cross sectional survey, and being able to measure incidence trends from at least two or more cross-sectional surveys. Moreover, there is no need for knowledge on excess mortality dynamics, nor need for expensive longitudinal studies [17].

A key challenge is that some HIV patients who have been long infected test 'recently infected'. It is crucial to note that Mean Duration of Recent Infection (MDRI) and False Recency Rate (FRR) are context specific and depend on the bioassay, type of epidemic and HIV virus subtype. Kassanjee et al. [4], Brookmeyer et al. [17] highlight the importance of properly accounting for FRR and MDRI and the one approach that consistently and correctly accounts for this is Kassanjee et al. [4]. This is not a matter of consensus as some scholars (mistakenly) argue that the false recents cancel out with the false non-recents (which don't exist, if the MDRI is correctly estimated). Recent infection testing algorithms (RITAS) were introduced and are structured such that they reduce the false recency proportion to a minimum [4, 17] without eroding too much of the MDRI. The RITAs mainly exclude, from the case definition of recent, those whose viral load is less than a specified threshold, which in practice means mainly those who are on ARTs, who account for most of the false recent results in a typical survey of an advanced epidemic. Very large sample sizes are required to give informative and robust incidence estimates, as the proportion of being recently infected among the positive is usually very low, and hence difficult to estimate with precision [49].

2.4 HIV Surveillance Tools

Several tools meant for HIV surveillance exist and may apply to specific epidemics (concentrated vs generalised epidemics), for example; Thembisa for the South African epidemic [51], whereas a new statistical model called Naomi derives estimates stratified by subnational administrative units, sex and 5-year age groups [52] and case surveillance and vital registration data (CSVAR) is most suitable for high and middle-income countries with low HIV prevalence [53]. This sections briefly recaps and discusses the HIV surveillance tools focusing mainly on HIV incidence estimation.

Spectrum/Estimation and Projections Package − Spectrum/EPP

The Spectrum/EPP model/package is supported by UNAIDS and was first derived from two software tools: Spectrum and Estimation and projection package [54]. The Spectrum/EPP software is a suite of modules each with specific functions. For example, EPP estimates HIV incidence rate primarily from historical prevalence data (Antenatal care (ANC) data scaled up to population level estimates using data from DHS [11] and PHIA [12]). The GOALS module evaluates the efficacy of interventions.

Spectrum/EPP is updated/modified based on the recommendations from the UNAIDS working group [54]. One recurrent theme /specific recommendation from UNAIDS working group biannual meetings is to update EPP/Spectrum to include age-specific incidence estimates, but the best one can do from EPP are 5 year age bins. Eaton et al. [7], modified EPP into be more adaptable to mature epidemics and yield the 5 year age bin/sex stratified incidence estimates.

EPP links the ANC data and population prevalence from the dynamical model through a random effects model and an extra term is used to cater for the bias of ANC data relative to data from the national population-based household survey [55]. For countries without population survey data, the ANC data is used and the confidence interval is estimated from the data derived from the other countries with population survey data [56]. Other interesting aspects in EPP include the estimation of sub national populations based on better data sources [8].

Currently, EPP fits the incidence trends based on four approaches discussed in Stover et al. [8]. which is a major advantage as it provides a suite of tools in one place - i.e., it includes all crucial units for HIV surveillance for easy access, and enables the HIV incidence among countries to be methodologically comparable.

Thembisa Model

The Thembisa model is a tool specifically designed for South African epidemic surveillance and it draws strength and motivation from key previous models, namely: the Actuarial Society of South Africa (ASSA), STI-HIV Interaction model, UCT (University of Cape Town) Paediatric HIV model and the National Strategic Plan ART Need model. These models were each tailored for the South African epidemic, but each had its strengths and limitations [51]. The Thembisa model synthesised the four models and updated them accordingly. For example, ASSA and STI- HIV interaction models failed to reflect HIV prevention and treatment beyond 2011. It also enabled the dynamical evaluation of Mother to Child Transmission (MTCT) which was not possible with UCT Paediatric model [51]. Since the first version of the Thembisa model, about six more versions have been released and each updated to suit context-specific issues. The latest version, Thembisa version 4.3 released in June 2020, made adjustments to the previous model based on the availability of data [57]. For example the updates included assumptions on fertility rates on HIV positive women, the alteration of some provincial calibration procedures, and revision of the viral suppression model after ART initiation [57].

Institute for Health Metrics and Evaluation - IHME

The Institute for Health Metrics and Evaluation (IHME) explores the past, present and future of the HIV/AIDS epidemic in the world, especially in hard-hit areas like sub-Saharan Africa [58, 59]. With the Millennium Development Goals (MDGs) in mind, they produced an online tool - MDG data visualisation - that provides the incidence and mortality yearly rates for each country from 1990 to 2013 [58, 59]. The data visualisation tool is made available through the Global Burden of Disease Study, and the rates are estimated from a modified UNAIDS Spectrum models (for mortality with/without Anti-retroviral Therapy) [59]. Models are adjusted based on the type of epidemic. For example concentrated epidemics are calibrated to fit the recorded data. Whereas for generalized epidemics they use a minimised loss function to select an epidemic curve that fits the prevalence and demographic data for all-cause mortality [59]. The implemented methodologies allow investigating/inferring the lives saved by the Prevention of Mother to child Treatment (PMTCT) and ART programmes [59].

INCTOOLS

This was initially introduced as an Assay Based Incidence Estimator (ABIE) spreadsheet online tool and since evolved into an **R** package - *Inctools* [60]. *Inctools* estimates population-level incidence from survey data with a biomarker ascertainment. The major function that does the incidence calculation uses the prevalence of HIV and the prevalence of recency estimates provided. When raw data is obtained (in the form of subject counts, and potentially stratification and weights) a pre-processing step is required to convert the counts to prevalence and recency estimates [60]. This is essentially a process which currently lies outside of *Inctools*. A naïve method is implemented internally. The findings of this thesis are guiding the developed of more sophisticated tools to perform this derivation of prevalence from raw survey data.

Inctools primarily implements the Kassanjee framework [60]. The standard errors are estimated either through bootstrapping or using closed-form error propagation formulas based on the delta method. Additionally, *Inctools* offers two other functionalities “incprecision” and “incpower” to calculate the power/precision and sample size of the required incidence estimate. In future, *Inctools* seeks to expand the age/time structure dynamical incidence estimator [3].

2.5 Conclusion

HIV incidence estimates are of great importance and we reviewed (not systematically) methods that various countries and organisations use to keep track of incidence, and incidence trends. Of great concern is to yield HIV incidence estimates that are accurate and precise independent of the epidemic. The main focus was on approaches (data collection/methods/soft wares) meant for population level incidence estimates.

There is no overall consensus on the most appropriate method to use when one is in possession of population level data. Important aspects to consider are the representativeness of data, accuracy and precision of the methods chosen to estimate the crucial input parameters, including the excess mortality after infection, and/or recency test properties.

There are various data sources that enable HIV incidence estimation depending on the adopted method. For example some methods rely on;

- ANC data and/or population level survey data - Thembisa and Spectrum/EPP,
- Entire history of the epidemic - back calculation methods.

Some of these methods were not discussed in the main text, but also yield HIV incidence estimates and are context specific and may not be applicable to generalized epidemics, for example,

- Case surveillance and vital registration data - (CSAR) uses vital registration data [53] and most recently
- A new statistical model called Naomi - that produces subnational estimates form population level survey data and ANC data [52].

Chapter 3

Estimating incidence in data with a time domain but no meaningful age structure

3.1 Introduction

Whereas a real population has both age and time structure, HIV surveillance data, from real surveys or thought experiments, may have simplified time-like degrees of freedom. For example, data may be from a single time point but have an age structure. A 'birth cohort' is a group of people born on the same date, leaving us with time structure that codes for age, but no internal variability by age. A population defined by an activity, such as sex work or blood donation, may have age structure which has no particular importance, but the prevalence of risk factors in that population may vary in important ways over time.

Independent of the data structure, the usual key epidemiological measures are still important i.e., questions on prevalence, incidence and incidence trends still need answers. This chapter explores how to get the most out of data with a single time-like dimension, in terms of optimal ways to obtain informative incidence and incidence trend estimates. Using both the frameworks of Kassanjee and Mahiane, we explore the informativeness of survey data under various conditions of simplified time/age structure. This leads to some useful observations that require more complex machinery to explore more deeply, which we do in the subsequent chapters.

3.2 Methods

3.2.1 Simulations

We used a non-mechanistic and unstructured SI-model to simulate an HIV epidemic. The input parameters were incidence, background mortality (death due to natural causes), and excess mortality (HIV (disease) induced death). Based on the prevalence output we simulated surveys at various time points. Our base case had constant parameters, but in order to interrogate study design options, we simulated a scenario with varying incidence, simulated yearly surveys and estimated the relative standard error associated with the Mahiane [3] estimator with an inter survey interval $\Delta t = 5$ years and sample size $= 1000$ at each time point. We also varied sample size and expressed the relative standard error of each age from 18 to 48 in steps of 2 as a function of sample sizes

3.2.2 Comparison of Dynamical vs Recency Approach

We compare the performance of the Kassanjee [4] and Mahiane [3] estimators based on their precision and bias in estimating incidence either from an instantaneous survey at the time of interest, or at the mid-point between two times of observation, respectively. We chose two symmetrically placed time points t_1 and t_2 respectively about t where, $t = \left(\frac{t_1+t_2}{2}\right)$ to achieve comparability of the two study designs. For the Mahiane [3] method, incidence is calculated at t from the two cross sectional surveys simulated at t_1 and t_2. To adapt the Kassanjee [4] method, we simulated a cross sectional survey at time t. To compare the estimators at a comparable level of study-subject interactions, the sum of the sample sizes in the two cross sectional surveys used in the Mahiane analysis was set equal to the sample size of the one cross sectional survey used in the Kassanjee analysis.

The incidence in a birth cohort (i.e. a population with a single time variable which codes for both time and age) using the Mahiane method (I_M) is given by;

$$I_M = \frac{1}{1-P} \cdot \frac{dP}{dt} + M \cdot P \tag{3.2.1}$$

We used a linear regression model to estimate the midpoint prevalence P and the rate of change of prevalence $\frac{dP}{dt}$ from the two simulated surveys at t_1 and t_2 and $\text{var}(I_M)$ is approximated using the delta method. Note M is the excess (disease induced) mortality.

The incidence estimate using Kassanjee method (I_K) ('recency') is given by ;

$$I_K = \frac{P(R-\beta)}{(1-P)\cdot(\Omega-\beta\cdot T)} \tag{3.2.2}$$

We estimated $\mathrm{var}(I_K)$ using the delta method assuming no covariance between P and R (which is probably true in the case of simple random sampling). In Equation 3.2.2, I_K - incidence estimate, R - prevalence of recency among the HIV positive, P - HIV prevalence, β the false recency rate (FRR), Ω - mean duration of recent infections (MDRI) and finally T - time cut-off.

3.2.3 Incidence Trends in a Birth Cohort - Application to First Time Blood Donors

To explore the idea of continuous sampling (ongoing surveillance) we simulated 500,000 first time donors presenting over a period of 10 years- as occurred in South Africa over a period of widening of the donor pool to include other racial groups [61, 62] beyond the previously mainly white donor pool. We simulated testing for HIV and for 'recent infection' with a test like the Sedia Lag Elisa. Data was analysed for incidence either in 'period bins' or continuously, in accordance with a suitably adapted form of the method of Kassanjee et al. [4] - Equation 3.2.2.

Method 1 - Time Binning

Suppose there exists first time donor data from a time period t_1 to t_2 , we define a disruption time t_d (point or period when questionnaire adjustments or intervention programs occurred). The individual level data is divided into two-time intervals $t_1 - t_d$ and $t_d - t_2$. For each interval we summarise the data into prevalence $P = \frac{n_P}{n_P+n_N}$ and prevalence of recency $R = \frac{n_R}{n_P}$. Where, n_N is the total number of FT-donors who test negative for HIV, n_P is the total number of FT-donors who test positive for HIV, and n_R is the number of individuals who test recently infected among the HIV positive blood-donors.

Populating the incidence estimator yields I_1 and I_2, the incidence estimates for the first and second intervals (bins) respectively and for each time bin/interval there exists midpoints defined by m_1 and m_2, which are used as reference points to calculate the incidence slope and associated standard error as shown in Equations 3.2.3 and 3.2.4 respectively,

$$s_{l1} = \frac{I_1 - I_2}{m_1 - m_2} \tag{3.2.3}$$

And the standard error

$$\text{se}(s_{l1}) = \sqrt{\frac{1}{(m_1 - m_2)^2} \cdot [\text{var}(I_1) + \text{var}(I_2)]} \tag{3.2.4}$$

Method 2 - Continuous Time Treatment

We fitted a regression model to the simulated data, and predicted P and R at disruption time (t_d), and to investigate the trend in HIV incidence we evaluated the derivative of Equation 3.2.2 given by Equation 3.2.5 (defined as the rate of change of incidence at disruption time with respect to time).

$$s_{l2} = \frac{dP}{dt} \cdot \frac{(R - \beta)}{(1 - P)^2 \cdot (\Omega - \beta \cdot T)} + \frac{dR}{dt} \cdot \frac{P}{(1 - P) \cdot (\Omega - \beta \cdot T)} \tag{3.2.5}$$

Where $\frac{dP}{dt}$ and $\frac{dR}{dt}$ are the derivatives of P and R with respect to time t.

The standard errors are estimated through 10000 bootstrap samples or the delta method in Equation 3.2.7 we assume $cov(P, R) = 0$.

$$\text{var}(s_{l2}) = \left[\frac{\partial s_{12}}{\partial P}\right]^2 \cdot \text{var}(P) + \left[\frac{\partial s_{12}}{\partial R}\right]^2 \cdot \text{var}(R) \tag{3.2.6}$$

$$\begin{aligned}\text{var}(s_{l2}) = &\left[\frac{dP}{dt} \cdot \frac{2 \cdot (R - \beta)}{(1 - P)^3 \cdot (\Omega - \beta \cdot T)} + \frac{dR}{dt} \cdot \frac{1}{(1 - P)^2 \cdot (\Omega - \beta \cdot T)}\right]^2 \cdot \text{var}(P) \\ &+ \left[\frac{dR}{dt} \cdot \frac{1}{(1 - P)^2 \cdot (\Omega - \beta \cdot T)}\right]^2 \cdot \text{var}(R)\end{aligned} \tag{3.2.7}$$

Data Simulation: Method Testing and Validating

We used the test parameter values suggested in Grebe et al. [9] i.e. $\Omega = 207$, $\beta = 0.001$, $T =$ 2years and based on the parameterisation of the Consortium for the Evaluation and Performance of HIV Incidence Assays (CEPHIA) data.

Below is the generic algorithm for the simulation of FT-donor data;

1. Uniformly distribute the N first time donors to the donation times between t_1 and t_2 such that the time step δ , the time between consecutive donations is given by;

$$\delta = \frac{t_2 - t_1}{N}$$

2. Assign an HIV status to each donation based on $P = P_0 + P_1 \cdot t$, P the prevalence as a function of time, where P_0 is the intercept and P_1 is the gradient of the prevalence.

3. For HIV positive cases we assign the HIV recency status using Equation 3.2.8 and for HIV negative cases we assign Not Applicable (NA).

$$R = \frac{(1-P)\cdot(\Omega - \beta \cdot T)}{P} \cdot I_K + \beta \tag{3.2.8}$$

For simplicity we set the midpoint as the disruption point (questionnaire changes or a new intervention strategies is introduced). Bootstrap was used to calculate the standard errors and we show the distribution of p values for each method (binning versus continuous) for both constant and time varying incidence.

3.3 Results

3.3.1 Reproducibility/Uncertainty

To assess the reproducibility of the Mahiane estimator in the context of a pure birth cohort, we used a hypothetical epidemiological scenario with incidence (0.02 p.a.), excess mortality (0.1 p.a.), and background mortality (0.01 p.a.). The cross-sectional surveys were simulated at times (years) $t_1 = 17$ and $t_2 = 23$ with sample sizes of 5000, to estimate midpoint incidence ($t = 20$ - not that it varies with time in our case). Figure 3.1 presents a histogram that assesses the reproducibility of the Mahiane estimator for 10000 iterations. For 9538 out of 10000 iterations, the 95% confidence intervals included the simulation's true incidence value. In this scenario the Mahiane analyses very slightly underestimates the true incidence, because of the smoothing approach employed, which assumes the prevalence is a linear function of age whereas the simulated prevalence is concave.

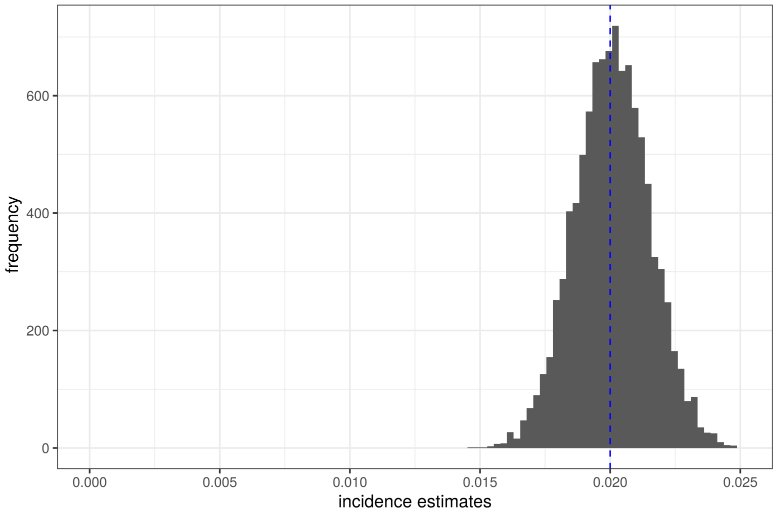

Figure 3.1: *10000 incidence estimates from bootstrap samples*. The distribution of incidence estimates for 10000 iterations using the Mahiane estimator (blue line - true incidence).

Inter Survey Interval

To demonstrate the intrinsic tradeoff of the standard error and bias, in determining the optimal inter survey interval, we used the hypothetical epidemiological scenario presented earlier and expressed the RMSE as a function of inter survey interval.

Figure 3.2 depicts how various values of Δt result in variation of the RMSE, relative bias, and relative standard error; for this specific scenario, the optimal inter survey interval is between 5 and 8 years. Practically, we suggest interpreting this as a 5 year interval being optimal; as the overall error reduction from waiting longer (up to 8 years) is negligible. Optimal Δt implies an acceptable trade-off between bias and standard error.

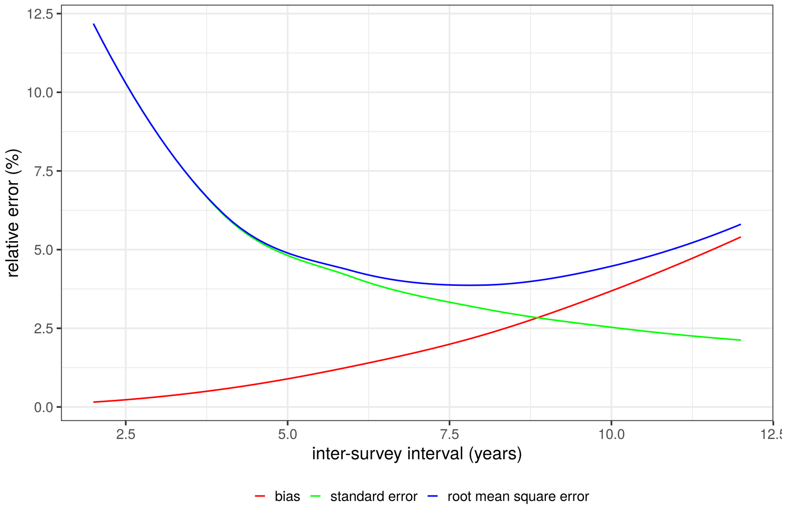

Figure 3.2: *RMSE, relative bias, and relative standard error as functions of the inter survey interval*. The figure expresses the relative errors as function of the inter-survey interval for a specific epidemiological scenario with incidence = 0.02, excess mortality = 0.1, and background mortality = 0.01 (created arbitrary)..

Accuracy and Precision of the Estimators in Mature and Early Epidemics

For a hypothetical scenario, with an underlying incidence of 0.05 and 0.01 in early (left) and mature (right) epidemic respectively, we show (Figure 3.3) the relative errors (RMSE, bias and RSE) as functions of Δt for the Mahiane (solid lines) and Kassanjee estimator (dashed lines).

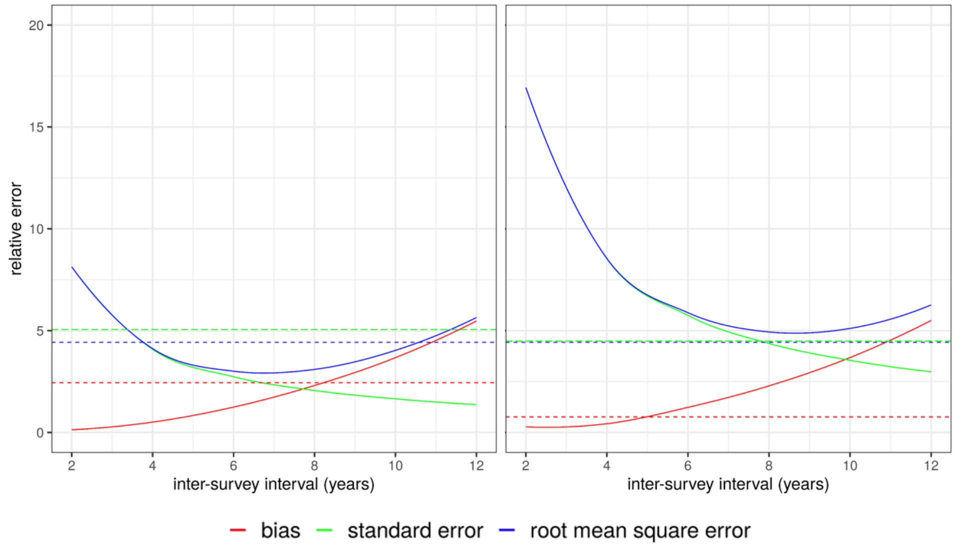

Figure 3.3: *Method performance in estimating HIV incidence at early vs mature epidemic stages.* The figure compares the performance of two incidence estimators (Kassanjee and Mahaiane) at differing epidemic stages early (left) and mature (right) for a range of inter survey intervals (only applicable to Mahiane method). The horizontal dashed reference values represent Kassanjee analysis solid lines the Mahiane estimator.

Based on Figure 3.3, (left panel), the Mahiane estimator is more precise in estimating the incidence in the early epidemic, whereas in the right panel we see that the Kassanjee estimator is more precise in the mature epidemic.

For a more nuanced notion of context, we simulated a birth cohort with a simple time varying incidence, and calculated, over a range of ages, the precision of the Mahiane and Kassanjee estimators, at comparable effort - and at a constant inter survey interval of 5 years for the Mahiane analysis. The results are in Figure 3.4. We also investigated the sample sizes required to attain a specified RSE (15%) for the Mahiane estimate, at each age, as shown in Figure 3.5. Note the required sample size is 12300 for 48 year olds and 1400 for 26 year olds!

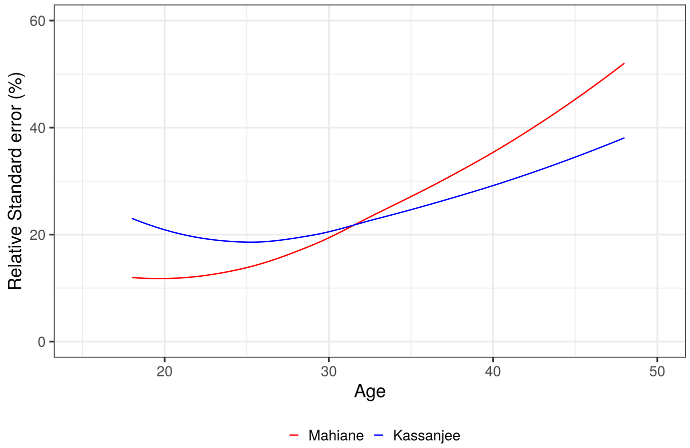

Figure 3.4: *Relative standard error (RSE) versus age*. The figure shows the relative standard error as a function of age for a specific scenario where incidence rises linearly from (1 to 5)% per annum from ages 14 -25, declining linearly to 2% p.a. by age 50. Background mortality 1% p.a. Excess mortality rising linearly from 1% p.a.- 5% p.a. from ages 14 -50, MDRI is 180 days, and FRR is 0.2%. We expressed the RSE as a function of age for both methods (Mahiane and Kassanjee).

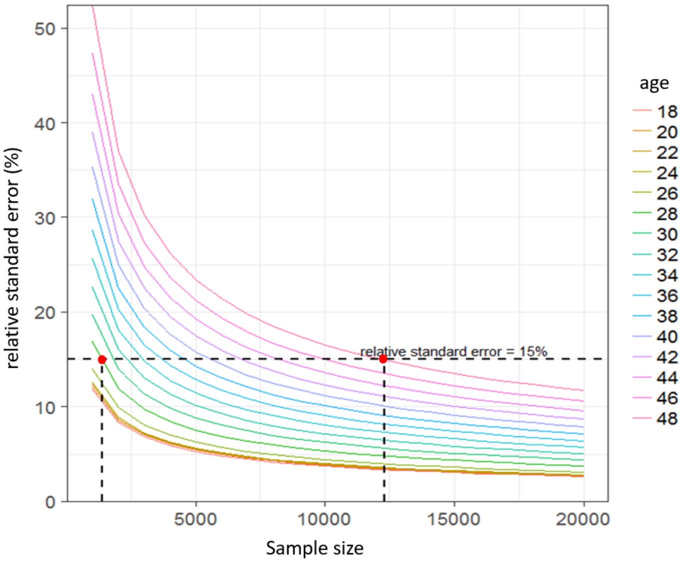

Figure 3.5: *Relative standard error as a function of the sample size for selected ages*. The figure depicts the required sample sizes for various ages to attain a reasonable RSE (of 15%). For example, a sample size of 12300 is required for 48 year olds versus 1400 for 26 year olds i.e., An extra 10900 sampling units are required to obtain the desired RSE.

3.3.2 Incidence Trends in a Birth Cohort - Application to First Time Blood Donors

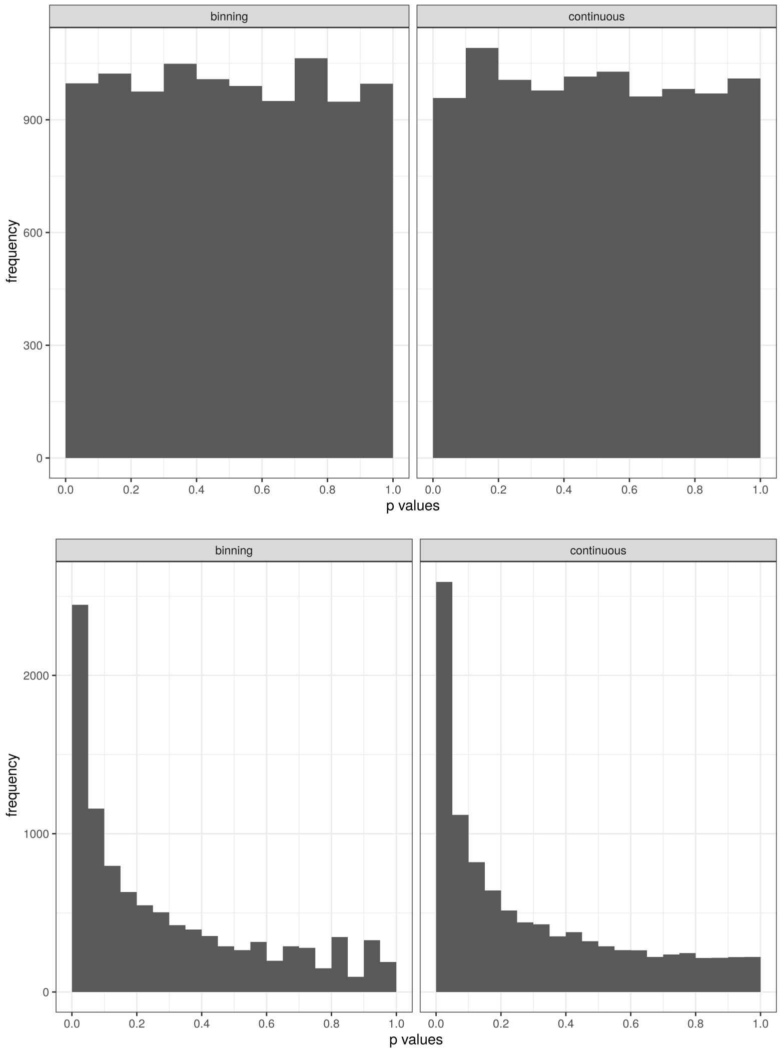

Figure 3.6: *Distribution of p-values for constant incidence (no disruption) vs non-constant (a disruption is introduced) both methods are suitable for their intended use.* Constant Incidence yields p values uniformly distributed between 0 and 1 (top row), contrary, for non-constant incidence the p values are clustered closer to zero (bottom row).

In a simulation of first time blood donors being screened with a test for HIV and a test for 'recent infection' applied to positive specimens, using a simple linear fit to prevalence observations leads to the expected uniform distribution of p values (see Figure

3.6, upper panels). Adjusting the incidence to vary linearly with time show a slight advantage of continuous regression versus binning, with a higher proportion of p values less than 0.05. For 10000 iterations with a true slope of 0.2% p.a the methods estimate, an incidence slope and the 95 percentile range of 0.2004 (0.1310, 0.2698)% p.a - continuous approach and 0.2006 (0.1190, 0.2822)% p.a - binning approach.

3.4 Discussion

Before heading into much more complex simulations and calculations, we focused this chapter on how incidence can be estimated from a population with just the time aspect (where the age ceases to be an interesting structure) or vice versa.

We demonstrated how one can use either the Mahiane [3] or the Kassanjee [4] frame work to estimate incidence at a given time (or age) point. Both Mahiane and the Kassanjee frame work are not meaningfully biased in these simulated scenarios, but depending on the epidemic context, one or other may be more informative at comparable effort. We noted the trade off, applicable to the Mahiane estimator, between bias and precision, which naturally emerges from varying the time interval between surveys. The classic Kassanjee analysis only requires one survey per estimate, so this trade off is not applicable.

In a simplified scenario with age dependent incidence we see that younger ages favoured the Mahiane [3] estimator, and older ages the Kassanjee estimator [4].

To address the need to detect incidence differences, we developed an algorithm that we validated using simulated data and compared two approaches; 1. Binning of observations into an earlier and a later phase. 2. Continuous variability of time. In a simulation of first time donors, we showed that the continuous approach is very slightly more informative about changes in incidence.

Chapter 4

Computational Platform for Scenario Simulation and Analysis benchmarking

4.1 Introduction

This chapter presents aspects of the methodology deployed for the simulation of population dynamics and surveys. We cover the considerations taken into account in choosing the implemented approach, and shed light on the mathematical and computational details. Given our objective, which is to test the performance of approaches to obtaining and analysing survey data, a non-mechanistic model is most applicable. In seeking to test, optimise, validate, and benchmark existing and novel HIV incidence estimation methods, we need merely to have access to a simulated scenario in which we know, and can readily control, the actual incidence and prevalence, so as to produce contexts which resemble real life applications. Hence mechanistic/predictive models are of no particular value to us, and merely complicate calibration. As we do not specifically model heterogeneity, so we implemented an aggregated simulation approach, but the platform does simulate the generation of individual-level datasets.

Specifically, this chapter covers the methods and thoughts that went into;

- Building blocks of the customised HIV epidemic simulation platform:
    1. The population renewal equations and their solution, including detailed accounting for age, time, and time since infection.

2. Integrating over times since infection to obtain observable prevalence of infection, prevalence of 'recent infection', and emergent mortality as a function of age and time
3. The design of the simulation platform in **R**

- Testing (Section 4.6)
- Simulation of cross-sectional surveys

4. Generating survey data.
5. Analysing simulated data exactly as if it were real survey data

- The details of the investigation (Section 4.9)
- Investigating computational demands of the investigation (Section 4.10)

4.2 Population Renewal Equations

In this section, we give a mathematical description of the population dynamics that we considered in creating the simulation platform. We adopt a standard susceptible - infected (SI) compartmental model described by the SI model Equations 4.2.1 and 4.2.2.

$$\frac{\partial S}{\partial a} + \frac{\partial S}{\partial t} = -\left(\lambda(t,a) + \mu(t,a)\right) S(t,a) \tag{4.2.1}$$

Where,

- $\lambda(a,t)$ is the incidence rate defined as the rate of new infections at time (t) and age (a),
- $\mu(a,t)$ the 'background mortality' rate - the rate at which natural death occurs within the population time (t) and age (a), and
- $\frac{\partial S}{\partial t}$ and $\frac{\partial S}{\partial a}$ rate of change in the susceptible population with respect to time (t) and age (a), respectively.

$$\frac{\partial I}{\partial a} + \frac{\partial I}{\partial t} + \frac{\partial I}{\partial \tau} = -\left(\mu(t,a) + \epsilon(t,a,\tau)\right) \cdot I(t,a,\tau) \tag{4.2.2}$$

Where,

- $M(t, a, \tau)$ denotes the excess mortality (death) rate due to HIV, at age a, time t and having been infected for τ – time-since infection ($\tau = a - a_0$),
- $\frac{\partial I}{\partial a}$, $\frac{\partial I}{\partial t}$, and $\frac{\partial I}{\partial \tau}$ denotes the rate of change of the infected population with respect to age (a), time (t), and time since infection (τ), respectively.

Sensible boundary condition for the susceptible population can be defined in various ways. We will specify S(t,0) by asserting a net birth rate, without explicit reference to fertility, for a suitable range of times. We will specify $I(t, a, 0) = s(t, a) \cdot [1 - e^{(-\lambda(t,a) \cdot \delta t)}] \cdot e^{(-\mu(t,a) \cdot \delta t)}$, which denotes the total number of people who get infected, but do not die, in a time step δ.

4.3 Discretised Solutions to the population Renewal Equations

The PDEs are solved using the method of lines and the epidemiological rates are defined as piecewise constant in a given unit grid such that on an $[a : a + \delta] \times [t : t + \delta]$ the same epidemiological rate applies. Consequently, we use the midpoint approximation to estimate the survival probabilities and population counts at specific age and time intervals, for example, the incidence and background mortality rates that apply on a unit grid defined by $[a : a + \delta] \times [t : t + \delta]$, are given by $\lambda \left(t + \frac{\delta}{2}, a + \frac{\delta}{2}\right)$ and $\mu \left(t + \frac{\delta}{2}, a + \frac{\delta}{2}\right)$. Hence at any given age and time, the susceptible population is given by Equation 4.3.1;

$$S(t, a) = S_0 \cdot e^{-\sum_{i=1}^{n} \left(\lambda(t_i + \frac{\delta}{2}, a_i + \frac{\delta}{2}) + \mu(t_i + \frac{\delta}{2}, a_i + \frac{\delta}{2})\right) \cdot \delta} \tag{4.3.1}$$

Where, i is an index for the elements in the age/time vector, n is the length of the age/-time vector, S_0 is the initial population size at birth. Equation 4.3.1 implies that given a birth count - $S(t, 0)$ of S_0 (first term of Equation 4.3.2) and the cumulative survival probability in the susceptible population (second term of Equation 4.3.2), we can estimate the susceptible population in the birth cohort at any given age and time.

Equation 4.3.2, gives the total number of people infected at age $a - \tau$ and time $t - \tau$ (infected at time τ previously).

$$I(t - \tau, a - \tau, 0) = S(t - \tau, a - \tau) \cdot [1 - e^{-\lambda\left(t - \tau + \frac{\delta}{2}, a - \tau + \frac{\delta}{2}\right) \cdot \delta}] \cdot e^{-\mu\left(t - \tau + \frac{\delta}{2}, a - \tau + \frac{\delta}{2}\right) \cdot \delta} \tag{4.3.2}$$

Equation 4.3.2 facilitates the calculation of the total number of infected people who are aged a at time t, survived from natural or infection related death for τ years as shown in Equation 4.3.3.

$$I(t,a,\tau) = I(t-\tau, a-\tau, 0) \cdot e^{-\sum_{i=1}^{n}\left(\mu(t_i+\frac{\delta}{2},a_i+\frac{\delta}{2})+M(t_i+\frac{\delta}{2},a_i+\frac{\delta}{2},\tau_i+\frac{\delta}{2})\right)\cdot\delta}. \tag{4.3.3}$$

4.4 Integrating Out Unobservable Time-since-infection

The population states are simulated whilst keeping track of the time since infection τ, which is a reasonable epidemiological procedure for method development and benchmarking purposes, but τ is not observable and hence when proper accounting of the calculations has been done we aggregate over τ (where applicable) as described below.

Prevalence

The prevalence $P(t,a,\tau)$ is defined as the proportion of individuals still alive aged a at time t, who have been infected for a duration of τ times as shown in Equation 4.4.1;

$$P(t,a,\tau) = \frac{I(t,a,\tau)}{S(t,a)+I(t,a,\tau)} \tag{4.4.1}$$

Aggregating over τ yields $P(t,a)$ the proportion of infected individuals at age a at time t and note $I(t,a) = \sum_{\tau} I\,(a,t,\tau)$ and consequently,

$$P(t,a) = \frac{I(t,a)}{S(t,a)+I(t,a)} \tag{4.4.2}$$

Prevalence of Recency

The prevalence of recency $R(t,a)$ (Equation 4.4.3), estimates the proportion of individuals testing recently infected among the HIV positive. $P_R(\tau)$ gives the probability of testing recently infected given that you have been infected for a time τ for which we used the Weibull function with shape and scale parameter 5 and 0.5, respectively.

$$R(t,a) = \frac{\sum_{\tau=1}^{\tau_{max}} I(t,a,\tau)\cdot P_R(\tau)}{P(t,a)\cdot(S(t,a)+I(t,a))} \tag{4.4.3}$$

Based on the specified $P_R(\tau)$, MDRI (integral of $P_R(\tau)$, from 0 to T and (context specific) FRR (the proportion of individuals testing recently infected but have been infected for over time T (time cut off of recency) are estimated by Equation 4.4.4.

$$\mathrm{FRR}(a,t) = \frac{\sum_{\tau>T}^{\tau_{\max}} R(a,t,\tau)}{\sum_{\tau=1}^{\tau_{\max}} R(a,t,\tau)} \tag{4.4.4}$$

Averaged Excess Mortality

The platform takes in the excess mortality as a function of age - a, time - t and time since infection - τ denoted by $M(t, a, \tau)$, but the Mahiane [3] framework requires the excess mortality averaged over τ (times since infection) - $M(t, a)$, therefore we estimate $M(t, a)$ in equation 4.4.5.

$$M(t,a) = \frac{\sum_{\tau=1} [M(t,a,\tau)] \cdot P(t,a,\tau)}{\sum_{\tau=1} P(t,a,\tau)} \tag{4.4.5}$$

4.5 Simulation Platform

There are many ways to proceed from what has been said so far. For example, one can design the simulator so that $\lambda(a, t)$ emerges from some rules for how people of one age have sexual contact with people of other ages, the prevalence, including detailed distribution of times since infection (coding for infectiousness) providing for infectious pressure. We simply declare the function $\lambda(a, t)$ in a way that does not require looking at the population state - it is just a fully specified function of age and time.

The platform is implemented in **R** and is primarily made up of numerous units that are aggregated into a single function that is called once with the desired input parameters to get an age and time structured population status at given cross sectional survey dates. All the units can be executed independently and hence use of separate units of the platform is possible. The intended use of the platform is to simulate an entire population history at specified cross-sectional surveys and to ensure that the final output is as desired.

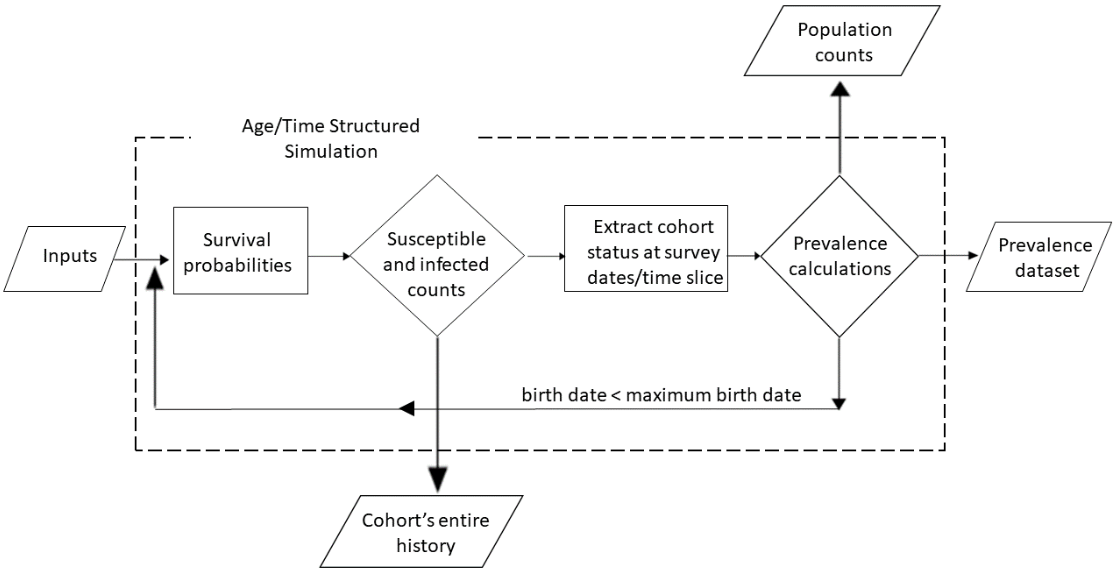
Population counts
Age/Time Structured Simulation
Inputs
Survival probabilities
Susceptible and infected counts
Extract cohort status at survey dates/time slice
Prevalence calculations
Prevalence dataset
birth date < maximum birth date
Cohort's entire history

Figure 4.1: *A depiction of how each calculation is linked to the next calculation in the platform.* The platform is composed of generic, independent, autonomous micro parts that are unified into major functions to yield an age/time structured population. Each units summarises a group of functions found within the umbrella term given to the box.

Given a set of *epidemiological rates* and *parametric adjustments* one birth cohort is simulated at a time up to the specified maximum age.

Model Requirements - Inputs and Outputs

The epidemiological/demographic rates required are;

- Birth rates -$\beta(t)$ and Mother to child transition (MTCT)- $\beta_I(t)$: these are two parameters, are both functions of time t - $\beta(t)$, specifies the total number of individuals born at a given time and $\beta_I(t)$ specifies the total number of infants who test positive independent of the stage of infection out of $\beta(t)$.

  This is just a convenience to help interpret the population state numbers that emerge. We always, for survey purposes, treat the population as effectively infinite, but we may want to implement 'age weighting' in which case it is useful to have meaningful state variables form which we can define a relative population size for different age groups

- Incidence $(\lambda(t,a))$ - specified as a function of age (a) and time (t), extra incidence parameters are specified in the global environment. For example, the shape, and scale parameters of a log-normal distribution.
- Mortality - background and excess mortality - $\mu(t,a)$ and $M(t,a,\tau)$, respectively. The mortality is specified as a function of age (a), and/or time (t), and time since infection (τ) (only for excess mortality).
- Probability of testing recently infected - $P_R(\tau)$ is user defined and is a function of time since infection τ.

In addition to the rate functions the platform requires 'parametric adjustment' parameters (housekeeping parameters)

- Time step δ - is the discretisation step for age, time, and/or time since infection $\delta \in (0,1]$.
- Maximum and minimum date births - DOB_{min} and DOB_{max} specify the date of births of the oldest (DOB_{max}) to the youngest cohort (DOB_{min}) born in the population.
- Reporting bin - specifies whether/not the DOB vector should be a δ step or use the user defined reporting bin, if not specified the birth cohorts are simulated in steps of δ from DOB_{min} and DOB_{max}. For example, reporting bin = 1, means birth dates are spaced 1 year apart but the cohort states are still calculated in steps of δ. The advantage being the discretisation errors are reduced, but still improve the simulation's run time. For the cases were the output's age structure is reported in steps δ (time steps), a function to aggregate the output into the desired reporting bin is available.
- Time cut-off T - time threshold determines when one stops being classified as recently infected.
- Time cut off switch - τ cut-off specifies whether or not the individuals with $\tau > T$ be included in the calculation of the prevalence of 'recent infections' - 'recency'.
- Time slice (cross-sectional surveys) - this is either a single calender dates or vector of calendar dates on which the population prevalence are required.

- Maximum age - a_{max} the maximum age to which each cohort is simulated to, but note that the final output will display the maximum age that corresponds to the last cross-sectional survey (maximum specified time slice). For example if $DOB_{min} = 1980$ and we require the population status in 2015, implies the age range of the population is from 0 to 35 years. Internally the full calculation is done of the state parameters up to age a_{max} for each cohort in question.

Single Birth Cohort Simulation

Figure 4.2 shows how a single birth cohort is simulated based on the implementation of Equations 3 to 9 and a set of inputs provided.

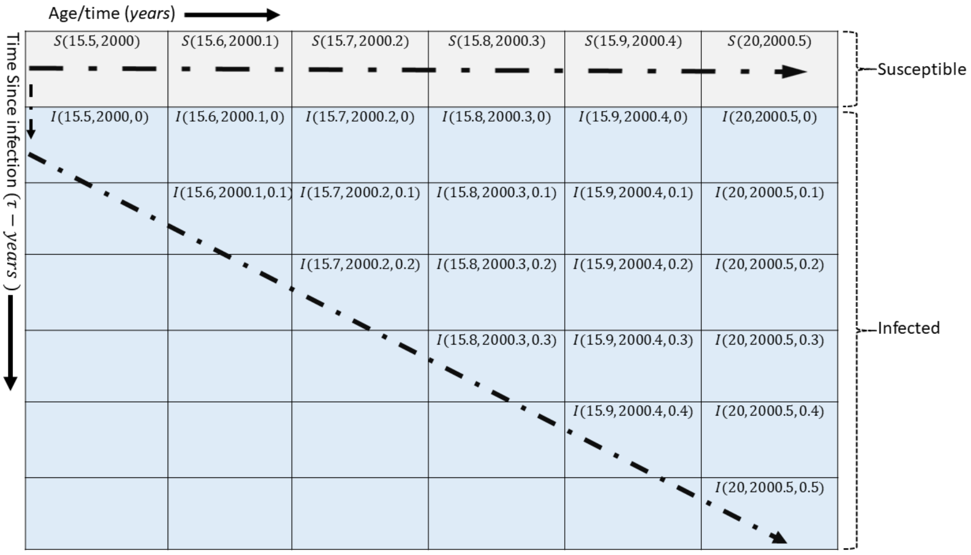

Figure 4.2: *Trajectory of an age cohort with a shared date of birth.* The diagram above depicts a detailed structure/trajectory for a single birth cohort starting from age 15.5 at time 2000, and how it is populated into the data structure for a time step 0.1. An uninfected individual (susceptible) who survives in the uninfected state moves horizontally as age/time passes, whereas an infected individual as time passes moves diagonally in the age-time (a,t) and time since infection (τ) plane. Those who get infected at a given age - a and time-t, for example, a $= 15.5$ and $t = 2000$ move into the infected state (row below the susceptible -$\tau = 0$).

To avoid storing large amounts of irrelevant population state data, once a birth cohort has been simulated to its maximum age, only the population state at specific time points of interest are saved and used to calculate prevalence, prevalence of recency, averaged excess mortality, and FRR, before the next birth cohort is simulated (*see Figure* 4.3 on structured simulations). The output is stored in a dataframe with the corresponding date of birth, age, cohort totals (alive individuals both susceptible and infected) and Non-applicable (NA) is assigned to non-existent population counts/prevalence.

Structured Simulation

Now consider another example where $\delta = 0.1$, Reporting bin = 1, first birth date in 2015 and last births at time 2020 and the required time slice is 2019 and 2020, then a total of 5 birth cohorts will be simulated and with ages ranging from 0.5 to 4.5. Note that for a given birth date the platform assigns all the births occurring at that time to the middle of the year or time step/reporting bin.

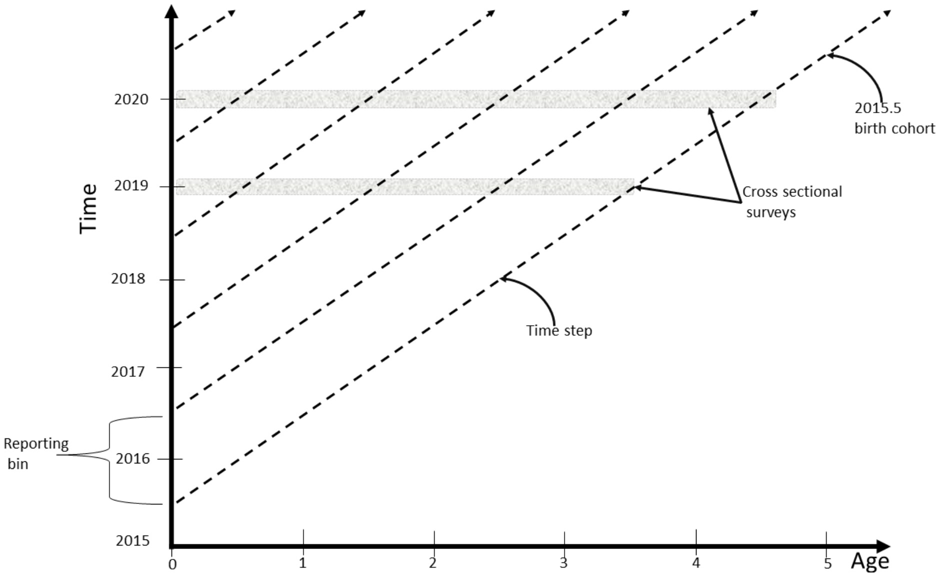

Figure 4.3: *Depiction of the simulation platform output*. For given 'parametric adjustment' parameters, the required time slices are 2019 and 2020 and specified dates of births 2015 to 2020. The output will correspond only to time slices 2019 and 2020 (Grey bands) and the corresponding ages to these time slice from each birth cohort is 0.5, 1.5, 2.5, 3.5, and 4.5 in 2020 and 0.5, 1.5, 2.5, and 3.5 in 2019 (no older ages are reported as there are cohorts simulated before 2015). This plot depicts a specific scenario with time step 0.1 and reporting bin 1, note that births happen in the middle of the year. At each cross-sectional survey, the platform yields a data frame of total population surviving (T), prevalence (P), recency R), averaged excess mortality ($M(a,t)$) and the age specific FRR.

4.6 Simulation Platform Testing

The simulation platform was extensively tested, validated, and benchmarked using custom tests and investigations in **R** before the planned comprehensive investigations. A few crucial tests are described below.

Correctness of Prevalence and Recency Calculations

Consider an epidemiological scenario where a birth cohort is exposed to constant incidence but no mortality then, the closed-form (exact) solutions for the prevalence of HIV (P) and prevalence of recency (R) at any point in time are given by;

$$P = 1 - e^{-\lambda \cdot t}$$

$$R = \frac{\left(e^{(\Omega \cdot \lambda)} - 1\right) \cdot e^{(-\lambda \cdot t)}}{1 - e^{-\lambda \cdot t}}$$

Assuming a constant incidence $\lambda = 0.01$, no excess and base mortality ($\mu = 0$ and $M = 0$) and a date of birth in 1995, we simulate the cohort's history until the age of 25 and compare the output to the closed-form calculations (our benchmark). Table 4.1 shows an extract of the simulation platform's output tabulated together with the closed-form calculations, including the absolute difference for P and R (ages 15.5 to 19.5).

Table 4.1: *Simulation platform output versus closed-form calculation (benchmark).* The Table compares simulation platform's output (time step = 0.01) to the closed form answer of the prevalence and 'recency' and shows the absolute relative difference of prevalence (ΔP) and 'recency' (ΔR).

	Prevalence			**Recency**		
Age	**Platform**	**Closed form**	ΔP	**Platform**	**Closed form**	ΔR
15.5	0.1435848	0.1435848	0	0.0274485	0.0274451	0.0003315
16.5	0.1521063	0.1521063	0	0.0256529	0.0256498	0.0003098
17.5	0.1605430	0.1605430	0	0.0240630	0.0240601	0.0002906
18.5	0.1688957	0.1688957	0	0.0226454	0.0226426	0.0002735
19.5	0.1771653	0.1771653	0	0.0213735	0.0213709	0.0002581

The absolute difference estimated for the P is negligible (≈ 0 %) at a time step of 0.01. Similarly, the R has a relative absolute difference of less than 0.02%. Figure 4.4 shows a detailed comparison of closed form calculation of P and R for all the simulated ages (0.5:24.5), and there clearly is no distinction between the closed-form calculation and simulation platform estimates.

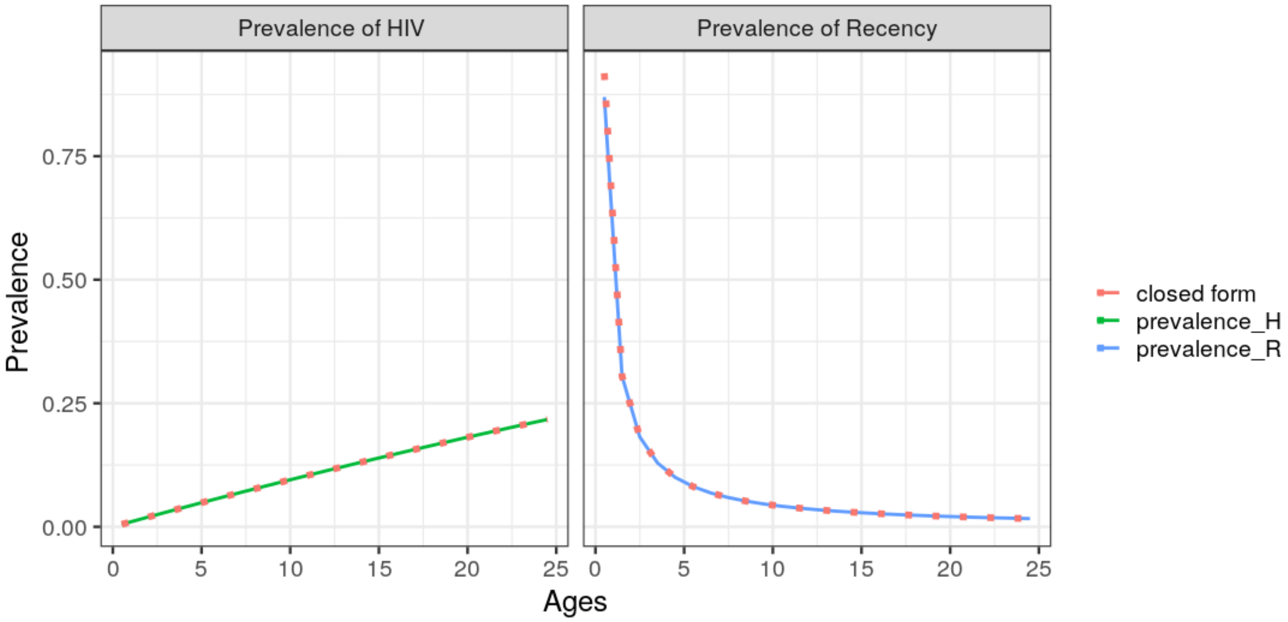

Figure 4.4: *Closed form versus simulation platform output.* A detailed comparison of the simulation platform output to the closed form calculation for a specific epidemiological scenario where incidence is constant and there is no mortality, the reporting bin is 1 and time step is 0.01.

Correctness of Aggregation of Mortality Over τ

We tested the aggregation over τ of the excess mortality calculated by the platform using two approaches:

1. **Test Case 1:** We used the previous test scenario of constant incidence but with constant base mortality (0.01) and no excess mortality. Calculating the average excess mortality gave $M = 0$ as expected.

2. **Test Case 2:** As a spin off of the scenario in test case 1 we introduced a constant excess mortality $(= 0.1)$ and again aggregating over τ yielded a value not greater than the input excess mortality but approaches the given value asymptotically (as expected).

Impact of Choice of Time Step

We used a standard Dell latitude 5490 laptop with RAM - 8 GiB, and processor - quad core Intel (R) core (TM) $i5 - 8350$ CPU at 1.70GH. It is satisfying to note that all desired calculations were doable on this very modest platform, but this did require some streamlining to avoid unnecessary storage demands, and even then, choosing a small delta value (0.001) can lead to inconveniently long run times.

In this section, we investigated and determined the effects of δ (the discretisation time step) on the accuracy of P and R.

- Run time - δ if it is too small the simulation run is increased.
- Accuracy of P and R - reasonably δ should be no more than than a modest fraction of the MDRI (Ω) and realistically available tests currently have an MDRI of about 0.5 years.
- Memory space - depending on the simulation specifications and the machine specifications where the code is being run may not be enough since there is limit on the size of an **R** object (maximum dimension of an array in **R** is $2^{31} - 1$) at present

Given this limitation (bullet 3), the platform does not yield the entire history of the population, but only the population status at pre-specified survey dates - called time slice(s).

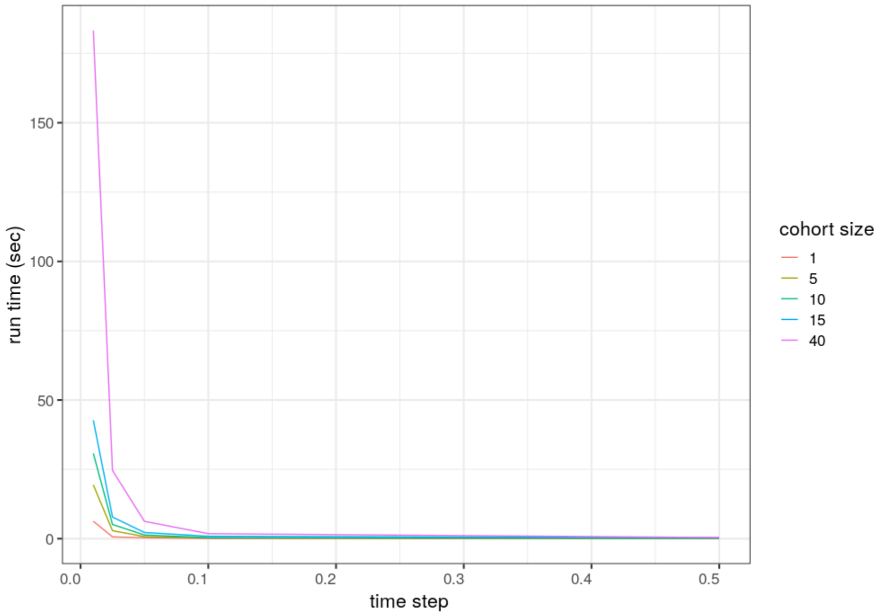

Figure 4.5: *Run time (seconds) of 5 arbitrary selected cohort sizes as a function of the time step.*. Cohort size refers to the number of birth cohorts being simulated and is equal to the length of the date of births supplied, we show a calculation of arbitrary chosen cohort sizes for which we measure the runtime.

Figure 4.5 shows the runtime of the number of cohorts (1, 5, 10, 15, and 40) per simulation as a function of δ. If the population dimension increases (decreased time step size or increased birth cohorts) then this results in an increased run-time. For example, in one cross-sectional survey, suppose we need 40 cohorts with maximum age 50 and time-step $\delta = 0.01$ then the run time is approximately 184 secs (see: Figure 4.5).

Table 4.2, shows the estimates of R at age 20.5 from the simulation platform versus the closed-form calculation. The simulation results are for a single birth cohort, $\delta = [0.01, 0.5]$. When $\delta \leq 0.1$ the percentage relative bias is between 0.012% and 0.013%, i.e., the R values associated with $\delta \in [0.01, 0.1]$ are all equal up to the 5th decimal place. Running big ($\leq$ 10 birth cohorts) simulations suggests settling for a time step (δ) of 0.1 which provides a reasonable trade-off i.e., faster run-time, minimum bias on R and enough disc space.

Table 4.2: δ and the associated run time of the simulation, and the estimated R (birth cohort).

Time step	Run time	Closed form R	Simulation R	Relative bias (%)
0.010	4.43520927	0.02022370	0.02022614	0.01207868
0.025	0.81270123	0.02022370	0.02022615	0.01212643
0.050	0.16710210	0.02022370	0.02022618	0.01229658
0.100	0.05958414	0.02022370	0.02022632	0.01295190
0.500	0.01957679	0.02022370	0.02136397	5.63830692

Effects of Discretisation on Incidence Estimates I_K and I_M Based on Raw Simulation Output

1. **'Recency'** - Using the previously described scenario of constant incidence (0.01) and no base/excess mortality we simulated an HIV epidemic and used *Inctools* (Section 2.4) to estimate the incidence using the simulation platform output values of P and R. We used the incidence estimate obtained from *Inctools* to quantify the error term in γ_8 (an error term in the weighted recency incidence estimator) and compared it to what is expected i.e., the error should be some value plus γ_8. γ_8 (Equation 4.6.2) and I_T (Equation 4.6.1 - the weighted incidence) are defined in Kassanjee et al. [4] and also given below;

$$I_T = I_K \left(1 + \left(\frac{\Omega}{\Omega - \beta \cdot T}\right) \gamma_8 - \beta \left(\frac{T}{\Omega - \beta \cdot T}\right) \sum_{k=1}^{7} \gamma_k \right)^{-1} \tag{4.6.1}$$

where I_K is the incidence estimate (as implemented in *Inctools* - **R**), the second term in parentheses captures the bias (which cannot be evaluated by an experimenter [4]) in the weighted incidence estimator I_K , γ_ks, where $k = 1, ..., 7$ are as defined in Kassanjee et al. [4]. Note since $\beta = 0$, then Equation 4.6.1 reduces to $I_T = I_K(1 + \gamma_8)^{-1}$ implying that $\gamma_8 = e$ and hence we only recap the definition of γ_8, given by Equation 4.6.2.

$$\gamma_8 = \frac{1}{\Omega_T} \int_0^T f_{N_s}(-t) \cdot P_R(t) dt \tag{4.6.2}$$

In the given scenario, $P_R(t) = e^{-\left(\frac{t}{\eta}\right)^k}$ and $f_{N_s}(-t) = e^{\lambda \cdot t} - 1$ therefore γ_8

$$\gamma_8 = \frac{1}{\Omega_T} \int_{-T}^{0} (e^{\lambda \cdot t} - 1) \cdot e^{-\left(\frac{t}{\eta}\right)^k} dt \tag{4.6.3}$$

Given our scenario and the fact that the susceptible population is depleting, we expect to overestimate the incidence by a factor of $(1+e)^{-1}$. Specifically, for incidence $I = 0.01$ and $\Omega = 0.500402$ and solving for γ_8 yields 0.002638103, and hence the bias factor should be $(1+0.002638103)^{-1}$.

The incidence estimate from *Inctools* is 0.01003 which, when we factor out the bias, reduces to 0.01000361 implying that the absolute relative bias for the scenario described above with a time step of 0.01 is 0.0361%. Varying the time step i.e., $0.01 \leq \delta \leq 0.1$ yields an insignificant change in the incidence estimate and hence it does not affect our choice to consistently use $\delta = 0.1$.

2. **Mahiane - 'Synthetic Cohort'** Using the same scenario (constant incidence and no excess mortality) we implemented the Mahiane method [3] to estimate the incidence at the age of 24.5 (arbitrary chosen), using a time step (δ) of 0.01. Using the prevalence $P(24.5)$ and $\frac{dP}{dt} = \frac{P(24.51)-P(24.49)}{(24.51-24.49)}$ the slope of the prevalence. The calculation yields an incidence of $I_M = 0.00997387$ (implying an absolute relative bias of 0.02613%) and hence, based on the output from the simulation platform, there are negligible discretisation errors to compromise the incidence estimation process. Table 4.3 captures the output from the simulation platform used for the calculation of the prevalence gradient at age 24.5.

Table 4.3: *Prevalence output from the platform for the estimation of the gradient.* A single birth cohort is simulated from a constant incidence and no excess/base mortality with a time step of 0.01 to enable estimation of the gradient of the prevalence for I_M and incidence is estimated at arbitrary age of 24.5.

Age	Total	Prevalence	Recency	Mortality
24.49	100	0.2172172	0.0165840	0
24.50	100	0.2172955	0.0165764	0
24.51	100	0.2173737	0.0165687	0

Effects Simple Random Sampling Incidence Estimates I_M

Similarly, we implemented a simple random survey (where everyone has an equal chance of being selected) and estimated the incidence. We set the survey sample size at 1 million and repeated the sampling process 10000 times, and for each bootstrapped sample, estimated the Mahiane incidence. Additionally, we compared the difference in the relative standard errors of the binomial approximation to the bootstrap approach, the absolute relative difference in the relative standard errors was 0.283%. Implying the standard error can be calculated using the binomial approximation. The resulting relative absolute discretisation error on the incidence estimates was 0.0131%.

4.7 Simulating Surveys

Simple Random Sample: Implementation

The survey implemented is an adaptation of simple random sampling; every individual in a given age group has an equal probability of being selected and the heterogeneity is only in the age. A survey simulation proceeds by using the simulation platform output and survey specifications.

We illustrate the survey algorithm using an example. Given the survey specification (Table 4.4) for ages 15 to 25 and prevalence data specifications from the simulation platform (Table 4.5) for a specified time $t = 2012$, with a total sample size requirement of 4000.

Table 4.4: *Example of survey specification.* The survey specifications are given in 5-year age bins that is ages 15-19 (bin 1) and 20- 24 (bin 2) and each bin is such that $a_{min} \leq a < a_{max}$ (unless or otherwise stated).

Age bin	**one**	**two**
Age min	15	20
Age max	20	25
Sample size	2000	2000

Table 4.5: *Example of population state platform output specifications.* The simulated population status for ages 15.5 to 25.5, from a time slice (cross sectional survey) of 2012. Note we chose a $P_R(\tau)$ that ensures FRR $= 0$.

Dates	Age	Population	Prevalence (P)	Recency (R)	FRR	Excess mortality
2012	15.5	816.2374	0.0016967	0.812364	0	0.000008
2012	16.5	802.9602	0.0141436	0.521970	0	0.000039
2012	17.5	789.5816	0.0440682	0.339266	0	0.000111
2012	18.5	776.1079	0.0900368	0.229016	0	0.000249
2012	19.5	762.5352	0.1467310	0.159521	0	0.000483
2012	20.5	748.8428	0.2086098	0.113861	0	0.000849
2012	21.5	734.9878	0.2713802	0.082875	0	0.001385
2012	22.5	720.9011	0.3321801	0.061316	0	0.002135
2012	23.5	706.4916	0.3890993	0.046057	0	0.003135
2012	24.5	691.6565	0.4407839	0.035137	0	0.004415
2012	25.5	676.2849	0.4865116	0.027239	0	0.006003

1. Using the age specific *population* in the simulated prevalence, we partition the age-bin sample size n_i (in this case $n_i = 2000$) in the survey specifications, among the ages in the age-bin.

2. Based on the totals column in Table 4.5 we calculate the proportion of each age in the age bin - age weight (W_a) shown in Table 4.6 (creates an age distribution within an age bin), for example $W_a(15) = \frac{population(15)}{bintotal}$.

Table 4.6: *Simulation platform and the survey specification combined.* The table highlights the internal calculations executed by the survey function to calculate the age weights for each age in the specified age bin.

Survey Specifications		Simulation Platform Output					Calculated	
Bin	**Sample size**	**Dates**	**Age**	**Total**	**Prevalence**	**Recency**	**Bin total**	**Age Weight**
15 - 20	2000	2012	15.5	816.2374	0.001696704	0.812364	3947.422	0.206777
		2012	16.5	802.9602	0.014143647	0.521970		0.203414
		2012	17.5	789.5816	0.044068168	0.339266		0.200025
		2012	18.5	776.1079	0.090036816	0.229016		0.196611
		2012	19.5	762.5352	0.146730963	0.159521		0.193173
20 - 25	2000	2012	20.5	748.8428	0.208609781	0.113861	3602.88	0.207846
		2012	21.5	734.9878	0.271380217	0.082875		0.204000
		2012	22.5	720.9011	0.332180077	0.061316		0.200090
		2012	23.5	706.4916	0.389099349	0.046057		0.196091
		2012	24.5	691.6565	0.440783856	0.035137		0.191973
				7550.302			7550.302	

3. Using a multinomial distribution, we generate an age specific sample size S_i such that $\sum_{i=15}^{<20} S_i = 2000$ (required sample size per age bin in the survey specification). The multinomial distribution is also used on S_i based on P and R to partition S_i into HIV negative (S), recent (R), and long infected (L) counts (as illustrated in Table 4.6).

Table 4.7: *Survey data derived from the survey specifications and simulation platform output.* The table depicts the final output of a survey implementation derived from the simulation platform output and survey specifications. Age weight refers to the proportion of the population in a given age relative to the age bin (5 year age bin).

Simulation Platform Output				Derived sample	Sample breakdown		
Dates	Age	Prevalence	Recency	Age Sample Size	Negative	Recent	Long
2012	15.5	0.001696704	0.812364	414	413	1	0
2012	16.5	0.014143647	0.521970	407	401	3	3
2012	17.5	0.044068168	0.339266	400	382	6	12
2012	18.5	0.090036816	0.229016	393	358	8	27
2012	19.5	0.146730963	0.159521	386	330	9	48
2012	20.5	0.208609781	0.113861	416	329	10	77
2012	21.5	0.271380217	0.082875	408	297	9	102
2012	22.5	0.332180077	0.061316	400	267	8	125
2012	23.5	0.389099349	0.046057	392	240	7	146
2012	24.5	0.440783856	0.035137	384	215	6	163

Variations of the algorithm implementation exist; for example, the algorithm yields individual based data if 'individual' is specified (as true). Additionally, in cases where bootstrapping the survey is not required the expectation functionality can be used in the algorithm (bypasses the multinomial sampling stage) and yields expected values (rounded to the nearest integer). This reduces the runtime considerable and relies on the delta method to estimate the standard errors.

4.8 Survey Data Analysis

We investigated the use of generalised linear models (GLM) to summarise the survey data into P, R, and $\frac{dP}{dt}(= \frac{\partial P}{\partial a} + \frac{\partial P}{\partial t})$ and quantified the errors (relative bias, relative standard error, and relative root mean square error) to determine the optimal ('one size fits most') analysis approach. Below we outline the sequence of steps followed to arrive at an optimal choice.

Determine the Inclusion/Exclusion Criteria/ Data Truncation Rule

The inclusion/exclusion radius r is a value that defines the region/area with the data points of interest (points used to estimate P, R, and $\frac{dP}{dt}$), therefore given an incidence estimation point (a_0, t_0) we specify r such that the data for subsequent step are specified by the range $a_0 - r \leq a_0 \leq a_0 + r$ and $t_0 - r \leq t_0 \leq t_0 + r$ (see Figure 4.6).

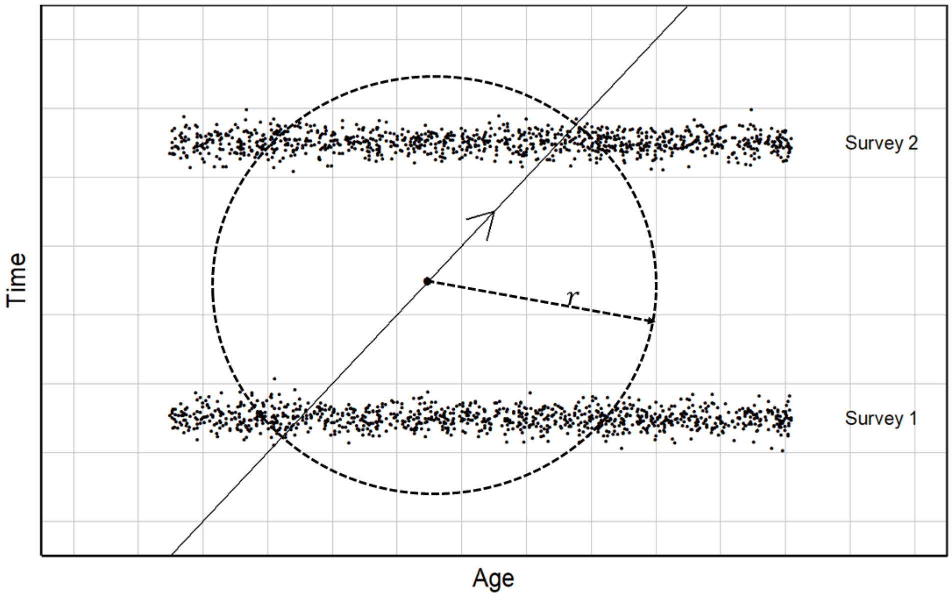

Figure 4.6: *Region with points of interest.* The circle with radius r (inclusion radius) defines the region with the points used in the model fitting.

Fit a Regression Model

We used GLM on the truncated data to estimate P, $\frac{dP}{dt}$, and R. The choice of link functions were 'identity' and 'logit', for P and complementary log log - ('clog-log') for R.

$$\begin{aligned}P(t,a) = P(t_0,a_0) &+ \frac{\partial P}{\partial a}\Big|_{t_0,a_0}(a-a_0)\\ &+ \frac{\partial P}{\partial t}\Big|_{t_0,a_0}(t-t_0) + \frac{\partial P}{\partial t \partial a}\Big|_{t_0,a_0}(t-t_0)(a-a_0)\\ &+ \frac{\partial^2 P}{\partial t^2}\Big|_{t_0,a_0}(t-t_0)^2 + \frac{\partial^2 P}{\partial a^2}\Big|_{t_0,a_0}(a-a_0)^2\\ &+ O(\Delta^3)\end{aligned} \tag{4.8.1}$$

$$\begin{aligned}P(v,z) = P(v=0,z=0) &+ \frac{\partial P}{\partial v}\bigg|_{v=0,z=0} \cdot v + \frac{\partial P}{\partial z}\bigg|_{r=0,z=0} \cdot z \\ &+ \frac{\partial^2 P}{\partial z \partial r}\bigg|_{v=0,z=0} \cdot z \cdot v + \frac{\partial^2 P}{\partial v^2}\bigg|_{v=0,z=0} \cdot v + \frac{\partial^2 P}{\partial z^2}\bigg|_{v=0,z=0} \cdot z \\ &+ O(\Delta^3)\end{aligned} \tag{4.8.2}$$

To estimate P, and $\frac{dP}{dt}$ using the 'identity' link function we used a variable transformation (45° $\frac{\tau}{8}$ anticlockwise rotation - *see Figure* 4.7). This enables the estimation of $\frac{dP}{dt}$ and its standard error in one step, instead of estimating $\frac{\partial P}{\partial a}$, $\frac{\partial P}{\partial t}$, standard errors, and their covariance separately (a simplification only applicable with link function identity). Therefore Equation 4.8.1 and Equation 4.8.2 give the functional forms of the fitted polynomial in the case of no variable transformation and variable transformation, respectively.

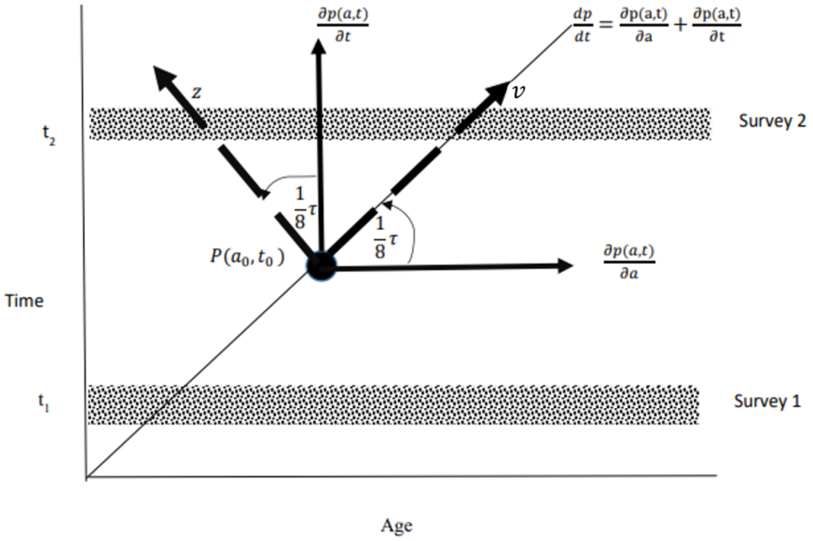
$\frac{\partial p(a,t)}{\partial t}$
$\frac{dp}{dt} = \frac{\partial p(a,t)}{\partial a} + \frac{\partial p(a,t)}{\partial t}$
z
v
t_2
Survey 2
$\frac{1}{8}\tau$
$\frac{1}{8}\tau$
$P(a_0, t_0)$
$\frac{\partial p(a,t)}{\partial a}$
Time
t_1
Survey 1
Age

Figure 4.7: *Variable transformation.* The Figure shows the axis transformation about the point of interest (a_0, t_0) where incidence is to be estimated and when link function is 'identity' in GLM, and note $\tau = 360°$ and only the points within the region defined by r are transformed. The solid arrows represent the original age and time axis whereas the solid arrows represents the new axis after transformation.

Similarly the logit link function can be used to estimate the prevalence and its gradient, but in this case axis transformation offers no advantage and hence the data is used as is. The details of the logistic regression are discussed in greater detail in Chapter 6.

The recency (R) is estimated using the GLM methods with link function 'clog log' (unless stated) and both the estimate of R and its standard error are readily available from the GLM fitting process and hence no extra calculations are required.

Incidence Estimation

I_K, and I_M are as previously introduced in Chapter 3, and I_{Opt} is the optimally weighted incidence estimator derived from augmenting I_K, and I_M, using the inverse variance method. Not much about these incidence estimators is covered in this chapter as each one of them has a Chapter dedicated to the method.

We populate the incidence estimators I_K, I_M and I_{Opt} previously discussed, with P, $\frac{dP}{dt}$, and R. The optimal inclusion distance and polynomial order are determined by varying the various permutations of the two qualitative choices and at each step we determine the relative errors of I_K, and I_M.

The standard error for each method is estimated through the delta method or repeating the survey 10000 times (resample one survey data 10000 times) and for each sampled survey the incidence is estimated. The 95 percentile range is then estimated for the 10000 incidence estimates. This is ideal for complex surveys as the bootstrapping captures the complex sampling strategy and helps circumvent the challenges of estimating the various covariance pieces required in the standard error of the incidence. In cases where bootstrapping is not necessary the standard error and the covariance are estimated using the delta method (see Appendix 4.11). Where applicable further investigations are done using an optimal permutation of polynomial order and inclusion distance and these included;

- Inter survey interval for I_M
- Incidence trends (incidence differences from 2/3 cross sectional surveys)
- Post hoc averaging to improve incidence estimate's precision and compared to the true average incidence estimate

- appropriate weight measure .

4.9 Investigation Flow of Analysis

Figure 4.8 highlights the internal core of the computational machine. An execution is initiated by a complementary choice of 'scenario', which involves setting the initial population state, incidence and mortality, and survey or study designs. The simulation is run, and the model world survey conducted to 'generate data'. The selected method is used to 'Analyse' the modelled survey data into incidence (point estimates, confidence intervals, trends, posteriors, etc.).

If there are concerns about the robustness of formulas for variance, the generation of data sets can be repeated, as indicated in Loop 1. Having established performance in one scenario, Loop 2 enumerates a choice of scenarios on which performance is to be benchmarked. Loop 3 involves the selection and implementation of variations in qualitative analytical model choices, or parameters. The final step involves evaluating performance as a function of conditions, to determine the best choice for a kind of scenario, and describe easy or difficult scenarios for a given method.

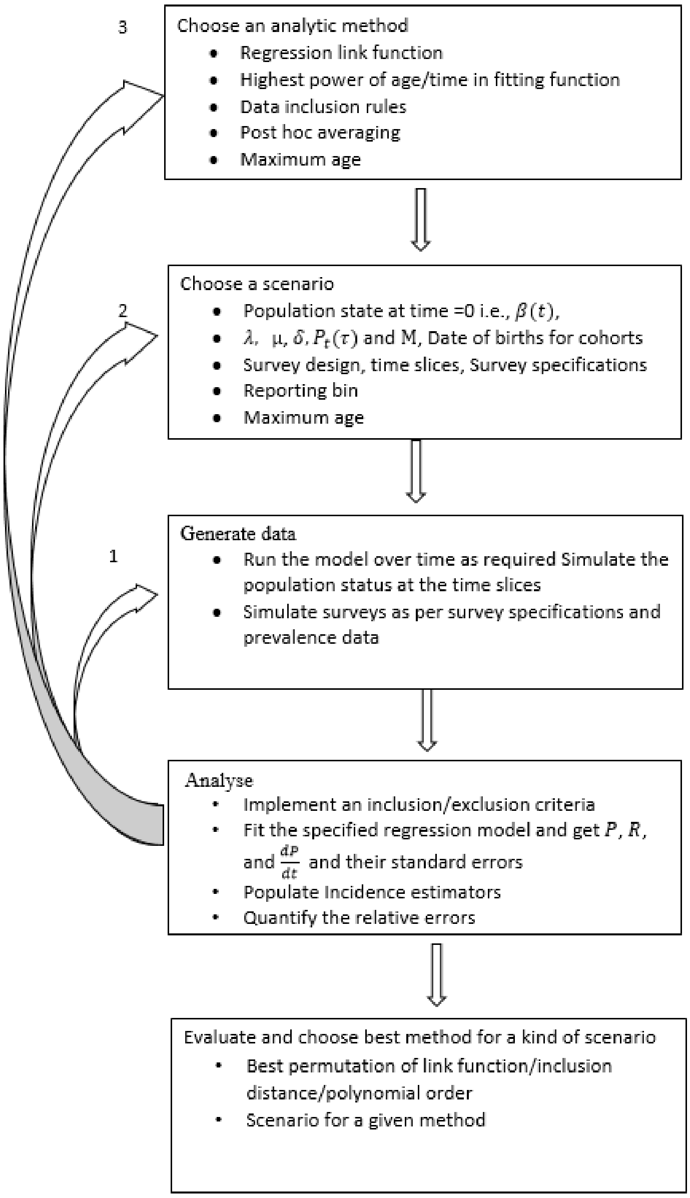
3
Choose an analytic method
Regression link function
Highest power of age/time in fitting function
Data inclusion rules
Post hoc averaging
Maximum age
2
Choose a scenario
Population state at time =0 i.e., $\beta(t)$,
λ, μ, δ, $P_t(\tau)$ and M, Date of births for cohorts
Survey design, time slices, Survey specifications
Reporting bin
Maximum age
1
Generate data
Run the model over time as required Simulate the population status at the time slices
Simulate surveys as per survey specifications and prevalence data
Analyse
Implement an inclusion/exclusion criteria
Fit the specified regression model and get P, R, and $\frac{dP}{dt}$ and their standard errors
Populate Incidence estimators
Quantify the relative errors
Evaluate and choose best method for a kind of scenario
Best permutation of link function/inclusion distance/polynomial order
Scenario for a given method

Figure 4.8: *The Investigation flow.* Highlights the internal mathematical computation and core of the platform. We show how the different methods and scenarios are investigated and evaluated. Loop 1 denotes the random iterations to measure reproducibility, loop 2 points to the selection of a new scenario to evaluate the performance of the system under varied epidemiological/demographic functions and finally loop 3 is the selection of qualitative choice and parametric adjustment.

4.10 Profiling the Investigation Algorithm

We implemented a basic profiling algorithm to investigate the time associated with the key blocks in generating and analysing the data. The analysis algorithm has two main aspects that were of concern to us;

1. Survey implementation and
2. Modelling P, R, and $\frac{dP}{dt}$.

For each unit we investigated the run-times for 10^1, 10^2, 10^3, and 10^5 iterations to yield run time per execution of the survey and the parameter estimation process (call to GLM).

Table 4.8: Execution time in (seconds) for the survey implementation versus the call to GLM

	iterations	survey runtime (sec)	GLM runtime (sec)
1	10^1	0.72	0.1956592
2	10^2	4.22	0.5040493
3	10^3	37.46	3.2780149
4	10^4	376.98	31.6990335
5	10^5	3166.84	307.7739689

Table 4.8, Execution times for iterations for the survey implementation versus the parameter estimation, the analysis algorithm takes longer to generate surveys versus the call to GLM and the magnitude increases rapidly with increased iteration numbers.

To counter the runtime challenge we implemented parallelisation and in some instances, where bootstrap was not necessary, we used the delta method to approximate the 95 percentile range of the incidence estimates. The advantage being that delta method requires a single call of the algorithm with expectation set to TRUE.

4.11 Discussion/Conclusions

The simulation platform is versatile and is primarily meant to mimic the HIV epidemic without any social interactions incorporated (non-mechanistic, non-predictive and non-explanatory), but can be used to simulate any chronic condition epidemic. The platform is meant to test, compare, and validate incidence estimation methods i.e., facilitates method development. The platform takes epidemiological/demographic rates and parametric adjustment parameters to yield the population state at a given time and for all ages up to the maximum age.

This is not a recreation of an existing tool as we desired specific features and we could not find a tool with the desired specifications i.e., tracks the time since infection, allows the specification of $P_R(\tau)$, granular tracking of τ and tracks the excess mortality as a function of τ.

The high level points that were covered in this section include;

1. The simulation can run seamlessly on standard computers with no need for high performance or cloud computing as long the required output is of a reasonable size and the data store at each calculation point does not exceed the disc size limit in **R**.

2. The quantified discretisation errors were shown to be almost negligible ($\delta = 0.1$) compared to closed form approximations and hence the simulation output can be used in further analysis with discretisation errors being the least of our concerns. The objective was to ensure that the platform offers discretisation errors less than at least 1%.

3. P and R were estimated accurately and hence it translates to an accurate implementation of the other micro units that make up the platform -such as the counts (susceptible and infected populations) and their cumulative probabilities.

4. The incidence estimates I_M and I_K from the raw prevalence estimates with uncomplicated epidemiological rates yielded relative bias estimates less than 1% for both no sampling (prevalence values used as is) and simple random sampling.

5. The profiling investigation highlighted that the survey simulations greater run time compared to the data smoothing process. Hence we estimated the standard error using either bootstrapping (with parallelisation) and in some instances the delta method.

4.A Appendix: Parameterising Epidemiological/ Demographic Rates

All epidemiological rates are functions of either age, or time, or time since infection, and hence all the parameters crucial to the functional forms of the selected epidemiological/demographic rates, parametric choices, and qualitative parameters are predefined in a separate R script called *Global variable script*. The analysis script sources global available script first before executing the investigation in question. Users of the platform have the option to define their own functions, but should adhere to the requirement of ensuring that the rates are strictly functions of a, t, and τ. Below we present some of the utility functions used in the simulation;

Birth Rate

The total counts of birth is essential to the population simulation, and we do not liken it to any particular population growth rate in the world but treat this rate as a tool to enable us to execute the simulations. The implementation is such that all the births happen at a particular time (birth date) and if the time step is δ, then the births occur at $\frac{\delta}{2}$. All simulations used a constant birth rate - $\beta(t)$. If the functional form of the birth rate is complex (not constant) the other extra parameters will need to be defined in the *global variable script*.

Incidence

We use a log normal distribution to parameterise the incidence as a function of age a - $f(a)$ (Equation 4.A.1) and $f(a)$ is rescaled by R_s (mode for the specified log normal distribution) so that it lies between 0 and 1. Specifically:

$$f(a) = \begin{cases} 0 & a < a_0 \\ \frac{1}{\sigma_a \cdot (a-a_0)\sqrt{2\pi}} \cdot e^{\frac{-(\log(a-a_0)-\beta_a)^2}{2\sigma_a}} & a \geq a_0 \end{cases} \tag{4.A.1}$$

Where a_0 is the age at which incidence departs from 0 for the first time - 14 years in our case. β_a is the mean of the log distribution, σ_a is the standard deviation of the log distribution and we choose parameters $\beta_a = 2.3$ and $\sigma_a = 0.5$ so that the peak incidence is always attained at age 20.

Similarly, in the case were we parametrise the $P(t)$ (the peak incidence experienced at time t) to follow a log normal distribution, then $P(t)$ is given by;

$$P(t) = \begin{cases} 0 & t < t_0 \\ \frac{1}{\sigma_t \cdot (t-t_0)\sqrt{2\pi}} \cdot e^{\frac{-(\log(t-t_0)-\beta_t)^2}{2\sigma_t}} & t \geq t_0 \end{cases} \tag{4.A.2}$$

Where t_0 is the initial time when incidence is introduced, t is the time of interest, σ_t is the shape parameter (standard deviation of the log distribution) in time and β_t - is the mean of the log normal distribution. We set the parameters $\beta_t = 2.5$, $\sigma_t = 0.325$, $t_0 = 1985$ for times $t \in [1985, 2030]$.

We also made available the $P(t)$ function presented in Mahiane et al. [3], which is a piecewise linear function in time i.e., linearly increases in time up to the maximum specified incidence from a given time, remains constant at the maximum value for a stipulated interval, and then gradually declines for another given time interval until it reaches some stipulated incidence value and remains constant thereafter (see Mahiane et al. [3]). The overall incidence function is given by;

$$\lambda(t, a) = R_s \cdot f(a) \cdot P(t)$$

The incidence function is a function of a and t. The shape, scale, and other parameters are defined in the *global variable script* as previously stated.

Base mortality

The base mortality in all the simulations is parameterised using a linear function of age (only) starting at 1% to 3% per annum for ages 0 to 50 and remains constant thereafter. The platform requires that you supply the base mortality as a function of time and age and it is possible to use a function of either of the two variables.

Excess Mortality and Averaged Excess Mortality (Platform Output)

The platform requires the excess mortality as a function of age, time, and time since infection. We have several excess mortality utility function, namely;

1. Constant,

2. Function of age (a) and time since infection (τ) i.e., internally it is a function of age at infection a_i (see Mahiane et al. [3])

3. Function of age (a), time (t) and time since infection (τ) i.e., internally it is a function of age at infection a_i and t

Below we give details on the excess mortality used in the simulations that are meant for the Mahiane et al. [3] investigations. The excess mortality in bullet 2 and 3 above were parameterised using a Weibull functional form, with a shape parameter ($\omega = 2.28$) scale parameter, $\alpha(a,t)$ which is a function of age and time and specifies the median survival times for infected individuals.

We defined the median survival time using a linearly decreasing function of the age at infection (a_i) between the maximum (max_j) and minimum (min_j) median survival times and is $j = 1,2$ depending on the time-pre-treatment era ($j = 1$) versus treatment era ($j = 2$), respectively

$$\eta_j(a_i) = \begin{cases} max_j & a_i < a_0 \\ \frac{max_j - min_j}{a_0 - a_{max}} \cdot (a_i - a_0) + max_j & a_0 \leq a_i \leq a_{max} \\ min_j & a_i > a_{max} \end{cases} \tag{4.A.3}$$

Since $j \in (1,2)$, this implies that we have $\eta_1(a_i)$ and $\eta_2(a_i)$, such that the overall scale parameter of the excess mortality rate is a piecewise function given by Equation 4.A.4.

$$\alpha(a,t) = \begin{cases} \eta_1(a_i) & t_i < t_0 \\ \frac{\eta_2(a_i) - \eta_1(a_i)}{t_{max} - t_0} \cdot (t_i - t_0) + \eta_1(a_i) & t_0 \leq t_i \leq t_{max} \\ \eta_2(a_i) & t_i > t_{max} \end{cases} \tag{4.A.4}$$

Where t_0 is the time treatment starts and t_{max} is the time the median survival times become constant. For our simulations, we defined the excess mortality;

$$M(a,t,\tau) = \omega \left(\frac{1}{\alpha(a,t)} \right) \left(\frac{\tau}{\alpha(a,t)} \right)^{\omega - 1} \tag{4.A.5}$$

Note that $M(a,t,\tau)$ is a function of a, t and τ, and has internal functions($\eta_j(a_i)$ and $\alpha(a,t)$). $M(a,t,\tau)$, firsts calculates a_i and t_i and then parses them to the internal functions ($\eta_j(a_i)$ and $\alpha(a,t)$).

Recency Function $P_R(t)$

This function is meant to calculate the probability of testing recently infected given that you tested HIV positive. We parameterised this probability using a Weibull distribution function as it describes the behaviour of most biomarkers. The function is required only as a function of time (t) and the shape and scale parameters are specified in the global variable script. The MDRI is calculated in the *global variable script* and is dependent on your choice of recency function.

4.B Appendix: Global Variable Script

Below we show an example of a global variable script with all the epidemiological/demographic rates functions, their extra parameters, qualitative choice and investigation stipulations.

```
########################################################################
# Global Variables to Control
# - Demographic/Epi Simulation
# - Survey Simulation
# - Survey Data Analysis
########################################################################

####################################
# BEGIN Simulation Housekeeping
####################################

time_slice =  seq(1980, 2025, 2.5)
max_age = 50
min_birth_date = 1935
max_birth_date = 2025
time_step = 0.1
reporting_bin = 1

####################################
# END Simulation Housekeeping
####################################
```

```
####################################
# BEGIN Survey Design
####################################

# survey_speccs

alteredspecss  <- data.frame(dates = c(rep(time_slice[1], 10)),
samplesize = c(rep(samplesize, 2),samplesize, rep(samplesize, 7)),
agesmin = (c(0, 5, 10, 15,20, 25, 30, 35, 40, 45)),
agesmax = (c( 5, 10, 15, 20, 25, 30, 35, 40, 45, 50)))

####################################
# END Survey Design
####################################

####################################
# BEGIN Incidence Estimation Housekeeping
####################################
estimatemethod = ‘‘Mahiane"/ ‘‘Kassanjee"/ ‘‘Optimally"
recency = T
method_weight = NULL

# incidence prediction point
age_predict = c(18, 22,  32, 42)
time_predict = 2002.5

# incidence anchor points (when data is to be transformed)
agetimetrans = TRUE
anchorage = age_predict #15.5:45.5
anchortime = 2002.5

# Taylor series expansion
overall_weights = FALSE
prevtaylororder = 1:4
rectaylororder = 1:4
```

```
# GLM link functions
prevlink = "logit"
reclink = "cloglog"
psilon = 0.01

# Inclusion  criteria
inclusion_distance = seq(2,6, 1)
timecutoff = 5 # when method is distance
inclusionmethod = "distance" # can be circle/distance

# Mahiane method extras
excessmortality_se = 0

# Kassanjee methods extras
RSE_MDRI = 0.00001
FRR = 0
RSE_FRR = 0.00001
BigT = 730

# extras
expectation = T
iterations = 1
individual = F # aggregate if individual is TRUE
points_weights = F
####################################
# END Incidence Estimatin Housekeeping
####################################

####################################
# BEGIN   Demographic rate functions
####################################
####################################
#  Fertility Functions
# (GV_fertility)
```

```
fertility_function = constant_birth_rate
GV_fertility_constant = 10000

#  MTCT Functions
# (GV_mtct)

mtct_function = constant_pmtct_rate
GV_mtct_constant = 0

# constant_pmtct_rate
# GV_mtct_constant = 0

###################################
#  Incidence Functions
# (GV_inc)
incidence_function = lognormalage_time_incidence_stretched
#age specifications
GV_inc_min = 0
GV_age_critical = 21
GV_stretchfactor = 2
GV_inc_betat = 2.3
GV_inc_sigma2 = 0. 5

#time specifications
GV_inc_time_debut =  1985
GV_inc_betat = 2.5
GV_inc_sigma2 = 0.324575
GV_timestretch_factor <- 2
GV_time_critical <- 1998
###################################
# HIV-free mortality functions
base_mortality_function =  time_indep_age_linear_base_mortality #(Linear in Age)
GV_basemort_constant = 0
GV_basemort_agemin = 0
GV_basemort_agemax = 50
GV_basemort_min = 0.01
```

```
GV_basemort_max = 0.03

###################################
#  HIV-positive mortality functions

excess_mortality_function = timeage_instanteneous_excessmortality
GV_treat_starts = 2005
GV_treat_constant = 2015
GV_maximum_survival_scale = 40
GV_minimum_survival_scale = 15
GV_maximum_survival_scale_pretreat = 15
GV_minimum_survival_scale_pretreat = 5
###################################
# recency functions
# (GV_prt)
recency_function = weibull_recency
GV_prt_scale = 0.5
GV_prt_shape = 5
GV_prt_bigt = 2

# MDRI_calculation
MDRI  = (calculate_MDRI(recency_function = weibull_recency) * 365.25)
####################################
# END   Demographic rate functions
####################################
```

Chapter 5

Optimal accounting for age and time structure of HIV incidence estimates based on cross-sectional survey data with ascertainment of 'recent infection'

5.1 Abstract

Background

Many surveys have attempted to estimate HIV incidence from cross-sectional data which includes ascertainment of 'recent infection', but the inevitable age and time structure of these data has never been systematically explored - no doubt partly because statistical precision in such estimates is often insufficient to allow for satisfactory disaggregation. Given the non-trivial age structure of HIV incidence and prevalence, and the enormous investments that have been made in such data sets, it is important to understand effective ways to extract valid age structure from these precious data sets.

This chapter is available as a preprint: authors: Mhlanga, L. Grebe, E. Welte, A., Title: Optimal accounting for age and time structure of HIV incidence estimates based on cross-sectional survey data with ascertainment of 'recent infection', DOI: 10.21203/rs.3.rs-871044/v2

Methods

Using a comprehensive demographic/epidemiological simulation platform developed for this, and some wider, purposes (documented in more detail separately) we simulated a complex 'South Africa inspired' HIV epidemic, with explicitly specified 1) age/time dependent incidence, 2) age/time dependent mortality for uninfected individuals, and 3) age/time/time-since-infection dependent mortality for infected individuals. In this simulated world, we conducted cross-sectional surveys at various times, and applied variants of the recent-infection-based incidence estimation methodology of Kassanjee et al. We analysed in considerable detail how to smooth, and average over, the age structure in these surveys to produce the incidence estimates, paying attention to the fundamental trade off between bias and statistical error.

Results

We summarise our detailed observations about incidence estimates, generated by various age smoothing or age disaggregation procedures, into a straightforward fully specified 'one size fits most' algorithm for processing the survey data into age-specific incidence estimates: 1) generalised linear regression to turn observations into 'prevalence' of 'infection' and 'recent infection' (logit, and complementary log log, link functions, respectively; fitting coefficients of up to cubic terms in age/time); 2) a 'moving window' data inclusion recipe which handles each age/time point of interest separately; 3) post hoc age averaging of resulting pseudo continuously fitted incidence; 4) bootstrapping as a generic variance/significance estimation procedure.

Conclusions

As far as we are aware, this is the first analysis of several fine details of how age structure in cross-sectional surveys interacts with recency-based incidence estimation. Our proposed default estimation procedure generates incidence estimates with negligible bias and near-optimal precision, and can be readily applied to complex survey data sets by any group in possession of such data. Our code is available, in part freely through the R computing platform, and in part upon request.

5.2 Introduction

Population-level cross-sectional surveys, including HIV status determination, are conducted routinely in many Sub-Saharan countries. Within the last two decades, many of the surveys include administering a 'recency' test to consenting individuals that have tested HIV positive. Defining a transient 'recency infection' state, among the HIV positive group, allows for the derivation of an HIV incidence estimator that resembles that of transient conditions.

Various methodologies for incidence estimation, based on 'recency' ascertainment [4, 38, 50, 63, 64] have been proposed. We will use the framework of Kassanjee et al.[4], which, we would argue, is the formally correct approach.

We envisage a 'recency' state that is fundamentally defined through standardised and validated objective laboratory procedures, sometimes known as a Recent Infection Testing Algorithm (RITA), or Test for Recent Infection (TRI). For technical details, we strongly recommend a close reading of the seminal derivation of the estimator [4] and for initial efforts to investigate age structure, we suggest looking at Grebe et al. [9].

A typical RITA (applied only on sensitively and specifically classified HIV infected respondents) defines'recency infection' as having:

1. A lower-than-threshold immunological marker (like antibody titre, avidity, or HIV-specific component fraction) and
2. A non-negligible Viral Load, defined by some threshold

These two typical components of the test serve the following functions:

1. The serology marker acts as a rough biological clock indicating duration of infection, and
2. The viral load marker rules out stable treatment (addressing the fact that consistent viral suppression typically rolls back the naive infection-time clock)

Details of thresholds are a subtle matter involving challenges in the development of an appropriate dynamic range for the assaying procedures employed, and some optimisation based on analysis of results obtained on substantial panels of well curated specimens [65].

The 'recency infection' case definition is reflected in the estimator via two parameters:

1. Mean Duration of Recent Infection (MDRI): the average time individuals are classified as recently infected on a given RITA, all while having been infected for a time less than some convenient bookkeeping cut-off T [65].

2. False Recent Rate (FRR): the proportion of the long-term infected individuals (those infected for longer than the bookkeeping recency cut-off T) that are ('falsely') classified as recently infected [9, 65, 66].

MDRI and FRR are, unfortunately, context-dependent, varying by such factors as dominant circulating virus subtypes, antiretroviral treatment coverage, and detailed epidemiological factors like current and recent history of incidence. For a RITA to be of significant value in the context of realistic survey sizes and currently envisaged contexts of application:

- The MDRI needs to be of the order of half a year [74], and subject only to minor variation between times and places for which incidence estimates are to be compared.

- The FRR needs to be below 1% [74]. i.e. the probability of false recent result, among 'definitely long infected cases', meaning infected for longer than T, must be reliably known to be less than 0.01.

There are a number of significant loose ends on matters of optimisation of analysis. This despite the fact that: HIV incidence estimation has been of high interest for several decades, considerable work has been done to extract such estimates from data sets gathered at great effort and cost, and there is a semblance of consensus that the Kassanjee incidence estimator (I_K) is the only formally rigorous and consistent approach to such estimation.

In particular, there is no general understanding of how to analyse and interpret the non-trivial age structure of HIV survey data. In outline, the present work has the following high-level components:

1. Simulating 'realistic' epidemics and cross-sectional surveys;

2. Applying either categorical criteria or smoothing algorithms to the survey data, in order to infer (age- and time-structured) prevalence of HIV infection, and prevalence of recent infection amongst HIV positive subjects;

3. Estimating incidence, and incidence differences/trends, from these smoothed functions, using the Kassanjee framework;

4. Evaluating the relative merits of various smoothing and averaging schemes, by comparing estimated with the known incidence parameter values in the simulations; and

5. Proposing a generic one-fits-most approach to the main use-cases of incidence estimation.

This paper is the first of three companion pieces looking at a closely related set of variations on the theme of smoothing survey data to optimally extract the age/time structure for the purposes of estimating HIV incidence.

5.3 Methods

5.3.1 Computational Environment

All computations were performed in the **R** system for statistical computation [67], and required only an ordinary laptop/pc hardware platform. Core evaluation of the Kassanjee estimator and its variance was largely performed by functionality in the **R** package inctools [60], available on CRAN.

5.3.2 Simulations

We used a customised simulation platform that requires only emergent epidemiological rates:

- birth rates as a function of time,
- incidence as a function of age and time ,
- background mortality as a function of age and time, and
- disease associated mortality as a function of age, time and time-since-infection.

This means we do not need to specify (i.e. 'make assumptions about') mechanistically detailed processes like contact rates, mixing rules, etc. in order to simulate an HIV epidemic that resembles what has been observed in generalised epidemics, such as, in South Africa. This platform, which is described in detail separately [68], allows us to track the age, time, and time-since-infection structure of the prevalence of infection/disease (interpreted as HIV), as well as the age and time structure of the prevalence of 'recent infection' (sometimes abbreviated to 'recency'). Using these simulated prevalences, we then simulated cross-sectional surveys (notably in 1990, 1995, 2000, 2005, 2010, 2015, and 2020) with varied sampling densities as a function of age (notably: uniform sampling density per year of age, and sampling density proportional to population density per year of age). The details of the functional forms which we used for the age, time, and, where applicable, time −since-infection dependence of the demographic process parameters are found in Appendix 5.A.

5.3.3 Estimating Incidence from One Cross-sectional Survey

$$I_K = \frac{P(R-\beta)}{(1-P)(\Omega-\beta\cdot T)} \tag{5.3.1}$$

Where, P is the HIV prevalence, R is the prevalence of recency among the positive, Ω is the Mean duration of recent infections(MDRI) and β, is the false recency rate (FRR), and T is time cut-off. A delta method-based formula for variance/standard error of incidence estimates has been derived, and this replicates very closely the values obtained by either bootstrapping a data set, or outright repeating (the simulation of) the entire survey process.

5.3.4 Estimating an Incidence Difference from Two Cross-sectional Surveys

When there are two cross sectional surveys, the Kassanjee estimator can be directly applied separately to the data set from each survey. However, the estimation of incidence differences requires some care, as the two estimates are usually not entirely independent. Most typically, even if the prevalence estimates are completely independent, at least the estimates of the recency test properties (Ω and β) are not independent. The point estimates of MDRI and FRR may be exactly the same numbers, derived from the same background analysis. In this case, they would be perfectly correlated. The FRR is almost inevitably somewhat different between any two contexts, but since this difference may not be directly estimated, it may still be rational to treat it as 'estimated once'-and this is how we proceed in the present analysis. Alternatively, MDRI and FRR may be similar numbers estimated by slight contextual adaptations to a shared base estimate obtained for a shared biomarker - in which case they would be significantly correlated in a way that would need to be analysed on a case-by-case basis. If the two recent infection tests are based on different biomarkers, whose properties are estimated on independent (or sufficiently large) specimen panels, it would be reasonable to treat MDRI estimates as independent.

In the present investigation, we consider the possibility of analysing the combined data set from two surveys in one single regression, to obtain smoothed prevalence from as much data as possible. In this case, the correlation in prevalence estimates will be complicated, and probably not readily estimated by means other than brute force bootstrapping. However, rather than estimating the correlation as an independent parameter, and then propagating the implications of that parameter into an incidence difference variance formula, we propose that it is generically more robust, and always computationally feasible, to generate bootstrapped data sets by resampling the full data, and thus generating a large number of incidence difference point estimates, from which a standard error can be obtained. This makes the most sense for real world data sets, where a little computational delay is the least of many challenges faced by investigators.

5.3.5 Binning Approach to Estimating P and R

Typically, one uses all the data from the survey to produce an incidence estimate for the entire population, or one divides the data according to 5 year-age bins. An exception is Grebe et al. [9]. The reason for this is largely the size of the data set, which usually leads to very uncertain incidence estimates for small age ranges. When binning data, we use a binomial exact function in **R**.

5.3.6 Regression Approach to Estimating P and R

A key point in the present investigation is the exploration of the performance of various regression models to summarise the survey data into $P(a,t)$ and $R(a,t)$. We considered linear models and differing in:

- the link function,
- observation/data inclusion/exclusion criteria, and
- polynomial order in powers of age/time.

5.4 Results/Discussion

All incidence estimation was done in the simulated 'South Africa −like' epidemic alluded to above, and described in more detail in the Appendix 5.A

5.4.1 Ignoring Age Structure

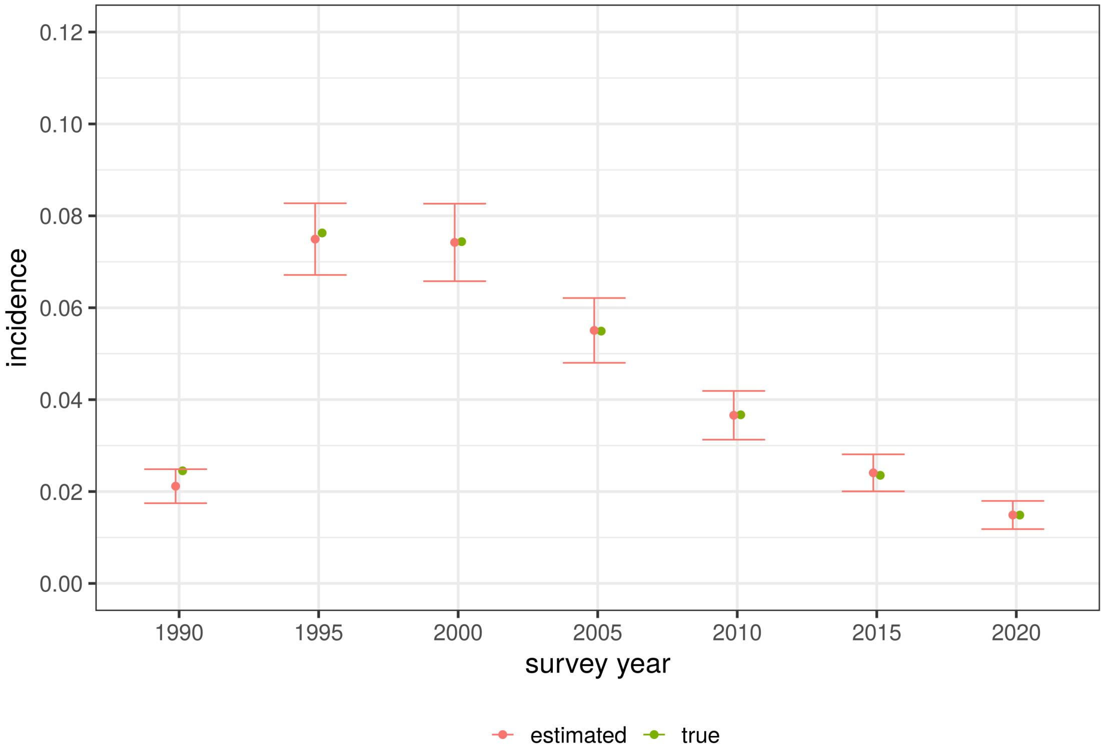

Figure 5.1: *Representative incidence estimates derived by treating an entire survey data set as one large age bin (red) shown alongside the 'true' incidence of the surveyed population i.e. age-weighted to the susceptible population (green)).* From each five−year age bin (15 − 19, 20 − 24, . . . , and 40 − 44) 4 000 individuals were sampled with equal probability (total sample size of 24 000).

Figure 5.1 shows a common way in which incidence estimates are derived and presented. All the data from a cross−sectional survey is used to derive one incidence estimate, without regard to age structure. We see that this approach has little bias and reasonable precision. An important caveat when treating a survey data set as one large age range is that the complex age structure of the incidence is hidden and no highly-at-risk ages can be identified.

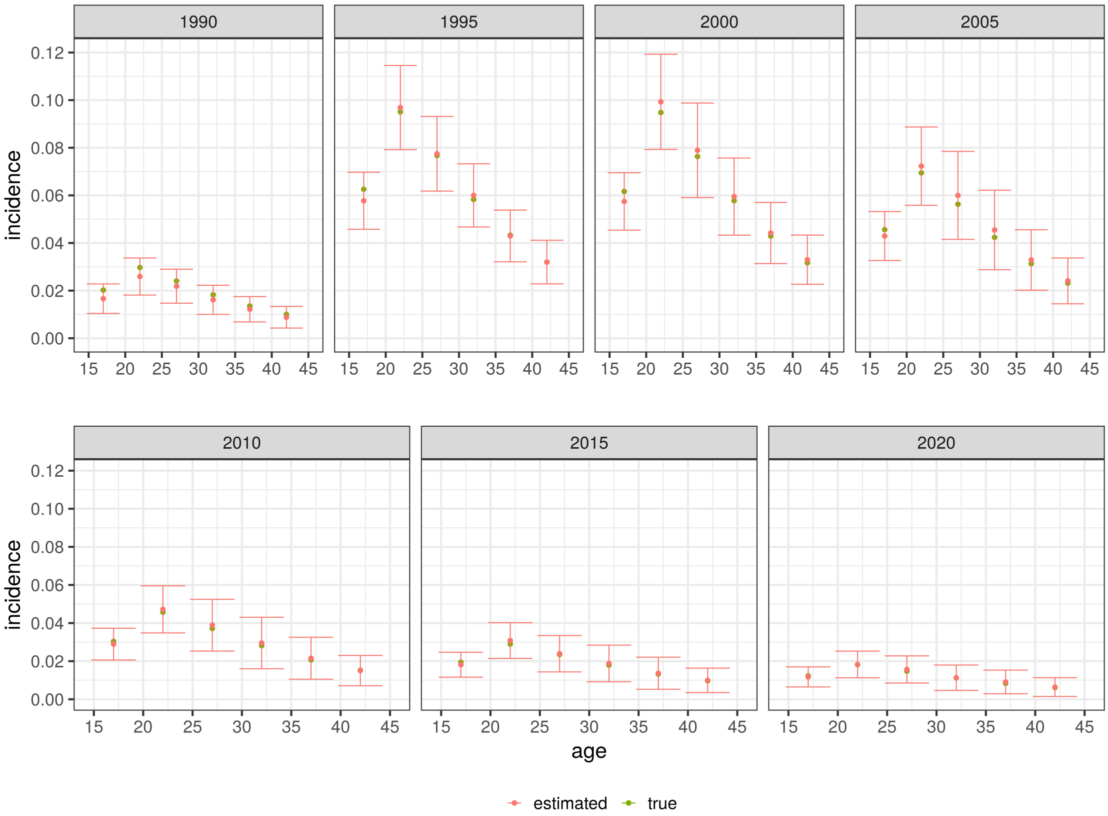

Figure 5.2: *Incidence estimates derived by decomposition of each survey data set into 5-year age bins (red) shown alongside the 'true' incidence within each age bin i.e. age−weighted to the susceptible population within each bin (green).* From each five−year age bin (15 − 19, 20 − 24, . . . , and 40 − 44) 4 000 individuals were sampled with equal probability (total sample size of 24 000).

5.4.2 Regression

A crucial part of the present investigation is understanding how one might extract prevalence estimates (of HIV and recency) from survey data in the form of well fitted functions of age, and how this might be 'optimised' for the purposes of estimating incidence. We considered permutations of link function, data inclusion rules, and polynomial order (powers of age) of the fitting function. Based on preliminary investigations, we:

- set the default link functions of P and R to logit and *complementary log log*, respectively,
- computed P and R separately for each integer age, by performing a fit of data 'sufficiently close' to the age of interest (defined simply by an age difference cut−off),and
- explored in great detail the choice of inclusion distance and polynomial order. To avoid undue proliferation of permutations, we always used the same values of these parameters for the calculation of both P and R.

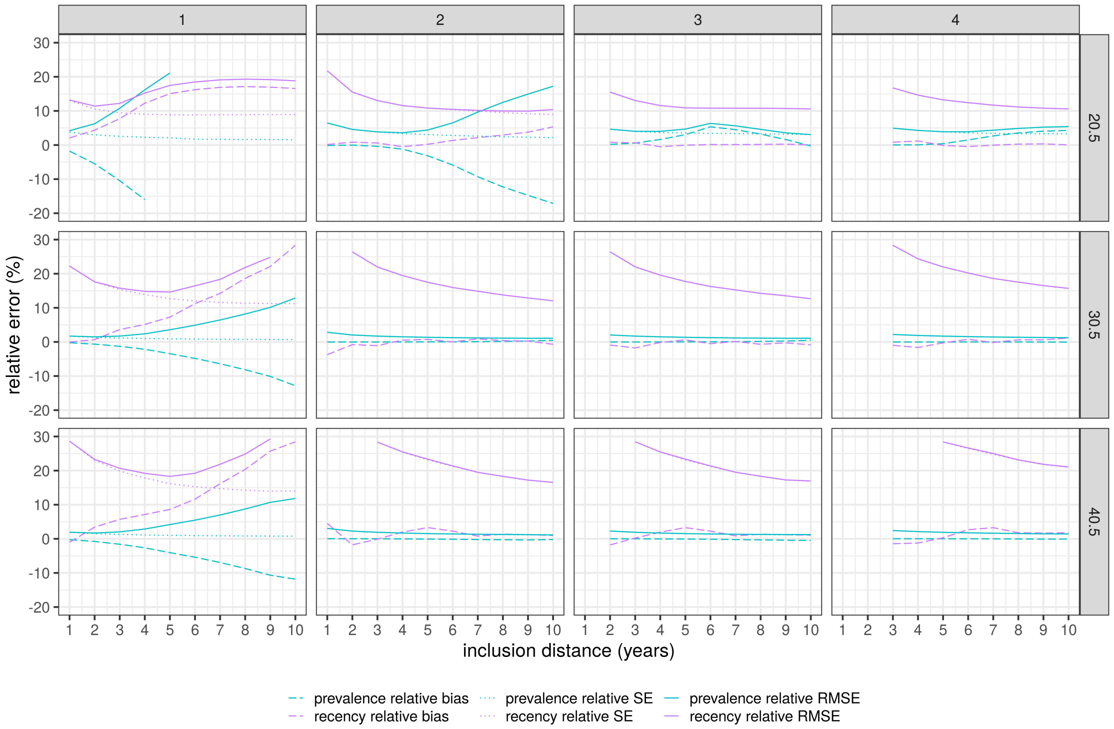

Figure 5.3: *Relative bias— relative standard error— and relative root mean square error— at age 20.5— 30.5— and 40.5 (rows)— using polynomial orders* 1 − 4 *(columns)— in each case as a function of data inclusion radius (x−axes).* These errors are based on the cross−sectional survey simulated in 2015— with a sampling density of 4000 per 5−year age range.

We executed all combinations of polynomial order and inclusion distance, for each integer age from 15 to 44, using data from each of the surveys conducted in 1990, 1995, 2000, 2005, 2010, 2015, 2020. This led to too many individual results to present here. We demonstrate some key features in the body of this article, and display additional results in the Appendix 5.A.

Figure 5.3 shows (percentage) relative 'errors' associated with estimating P and R, using variations (defined by age polynomial order and age inclusion radius) on this regression approach, applied to the 2015 survey data. The errors are colour coded turquoise and lilac for P and R, respectively. Each line type represents one of the 3 relative errors: dotted – standard error, dashed – bias, and solid – total root mean square error. Note that only bias has a meaningful (plus/minus) sign. Each row shows relative errors for a single age, as labelled. Each column corresponds, as labelled, to the polynomial order 1 – 4, and the x–axes of individual plots represent the inclusion distance.

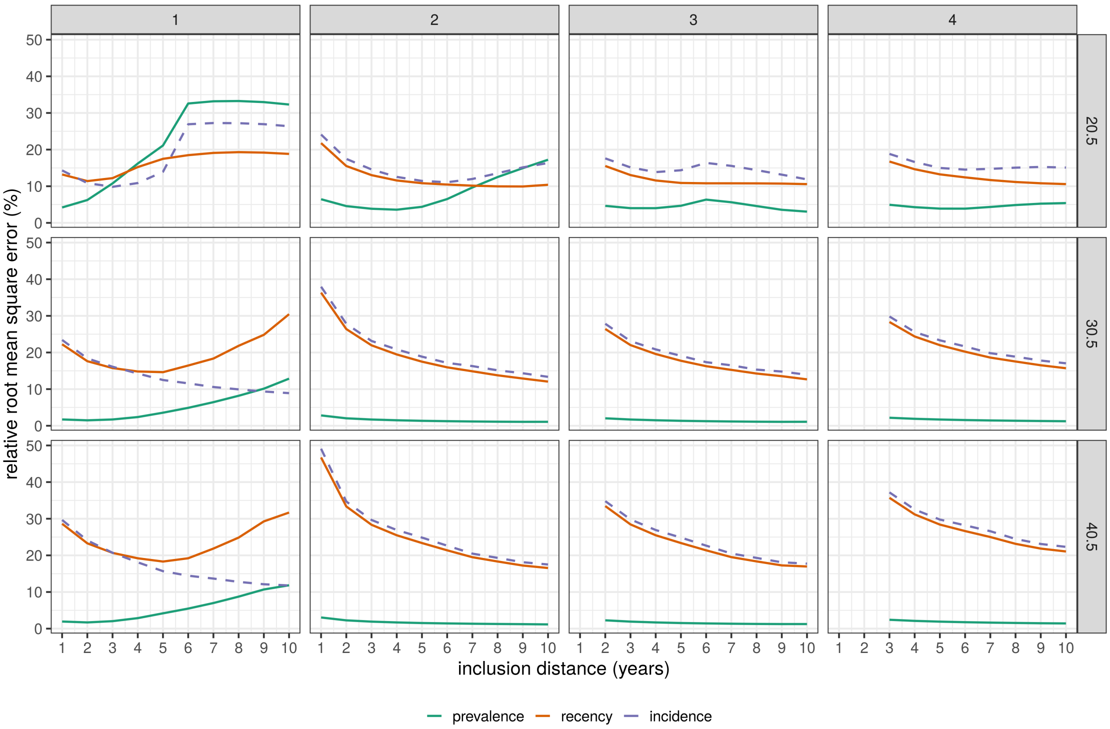

Figure 5.4: *Relative root mean square error for estimated prevalence, recency, and incidence.* These errors are based on the cross-sectional survey simulated in 2015 on the canonical scenario- with a sampling density of 4000 per 5-year age range.

Figure 5.4 shows just the relative root mean square errors, but in addition to the errors in Prevalence and Prevalence of Recency among positives, also shows the error in the estimate I_K, for the same ages presented in Figure 5.3. Evidently R is the main source of errors in I_K as the relative root mean square error of R largely tracks that of I_K. Instead of generating a large number of plots similar to Figure 5.4, by displaying individual results from all permutations of polynomial order, inclusion distance, age, and survey round (in our canonical scenario described above) we can summarise the relative root mean square errors into one plot, showing the distributions of relative root mean square error for P, R and I_K, as shown in Figure 5.5.

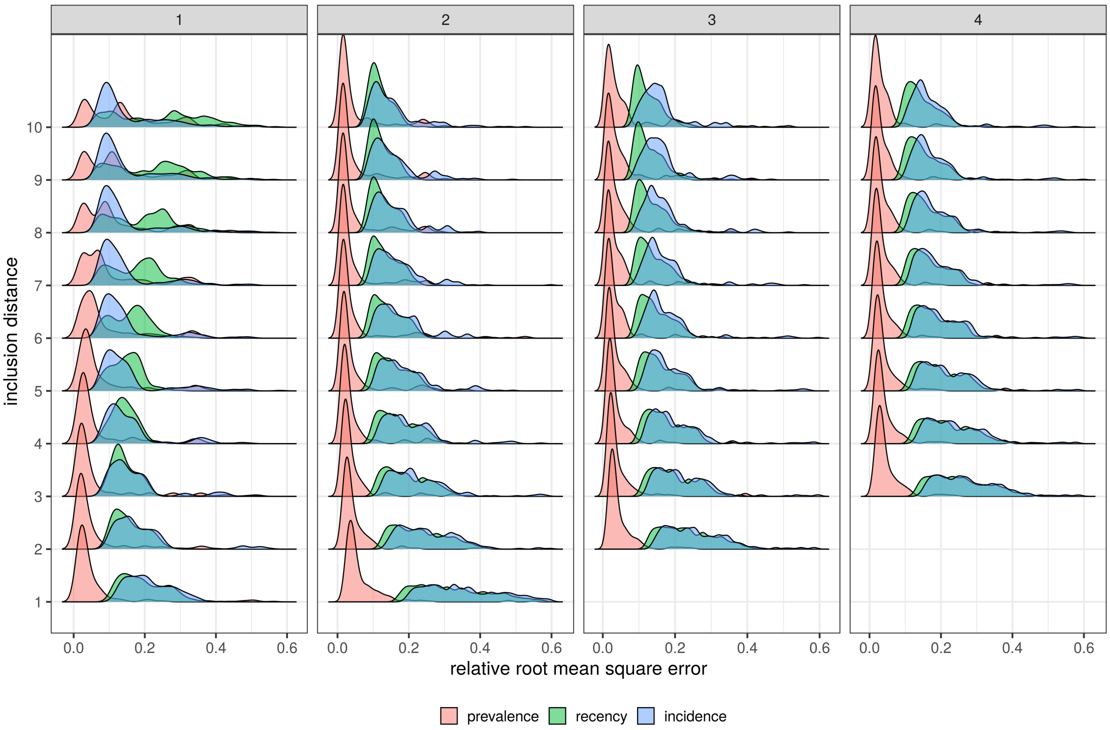

Figure 5.5: *Distributions of the relative root mean square errors of Prevalence- Recency- and Incidence- over the canonical permutation of survey dates and ages- shown separately for each choice of polynomial order of regression formula (columns 1—4) and choice of data inclusion distance in the age direction (row label).* These estimates are based on the canonical scenario and surveys simulated throughout this section- comprising 30 integer ages and 7 surveys- for a total of 210 estimates to construct each of the distributions.

Even at the generous sample size used here (4000 individuals per 5 year age range) the standard error of incidence estimates consistently exceeds the bias. While this is true for both the underlying estimates of prevalence and recency, it is well known that the standard error in the prevalence of recency is the most important source of the standard error in incidence estimates.

Comparing different choices for inclusion radius and polynomial order of fitting function, only the linear fit shows a stark deterioration as the inclusion radius becomes 'too large'- a manifestation of the bias we saw in Figure 5.3. There is little to choose between the third and fourth order fitting procedures, but we noted, by looking at estimated standard errors across repeats of surveys, that the standard errors are more tightly clustered for the cubic fitting. We henceforth use, unless specified otherwise, a cubic polynomial order and an inclusion window of 10 years around the age of interest.

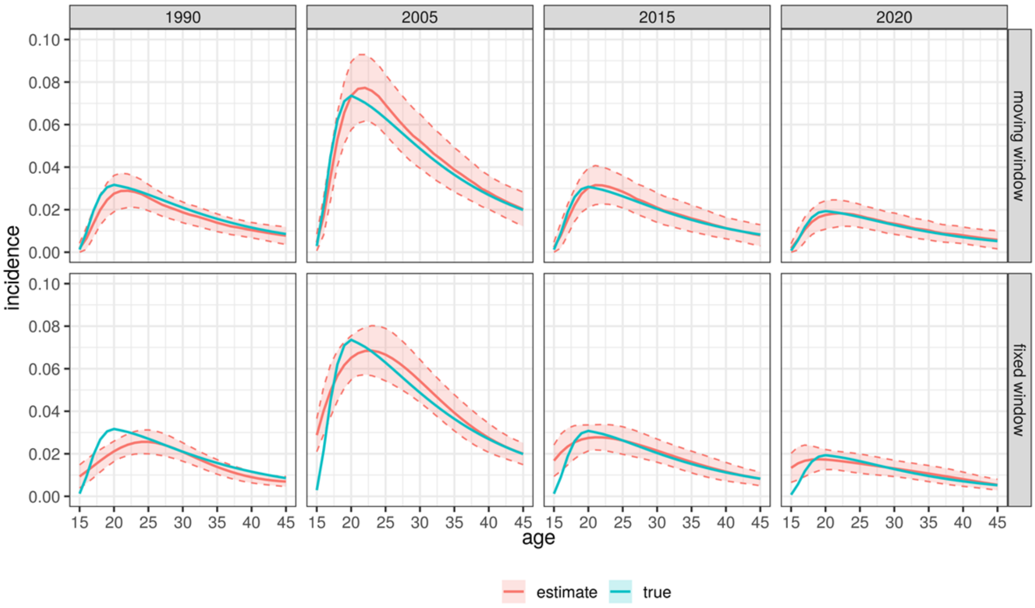

Figure 5.6: *Comparison of the incidence estimates to the true incidence at selected epidemic stages.* Shows the comparison of incidence estimates to the true incidence at epidemic stages simulated in 1990- 2000- 2005- 2015- and 2020 (moving window vs fixed window).

In Figure 5.6 we see a comparison of this 'moving data inclusion window' (cubic with 10 years on either side of age of interest) with a fit performed over the entire range of the age data (15-45), for 4 epidemiological stages. The true incidence is also shown. The slight decrease in standard error achieved by using all the data in a single fit seems to be more than offset by the appearance of significant bias.

Table 5.1: *Age specific incidence estimates (in % p.a.) derived from the regression versus naive approach.*

Age	True	Regression Estimate	Naive Estimate	S.E.$\frac{Naive}{Regression}$
20.5	2.54	2.64 (1.70−3.59)	2.63 (1.61−3.66)	1.08
21.5	2.47	2.55 (1.97−3.13)	2.56 (1.53−3.59)	1.76
22.5	2.40	2.48 (1.87−3.10)	2.49 (1.46−3.52)	1.68
23.5	2.32	2.40 (1.82−2.98)	2.41 (1.36−3.46)	1.80
24.5	2.23	2.32 (1.80−2.83)	2.32 (1.28−3.37)	2.04
25.5	2.13	2.23 (1.70−2.75)	2.22 (1.19−3.24)	1.95
26.5	2.03	2.13 (1.55−2.71)	2.13 (1.09−3.17)	1.80
27.5	1.93	2.02 (1.46−2.59)	2.02 (0.97−3.07)	1.85
28.5	1.84	1.91 (1.34−2.47)	1.91 (0.89−2.93)	1.82
29.5	1.74	1.82 (0.89−2.74)	1.82 (0.77−2.86)	1.13

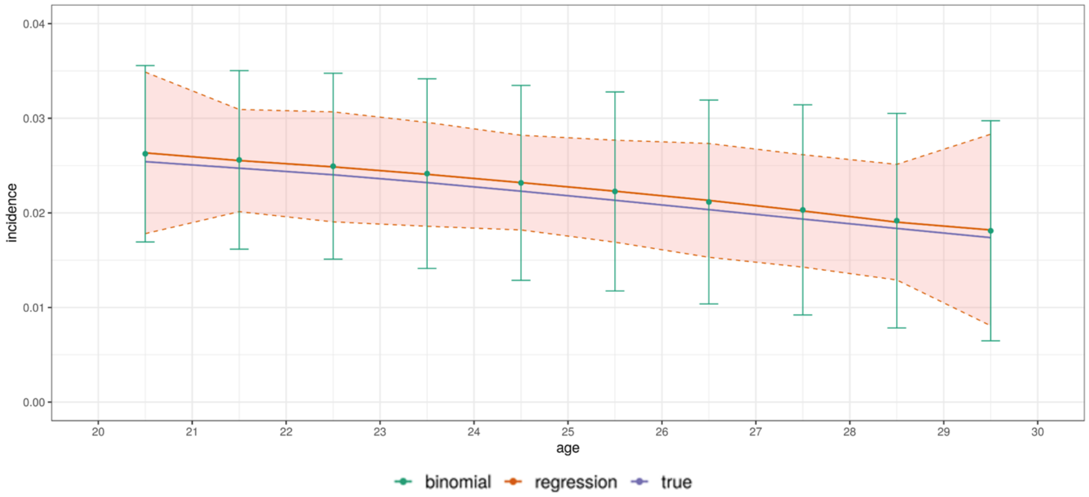

Figure 5.7: *Representative incidence estimate by age from focused survey of 24000 individuals in the age range 20-30.*

Table 5.1 presents a comparison of age specific incidence estimates derived from our proposed regression with the naive approach of estimating P and R for each age by simply using the observed prevalences in a one-year age bin. The mean and ranges summarise the point estimates from 10,000 repeats of the entire survey. Expectation values of the point estimates from these two approaches show precisely the same negligible bias, but as expected, the standard errors are substantially smaller for the estimates derived from the regression approach. Figure 5.7, is a graphical representation of Table 5.1, and compares the age specific incidence estimates derived from the binomial exact versus the regression approach for ages 20.5 to 29.5.

5.4.3 Post Hoc Age Averaging

Integer age specific incidence estimates for similar ages are fairly correlated as they are based on very similar data sets, but they are derived from age-specific customisation and hence each contain some different information. Hence, the question arises whether some averaging over these incidence estimates might provide a reduction in statistical error. To explore this, we performed variable window averaging of the integer age specific estimates obtained by our canonical regression. Figure 5.8 shows the standard errors of such estimates as a function of the averaging window, for a combination of ages and times, using our canonical scenario and survey times.

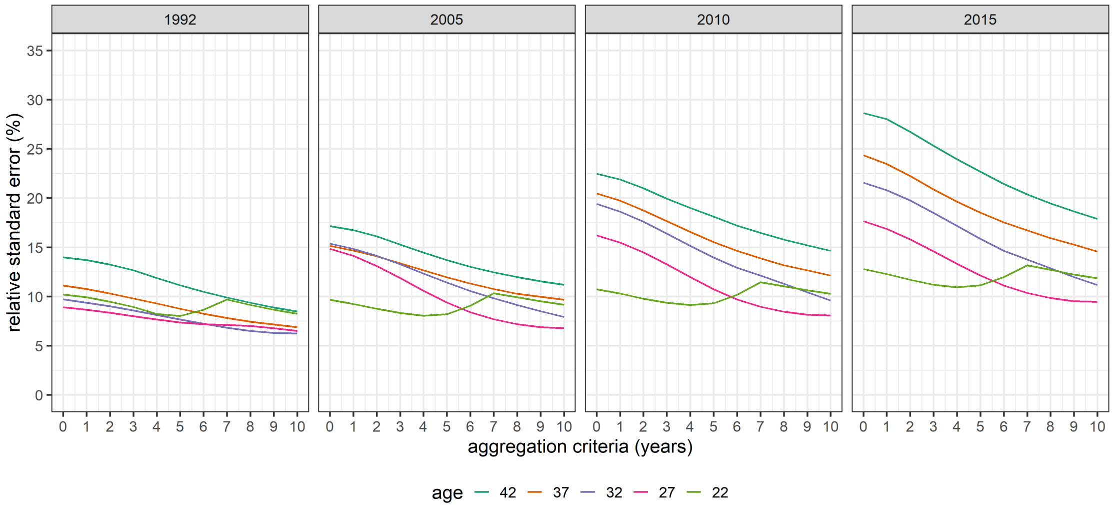

Figure 5.8: *Incidence's relative standard error as a function of the binning strategy for simulated times 1995- 2005- 2010- 2015- and 2020 and ages 22- 27- 32- 37- and 42.*

5.4.4 Incidence Trends

Increasingly, major population-based surveys with recency ascertainment, such as the PHIA surveys, are being performed in multiple rounds. Naturally, one would like to use such data sets to estimate changes in incidence over time.

Various attempts will inevitably be made to estimate mean or midpoint incidence between major surveys, based on such ideas as 'synthetic cohort' analysis [2, 3]. These methods do not necessarily require, or have any role for, recency data. In the present discussion, it makes sense to ask how such midpoint estimates would be obtained via the Kassanjee analysis, and how accurate and precise they are expected to be. We used pairs of surveys 5 years apart, from our canonical set of surveys, and performed a simultaneous age and time regression using all the relevant powers of age and time (including cross terms) consistent with the default cubic form chosen earlier. Given that there are only two time points in each regression, terms with higher order than linear in time are pointless and a sufficiently robust regression algorithm will detect this. Figure 5.11 shows the estimates obtained when fitting with polynomial order 3, with a moving window of plus minus ten years around each age at which incidence is being estimated. Note the almost absent bias and pleasing standard errors.

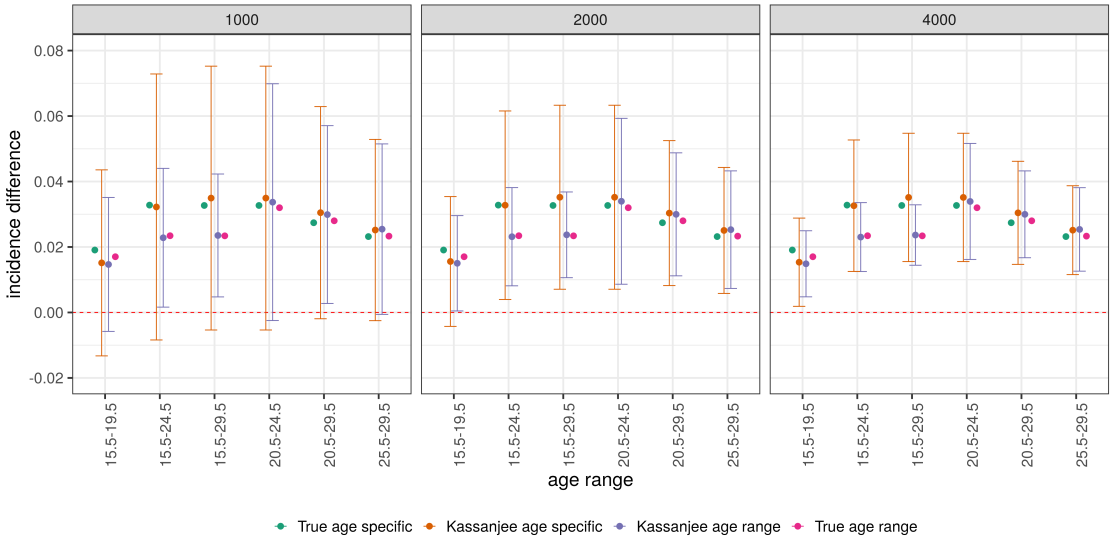

Figure 5.9: *Comparison of age specific and age range incidence differences.* Estimated from 2 cross sectional surveys five years apart (1993- and 1998) and simulated when incidence was increasing- the sample size is 1000- 2000- and 4000 per 5-year age range (total sample sizes are 8000- 12000- and 24000 respectively).

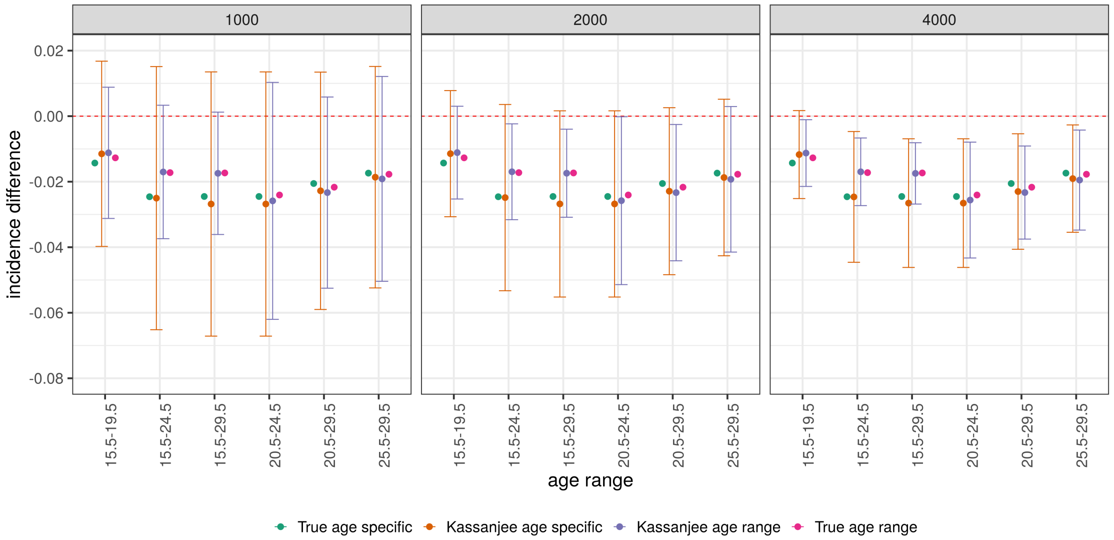

Figure 5.10: *Comparison of age specific and age range incidence differences.* Estimated from 2 cross sectional surveys five years apart (2010 and 2015) and simulated when incidence was steadily decreasing- with sample size is 1000- 2000- and 5000 per 5-year age range (total sample sizes are 8000- 12000- and 24000 respectively

5.4.5 Mean Incidence Between Survey Rounds:

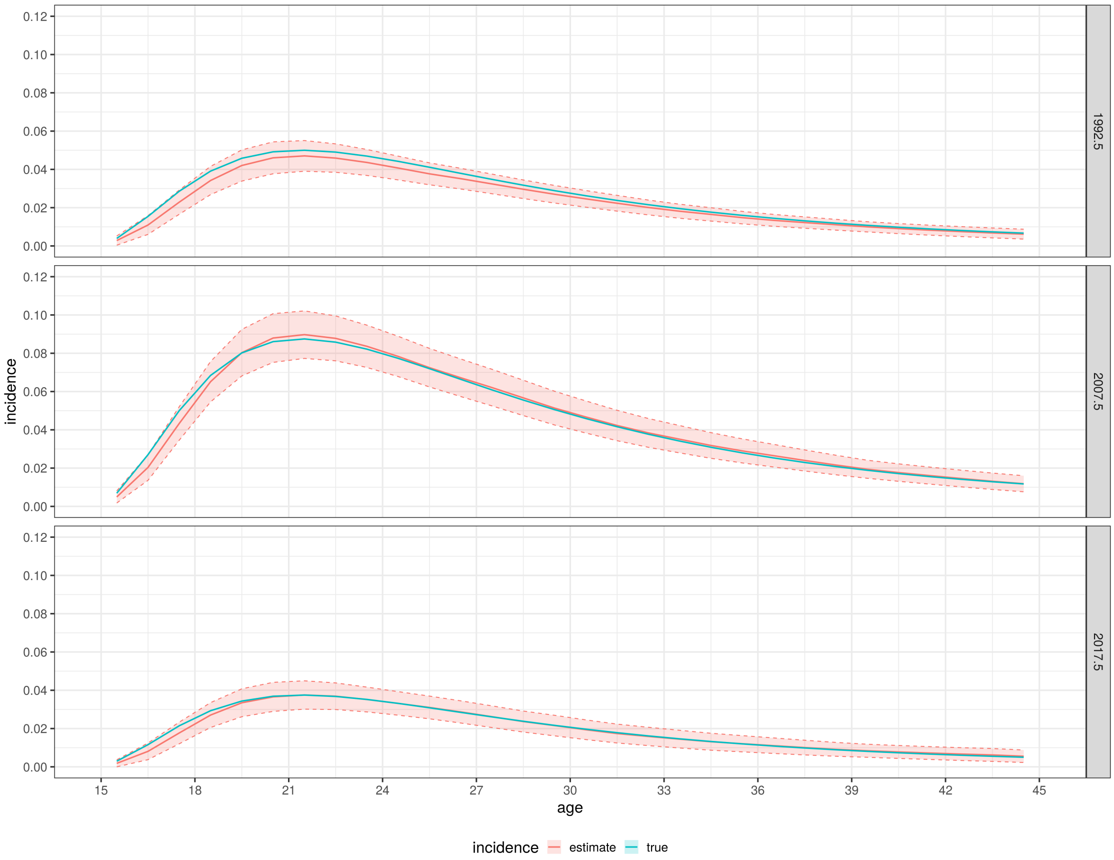

Figure 5.11: *Midpoint incidence estimates from two time points in our canonical South-Africa-like epidemiological scenario- alongside the true incidence at the corresponding time.*

5.5 Conclusion

Using some laboratory procedure to define 'recent infection' amongst HIV positive survey respondents is a widely practiced approach to generating population level incidence estimates without the need to do individual follow up or wholesale repeat of major surveys. A useful conceptual framework for

- Defining recent infection testing,
- Defining recency test performance characteristics, and

- How to combine these with survey-based 'prevalence' estimates into an incidence estimator with well described analytical inputs and computable variance

was fundamentally outlined by Kassanjee et al in 2012 [4] but this exposition did not address important details around managing age structure in survey data, which is known to be very important in the case of generalised HIV epidemics.
In the present work, we have shown

- That while ignoring age structure is technically valid, the age averaging implied by such an analysis hides important details that are of high interest epidemiologically
- How to select generically stable regression models which ultimately lead to robust age-specific incidence estimates that are close to optimal, given the information content of data sets such as are routinely generated by large population-based surveys which test for 'recent' HIV infection.

Specifically, we propose the following one-size-fits-most approach to implementing the Kassanjee estimator:

- Prevalence (of HIV and of recency) data can be generically fitted by a polynomial in age (and, where applicable, time) truncated at third or fourth order.
- It makes sense to estimate incidence separately for each integer age, by performing a fit of data sufficiently close to the age/time point of interest, in a 'moving window' data inclusion rule. We recommend, by default, inclusion of data from all ages no further than 10 years from the age of interest.
- To improve precision, age specific estimates can be aggregated into age range averages using a contextually appropriate range of ages.
- Statistical uncertainty is most reliably computed by bootstrapping the data in accordance with the sampling strategy, to generate realistic uncertainty in, and covariance among, the prevalence estimates.

Depending on how much data is available - by which we mean both the sampling intensity per survey, and the number of discrete survey rounds (typically separated by more than just one or two years), the following can be considered to be the primary fruitful applications of the Kassanjee incidence estimator:

- Estimating incidence from a single cross-sectional survey

- Estimating incidence changes from two cross sectional surveys conducted some years apart
- Estimating incidence differences between locales surveyed separately
- Estimating point (or mean) incidence at (or over) a time between two surveys

Much of our crucial **R** code for analysis is partly already freely available in the **R** package Inctools [60] available on CRAN (the standard community platform), and additional code is will find its way into later releases of inctools. The simulation code can be transferred under bilateral agreements until it is formally released in a separate **R** package. It will not be burdensome for analysts who are familiar with **R** to replicate our analyses, and adapt them to their specific needs in order to confirm or tailor our proposed algorithms from case to case.

In two companion papers to the present one, we further investigate:

1. Similar prevalence smoothing criteria [69]- in particular optimised for estimating the gradient of prevalence such as is needed for a robust ‘synthetic cohort’ type incidence estimate in the sense of Mahiane et al [3].
2. The optimal use of both the Kassanjee and Mahiane analyses on data sets to which both are applicable [70].

5.A Appendix: Epidemiological Rates in the Simulation Platform

We simulated a population starting in 1945, in order to have persons of all relevant ages when we start surveys in 1990. Incidence and mortality were chosen to yield a scenario superficially similar to the generalized HIV epidemic seen since then in South Africa

Fertility

Most of the calculations are not affected in any way by the fertility parameters of the simulation, since sampling is performed as if from an infinite population. Except where explicitly noted, we used an arbitrary, meaningless, constant birth rate. This results in some age structure due to mortality, and hence (minor) differences between various age weighted incidence averages:

- uniformly weighted,
- population age distribution weighted,
- susceptible population weighted

Incidence

The HIV incidence is dependent on age and time through a function which is the product of

1. a lognormal term that is only a function of age, with incidence being zero until age 14 (no mother to child transmission) and peaking at age 20 (see figure A1), and
2. a lognormal term that is only a function of time, becoming non zero from 1986, and peaking in 2000 (see Figure 5.12)

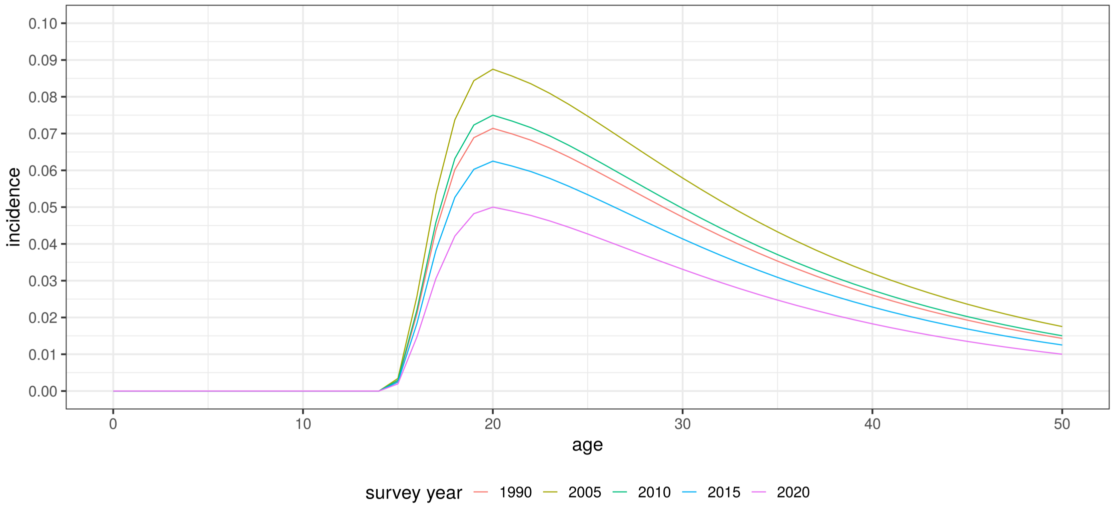

Figure 5.12: Incidence as a function of age $I(a)$ at selected times considered in the investigations

Background Mortality

Mortality among uninfected individuals is 1 percent per annum at birth, and climbs linearly with age, reaching 3 percent per annum at age 50. We do not survey the population over age 50.

Infection-associated Mortality

Upon infection, individuals experience an age-at-infection and time-since-infection dependent excess mortality which is a calendar time independent power (2.28) of time-since-infection. The prevalence which emerges from the interplay of incidence and mortality is summarized in Figure 5.13

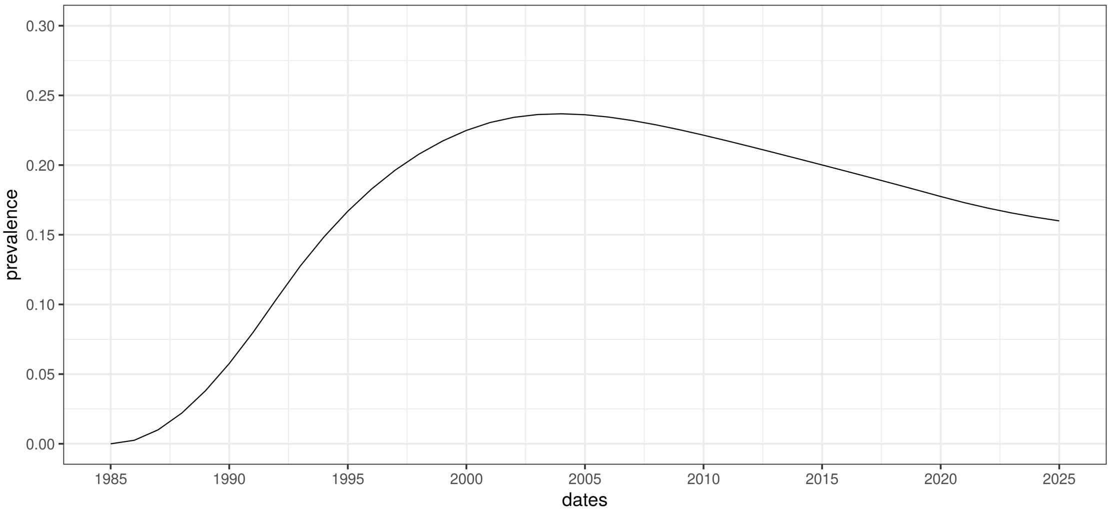

Figure 5.13: Age weighted prevalence output as a function of time $P(t)$ for (ages 15-45).

'Recent Infection'

After infection, individuals tested for 'recent infection' have a probability of giving the result 'recent' according to a Weibull survival curve with scale factor 0.5 (years) and shape parameter 5. This leads to a mean duration of recent infection (MDRI) of 167.7 days and a negligible false recent rate (FRR). This simplifies all our analysis by freeing us from the real-world problem of estimating the FRR. The interplay of all of the above mentioned parameters leads to a 'prevalence' of recent infection as shown in Figure 5.14. Note that this prevalence is only defined among HIV positives, not over the entire population.

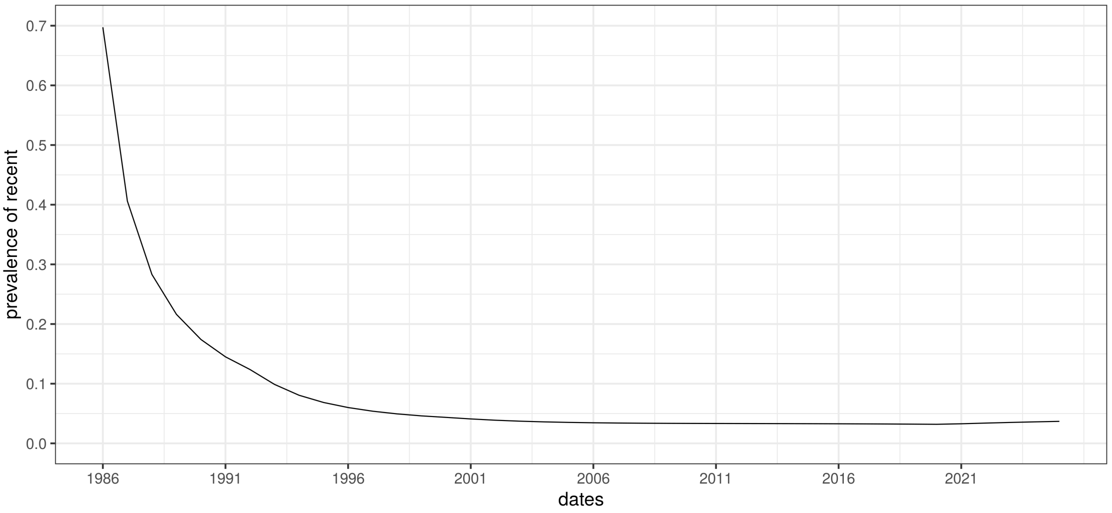

Figure 5.14: Age weighted prevalence of recent infection output as a function of time $R(t)$.

5.B Appendix: Further Error Analysis

Figure 5.5 in the body of the manuscript shows distributions of relative root mean square errors for prevalence, recency and incidence for various choices of survey data fitting parameters (polynomial order and data inclusion age range around age of interest). The following three plots present the underlying distributions of the relative standard errors (red), and relative bias (teal), of, respectively, prevalence (Figure 5.16), prevalence of recent infection (Figure 5.15), and incidence (Figure 5.17). The plots are based on a range of simulated epidemic stages (the standard times at which the cross-sectional surveys were simulated) and ages in the range 15 to 45. At each of these times and ages, we varied the polynomial order of the regression formula, as well as the data inclusion distance.

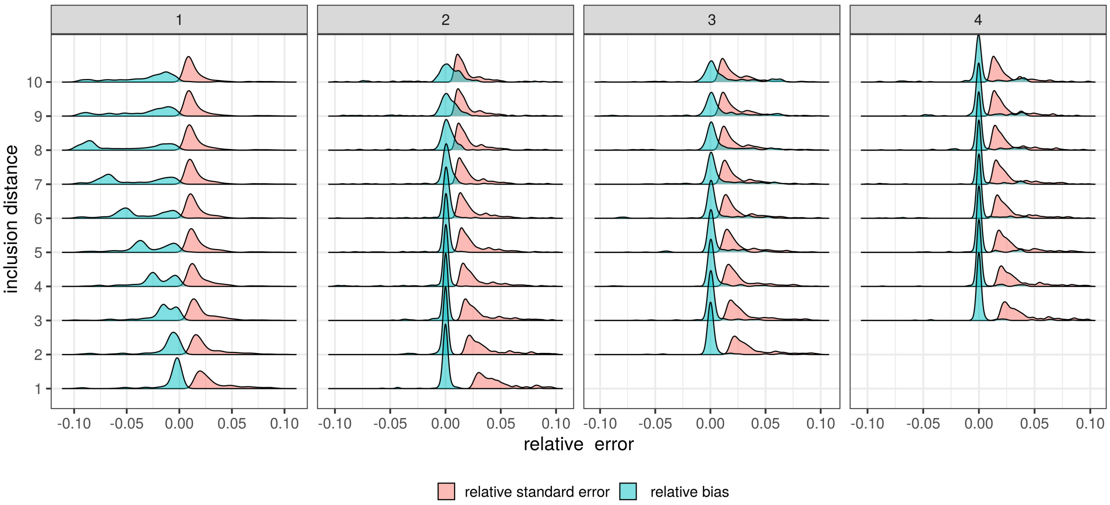

Figure 5.15: Relative standard error and relative bias of P (prevalence) estimates using the logit link function. Each facet represents the 4 polynomial orders (1-4) compared to each other and the x-axis represents the inclusion distance, sample $= 4000/5$ year age range.

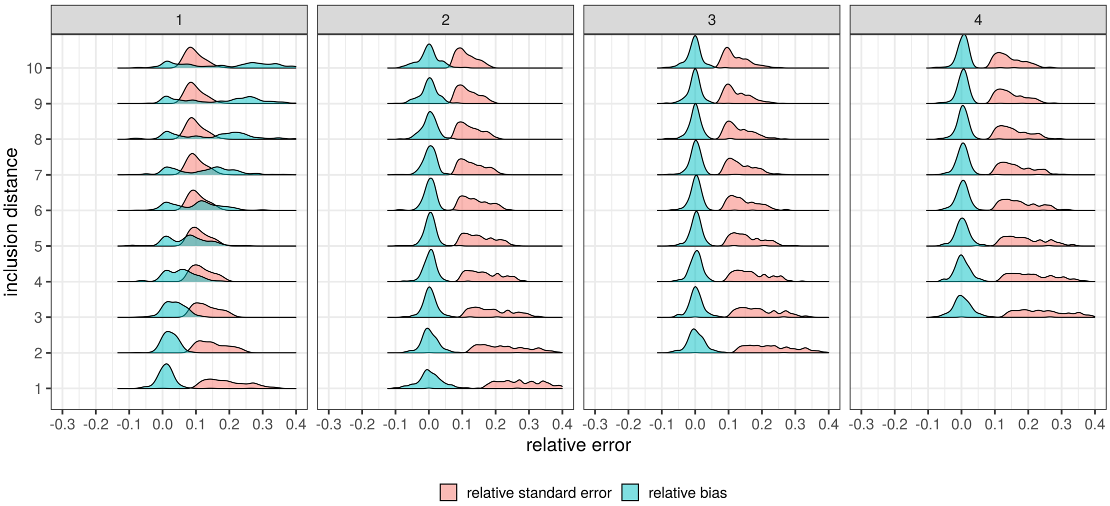

Figure 5.16: *Relative standard error and relative bias of R (recency) estimates using the clog log link function.* Each facet represents the 4 polynomial orders being compared and the inclusion distance is on the x-axis. The sample size was set to 4000 per 5 year age ranges

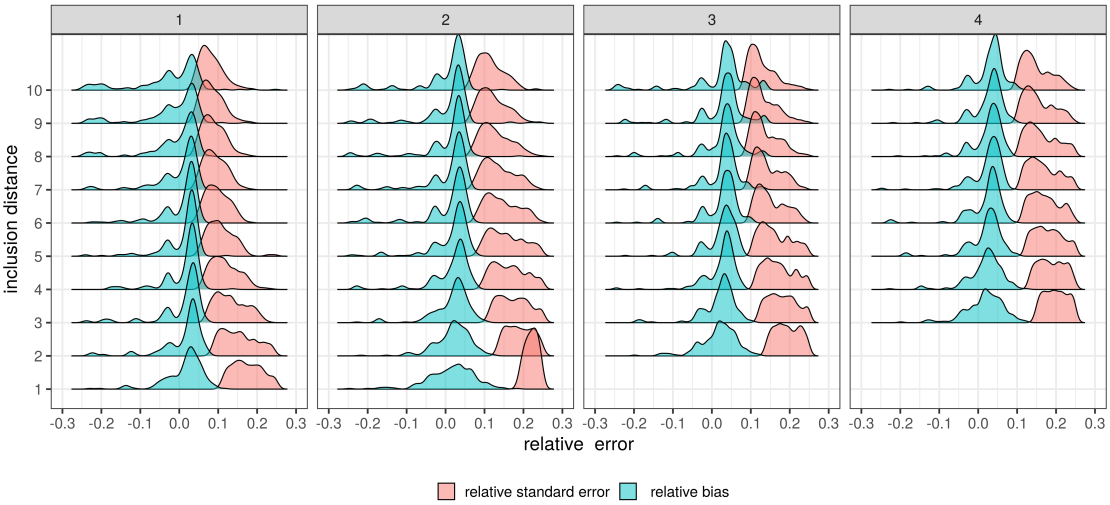

Figure 5.17: Relative standard error and relative bias for **incidence estimates** disaggregated by polynomial order and the inclusion distance on the x-axis.

Incidence Difference

For the purpose of evaluating the ability to detect incidence differences, we simulated 2 pairs of cross-sectional surveys; the first pair of cross-sectional surveys depicted an epidemic where incidence was rapidly increasing (1993 and 1998 in our canonical scenario); the second pair of cross-sectional surveys portrays an epidemic that is steadily decreasing (2010 and 2015 in our canonical scenario). For each cross-sectional survey, we independently estimated the age specific incidence and for the survey pairs (set 5 years apart) we estimated the incidence differences. For each analysis we varied the sample size to highlight the effect of sample size on yielding informative incidence difference estimates. Figures 5.9 and 5.10 in the body indicate differences in various proposed incidence age-range-averages, and Figures 5.18 and 5.19 show the underlying detailed integer-age specific estimates.

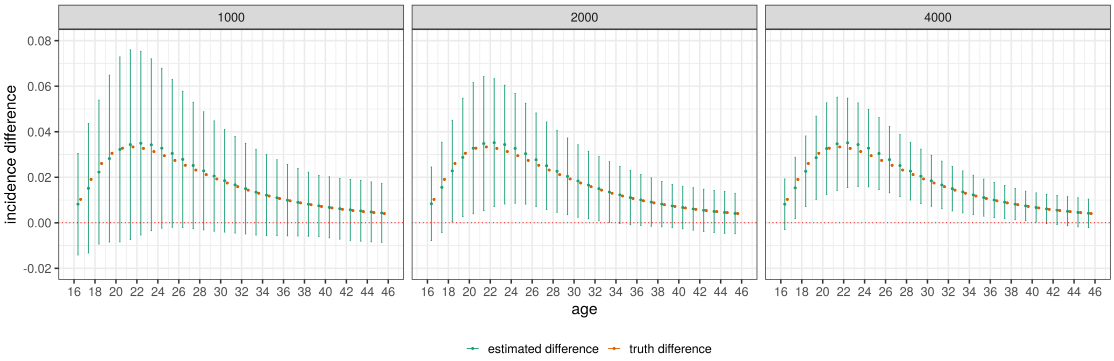

Figure 5.18: Incidence difference estimates calculated from two cross sectional surveys simulated in 1993 and 1998 with size of 1000, 2000, and 4000 per 5-year age bin.

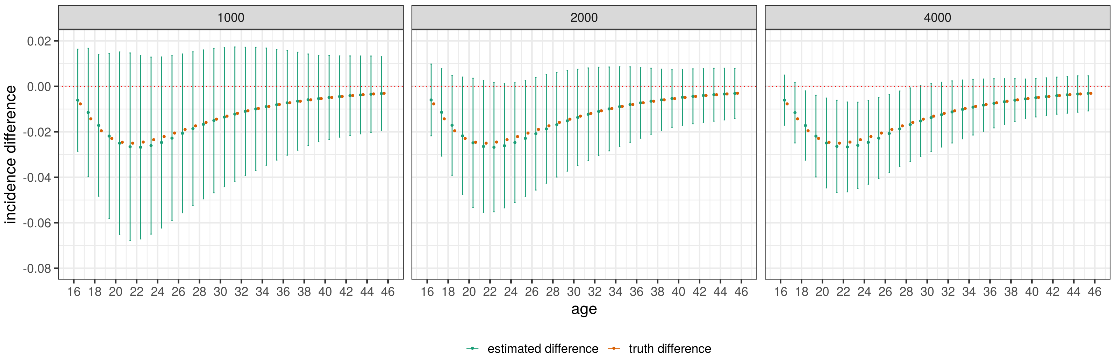

Figure 5.19: Incidence difference estimates calculated from two cross sectional surveys simulated in 2010 and 2015 each with sample size 1000, 2000, 4000 per 5-year age bin.

Chapter 6

Smoothing age/time structure of HIV prevalence, for optimal use in synthetic cohort based incidence estimation

6.1 Abstract

Background

Population-based surveys which ascertain HIV status are conducted in heavily affected countries, with the estimation of incidence being a primary goal. Numerous methods exist under the umbrella of 'synthetic cohort analysis', by which we mean estimating incidence from the age/time structure of prevalence (given knowledge on mortality). However, not enough attention has been given to how serostatus data is 'smoothed' into a time/age-dependent prevalence, so as to optimise the estimation of incidence.

Methods

To support this and other related investigations, we developed a comprehensive simulation environment in which we simulate age/time structured SI type epidemics and surveys. Scenarios are flexibly defined by demographic rates (fertility, incidence and mor-

This chapter is available as a preprint: authors: Mhlanga, L. Grebe, E. Welte, A., Title: Smoothing age/time structure of HIV prevalence, for optimal use in synthetic cohort based incidence estimation, DOI: 10.21203/rs.3.rs-959136/v2

tality - dependent, as appropriate, on age, time, and time-since-infection) without any reference to underlying causative processes/parameters. Primarily using 1) a simulated epidemiological scenario inspired by what is seen in the hyper-endemic HIV affected regions, and 2) pairs of cross-sectional surveys, we explored A) options for extracting the age/time structure of prevalence so as to optimise the use of the formal incidence estimation framework of Mahiane et al, and B) aspects of survey design such as the interaction of epidemic details, sample-size/sampling-density and inter-survey interval.

Results

Much as in our companion piece which crucially investigated the use of 'recent infection' (whereas the present analysis hinges fundamentally on the estimation of the prevalence gradient) we propose a 'one size fits most' process for conducting 'synthetic cohort' analyses of large population survey data sets, for HIV incidence estimation: fitting a generalised linear model for prevalence, separately for each age/time point where an incidence estimate is desired, using a 'moving window' data inclusion rule. Overall, even in very high incidence settings, sampling density requirements are onerous.

Conclusion

The general default approach we propose for fitting HIV prevalence to data as a function of age and time appears to be broadly stable over various epidemiological stages. Particular scenarios of interest, and the applicable options for survey design and analysis, can readily be more closely investigated using our approach. We note that it is often unrealistic to expect even large household based surveys to provide meaningful incidence estimates outside of priority groups like young women, where incidence is often particularly high.

6.2 Introduction

Population-level cross-sectional surveys, including HIV status determination, are conducted routinely in many Sub-Saharan countries [11, 12, 13, 14]. Within the last two decades, variations of such surveys have been executed multiple times in numerous countries, making it possible to track explicitly the dependence of prevalence on both age and time. Combining estimates of the age/time structure of prevalence with appropriate estimates of mortality facilitates what have often been called 'synthetic cohort' estimates of incidence. Various approaches to 'synthetic cohort' incidence estimate have been proposed [2, 3, 40, 41]. A recurring theme is the use of assumptions, or parameterisations, which effectively capture the idea that some combination of incidence, prevalence and mortality is constrained to be piecewise constant over ranges of age and/or time.

The approach of Mahiane et al. [3] requires no such simplifications, and the core estimator is nothing more than a rewriting of the minimal population renewal equation applicable to an irremissible condition. The inputs required for the estimator are:

1. An estimate of prevalence for the population at an age and time of interest
2. An estimate of the 'gradient' of this prevalence - defined as the rate of change of prevalence experienced by the single-age birth cohort to which the age and time of interest belongs
3. An estimate of the (net/average) 'excess mortality' experienced by the infected population at the age/time of interest. (Viewed in its most general form, we can reinterpret the excess mortality as a 'net excess attrition', which can theoretically be negative if there is substantial migration impacting prevalence)

The first two of these inputs are clearly to be based on particular survey data, and the third will typically have to be based on suitable background studies, sensibly adapted to the applicable context where the survey data has been obtained.

In the present investigation, we do not explore the problem of estimating this excess attrition rate. We focus on the smoothing of serostatus observations of survey respondents, to extract optimal estimates of the prevalence and prevalence gradient. As far as we are aware, there has been no previous investigation of the various trade-offs involved in choosing one or other approach for extracting the age/time structure of the prevalence for these purposes.

In outline, the present work has the following high-level components:

1. Simulating 'realistic' epidemics and cross-sectional surveys
2. Applying various smoothing algorithms to the survey data, in order to infer (age- and time-structured) prevalence of HIV infection.
3. Estimating incidence, and incidence differences/trends, from these smoothed functions, using the Mahiane [3] framework

4. Evaluating the relative merits of various smoothing and averaging schemes, by comparing estimated with the known incidence parameter values in the simulations.

5. Proposing a generic one-fits-most approach to the main use-cases of incidence estimation

The possible availability of 'recent infection' ascertainment is not considered here, but features in detail in two companion pieces, which together with this one, explore a closely related set of variations on the theme of smoothing survey data to optimally extract the age/time structure of prevalence, for the purposes of estimating HIV incidence.

6.3 Methods

Using a customised simulation platform, developed for this and some closely related investigations, and described in detail separately [71], we simulated a South-Africa-like HIV epidemic. The simulation platform generates scenarios defined by epidemiological and demographic rates (incidence, base and excess mortality - see functional forms in Appendix 6.5) and described by an age, time, and time since infection dependent population (density). The canonical simulation is run from 1935 to 2025. Each birth cohort is simulated to age 50.

We simulated cross-sectional surveys in stages (1992, 1997), (1995, 2000), (1998, 2003), (2000, 2005), (2005, 2010), (2010, 2015), and (2015, 2020). Sampling density was varied from 1000 to 4000 persons per 5 year age bracket. Incidence estimation is based on the estimator of Mahiane et al. [3]

$$I_M = \frac{1}{1-P} \cdot \left(\frac{\partial P}{\partial t} + \frac{\partial P}{\partial a} \right) + M \cdot P \tag{6.3.1}$$

Or,

$$I_M = \frac{1}{1-P} \cdot \frac{dP}{dt} + M \cdot P \tag{6.3.2}$$

where P is the prevalence of HIV, M is the differential mortality of the HIV infected population, and the derivative of prevalence captures the rate of increase of prevalence as seen from the point of view of a birth cohort which has reached the age of interest, at the time of interest. In terms of a traditional delta method expansion for statistical error:

$$\begin{aligned}
\mathrm{var}(I_M) = & \left[\frac{1}{(1-P)^2}\cdot\frac{dP}{dt}+M\right]^2\cdot\sigma_P^2+\left[\frac{1}{1-P}\right]^2\cdot\sigma_{\frac{dP}{dt}}^2+[P]^2\cdot\sigma_M^2+ \\
& 2\cdot\left[\frac{1}{(1-P)^2}\cdot\frac{dP}{dt}+M\right]\cdot\left[\frac{1}{1-P}\right]\cdot\sigma_{P,\frac{dP}{dt}}+ \\
& 2\cdot\left[\frac{1}{(1-P)^2}\cdot\frac{dP}{dt}+M\right]\cdot P\cdot\sigma_{P,M}+ \\
& 2\cdot\left[\frac{1}{1-P}\right]\cdot P\cdot\sigma_{\frac{dP}{dt},M}
\end{aligned} \tag{6.3.3}$$

Where σ_P, $\sigma_{\frac{dP}{dt}}$, and σ_M are the standard errors of prevalence, gradient of the prevalence, and excess mortality respectively, and the σ's with double subscripts are the indicated covariances and therefore the incidence's standard error is given by $se(\lambda)=\sqrt{\mathrm{var}(I_M)}$.

Our investigation is very similar in inspiration to that reported in our companion article [71], as it aims to find robust ways to perform regression of survey based HIV status observations, to derive prevalence as a function of time in a manner that can be substituted into an incidence estimator. The key difference is that in our prior work, we reported on an estimator (according to Kassanjee et al. [4]) which does not rely on an estimate of the gradient of prevalence, using instead, crucially, data on ascertainment of 'recent' versus 'non-recent' infection. In the present case, as far as processing of survey data is concerned, incidence estimation hinges crucially on the estimates of the gradient of prevalence.

Additionally, the Mahiane estimator requires an estimate of 'excess mortality' associated with infection. We defer discussion of how best to estimate this, but note:

1. Data from household surveys is generally not an appropriate source of mortality estimates.

2. Hence, appropriate estimation of the contextually applicable excess mortality is in practice an open ended problem.

3. In our simulated estimation challenges, we explicitly calculate the age and time specific excess mortality, at any required values of age and time, by averaging the differential mortality over all extant values of 'time since infection' which are manifested in the population, and we use this exact excess mortality in our estimates.

4. Hence, our estimates indicate the most optimistic application of the Mahiane estimator which is conceivable under the circumstances defined by the survey design.

The Appendix 6.5 provides additional details on how uncertainty in fitting parameters is propagated into uncertainty of incidence estimates, given the potentially non linear relation between these parameters and the prevalence and prevalence gradient which are required in the Mahiane [3] estimator.

Our approach to smoothing prevalence data in this work is essentially the same as in our companion piece focusing on recency data - namely to have a separate raw-data-to-estimate process for each value of age and time for which an incidence estimate is to be obtained. For any particular choice of age and time, then, we identify the data 'sufficiently close' to the age/time of interest - usually defined by all observations within a specified range of ages - and then fit a generalised linear model of some polynomial in age and time, using a logit link function by default.

Proceeding much as in our previous analysis of how to smooth survey data in prevalence and prevalence of recency, we proceed to consider also the estimated gradient of prevalence, by considering various permutations of the polynomial order of the fitting function and data inclusion algorithms. The use of a logit link function ensures stability of prevalence between 0 and 1, whereas an identity link function sometimes leads to fitting instability. We use the logit link throughout the present work, but note that other link functions may be perfectly stable in various real-data applications where it is not necessary to automate the production of a large number of variations on an analytical theme.

6.4 Results

6.4.1 Data Inclusion Distance and Polynomial Order of Fitting Function

Within the general approach of a 'moving window' data inclusion rule, fitting a generalised linear model (polynomial in age and time) to HIV status data is expected to show the following trade-offs:

- Increasing the data inclusion rule should increase precision at the cost of some bias
- Increasing the polynomial order should decrease bias at the cost of precision

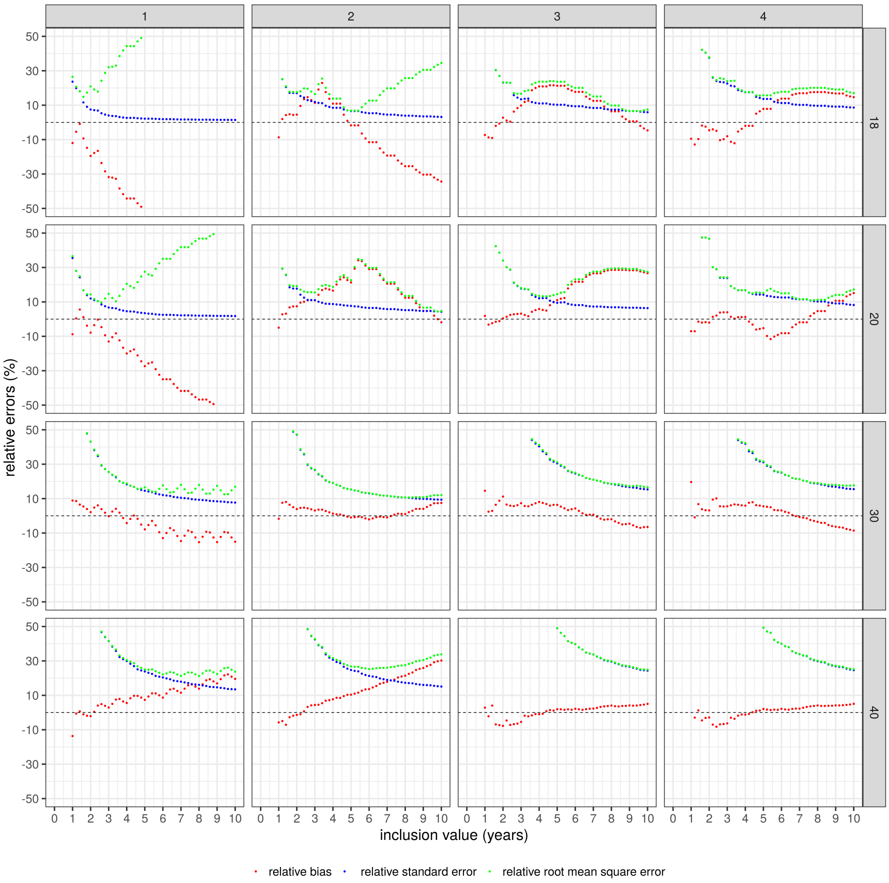

Figure 6.1: *Relative errors, as a function of the inclusion distance, for various ages and choices of the polynomial order of the prevalence fitting function.* The plots relative bias (green), relative standard error (red), and relative root mean square error (blue) for estimates of incidence at the indicated ages, in mid-2017 of the canonical scenario, based on surveys conducted at the beginning of 2015 and 2020 with a sampling density of 4000 individuals per 5 year age range.

Figure 6.1 shows the interaction of these trade-offs at a range of ages (a_0 = 18, 20, 30, and 40) and a single time (t_0 = 2017.5) using simulated data from 2015 and 2020. The curves indicate (percentage) relative errors (standard error (red), bias (green) and root mean square error (blue)) as functions of r (inclusion distance) shown separately for each polynomial order (linear, quadratic, cubic, or quartic in age, always terminated at linear in time as there are only two time points, and allowing all the arising cross terms). It appears that, for these cases:

- At least cubic terms are needed to avoid substantial bias
- Inclusion distances should be at least 5 years
- The younger ages are more problematic.

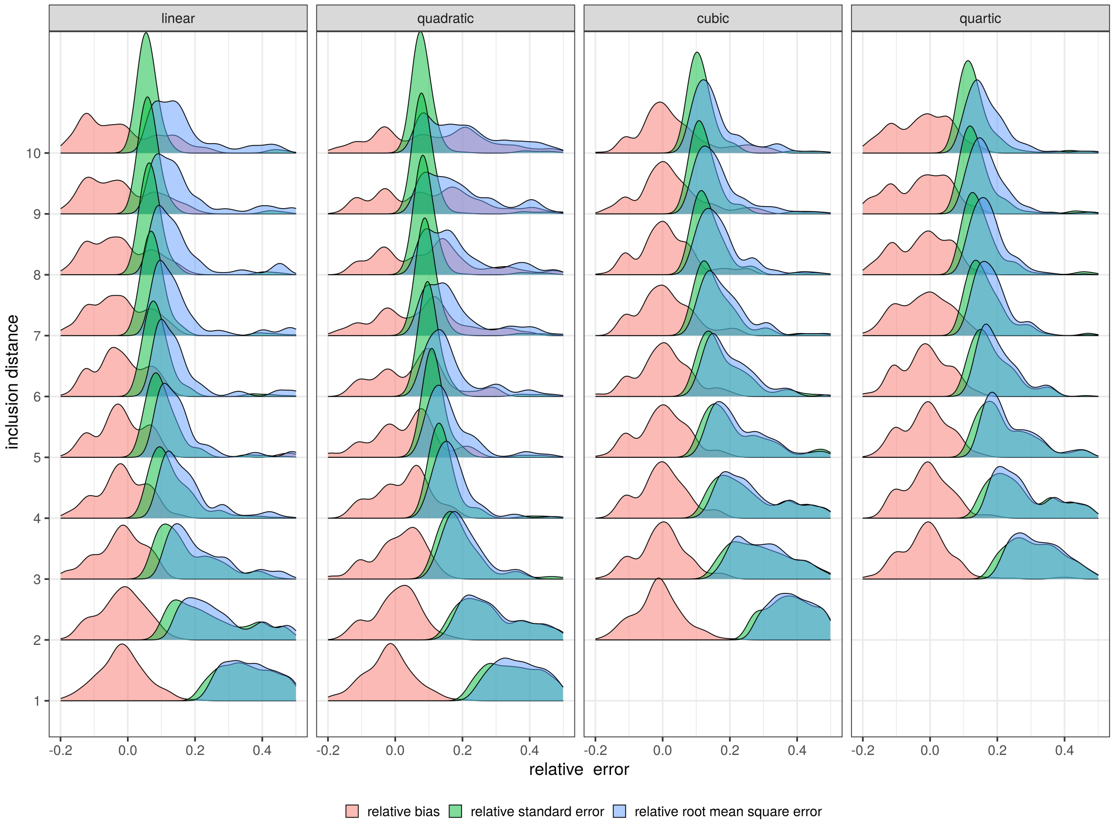

Figure 6.2: *Distributions of relative errors in incidence estimates, arising over the range of integer ages and inter-survey midpoint times in the standard canonical epidemiological scenario.* The plots show the relative bias, relative standard error and relative root mean square error. The inter survey intervals are each 5 years, and the sampling density is 4000/5yr age bin. Each facet shows the distribution of the relative errors generated for the indicated combination of polynomial order of the prevalence fitting function, and the data inclusion distance.

Rather than considering many more individual combinations of age, time, polynomial order and inclusion distance, we show, in Figure 6.2, the distribution of errors arising over various combinations of age (15-45) and times (1994.5, 1997.5, 2000.5, 2002.5, 2007.5, 2010.5, 2012.5, and 2017.5 - in each case based on a pair of surveys five years apart with the relevant time as the midpoint) in our canonical scenario. Each single density plot depicts the distribution of the relative error under consideration (relative bias/relative standard error/relative root mean square error) for the indicated choice of polynomial order and inclusion distance.

The cubic and quartic distributions show significantly smaller tales, indicating fewer 'poor' estimates, and as in Figure 6.1, it seems best to consider data inclusion windows of at least plus/minus 5 years from the age of interest. From now, by default, we will use a polynomial order of 3 and a data inclusion distance of plus/minus 6 years from the age of interest

6.4.2 Effect of Sample Size

For an inter-survey interval of 5 years, we investigated the effects of sample size on the overall errors and present the results as distributions over age/time points in Figure 6.3.

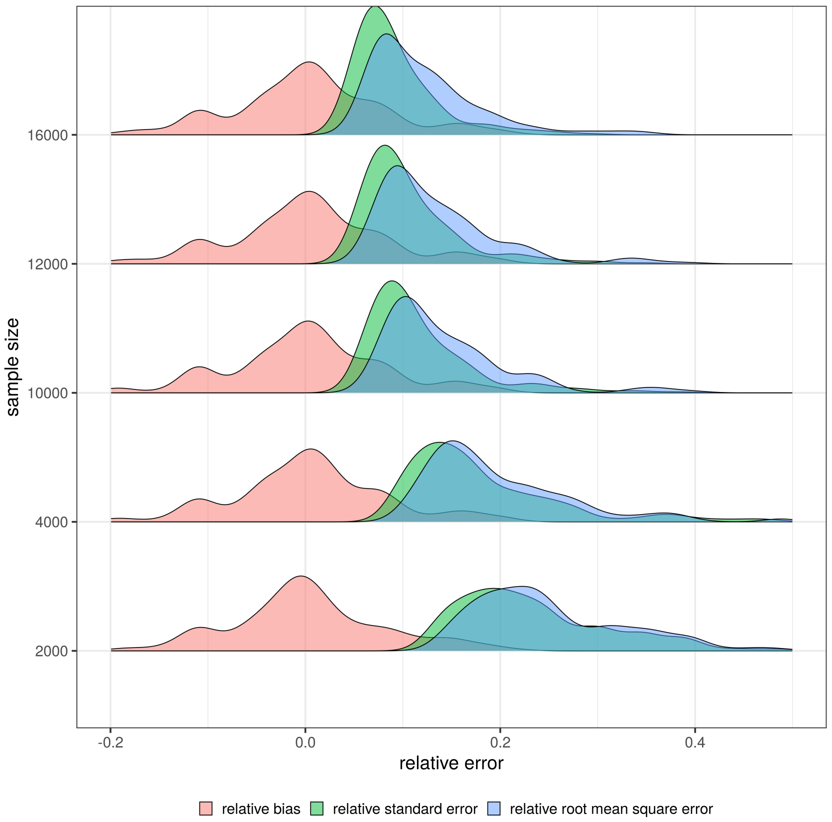

Figure 6.3: *Density plots of the overall errors for varying sample sizes.* The effect of sample size on the incidence estimates is summarised by the distribution of the relative errors. The sample sizes varied are 2000, 4000, 10000, 12000, and 16000 per 5 year survey bin and the inter survey interval is 5. The distributions are based on 248 data points (8 midpoints for ages 15:45).

As expected, the sample size only has an effect on the standard error, not the bias. Even at simulated sampling densities well beyond what has ever been seen in the real world (more than 10,000 individuals per 5 year age bin) the net root mean square error is dominated by the standard error rather than bias.

6.4.3 Inter-survey Interval

Figure 6.4 shows relative error density plots for a range of indicated inter-survey intervals, at which incidence estimates are again generated for the canonical 248 combinations of age and time.

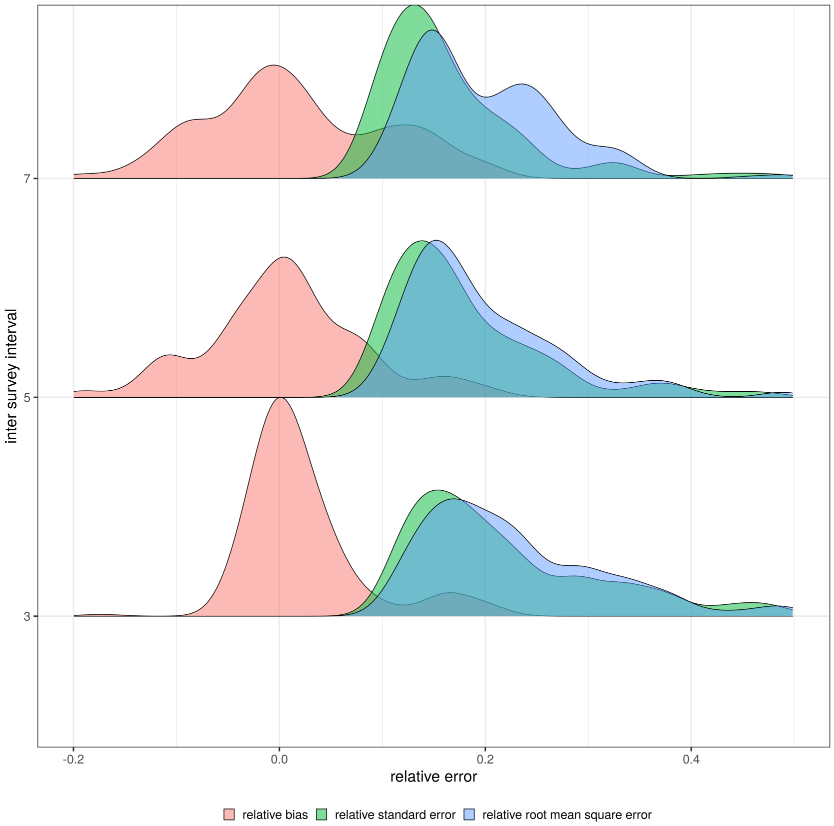

Figure 6.4: *Density plot of the relative errors (bias, standard error, and root mean square error) for 3 different inter-survey intervals (3, 5, and 7).* The Figure shows the distribution of the relative errors for a range of simulated pairs of survey each with a sample of 4000 per 5 year age bin. The surveys are simulated 3, 5, and 7 years apart, but have mid points 1994.5, 1994.5, 2000.5, 2002.5, 2010.5, 2012.5, 2017.5. The plot is based on a total of 210 data points (ages 15 to 45, for each of the 8 time points).

The inter-survey interval of 3 stands out as the one with clean tails on the bias, but a more substantial tail in the distribution of standard errors as the short time between surveys means the prevalence has changed less, and it is hence harder to estimate the prevalence gradient. At an inter-survey interval of 7 years, bias begins to become significant.

6.4.4 Estimating Precision

When analysing real survey data, obtained by substantial investment of money and effort, we would generically propose that statistical error be estimated by bootstrapping the data, replicating sample clustering, stratification, and weighting, as appropriate. When investigating the performance of analysis algorithms on simulated data, one may want to consider many permutations of design features, and be tolerant of such approximations as delta method expansion, which are unlikely to have substantial impact on the evaluation of algorithm optimisation. In fact, as shown in Table 6.1, there is no important difference between the numerically considerably more intensive approach of bootstrapping and the much more computationally compact delta method, which makes it easy to perform a great many simulations very rapidly without requiring more than a single standard PC or laptop. It would even be feasible to implement reliable calculations in browser based applications.

Table 6.1: *Bootstrap (10000) versus delta method* ***standard errors*** *for prevalence, gradient of the prevalence and incidence*. The table shows the prevalence, gradient of the prevalence and the incidence's standard errors, for selected ages 18, 20, 30, and 40 at time 2017.5.

	Prevalence			Prevalence Gradient			Incidence		
Age	Bootstrap	Delta	Ratio	Bootstrap	Delta	Ratio	Bootstrap	Delta	Ratio
18	0.289	0.286	1.01	0.209	0.230	0.900	0.218	0.227	0.96
20	0.400	0.414	0.96	0.231	0.230	1.00	0.248	0.239	1.04
30	0.518	0.514	1.01	0.327	0.330	1.00	0.469	0.467	1.00
40	0.492	0.492	1.00	0.307	0.310	0.989	0.407	0.412	0.98

6.4.5 Two-survey Midpoint Incidence Estimation

It is worth emphasizing that the classic application of the Mahiane estimation procedure is to estimate incidence at the mid-time between two cross sectional surveys conducted a few years apart. We focus, for the present analysis, on this application, and describe the pros and cons of various methodological details in this context. Other application scenarios will be considered in our third (and final) article in this series, where we explore the relative utility of adding recency ascertainment to surveillance scenarios where it might be hoped that the Mahiane analysis provides useful estimates by itself.

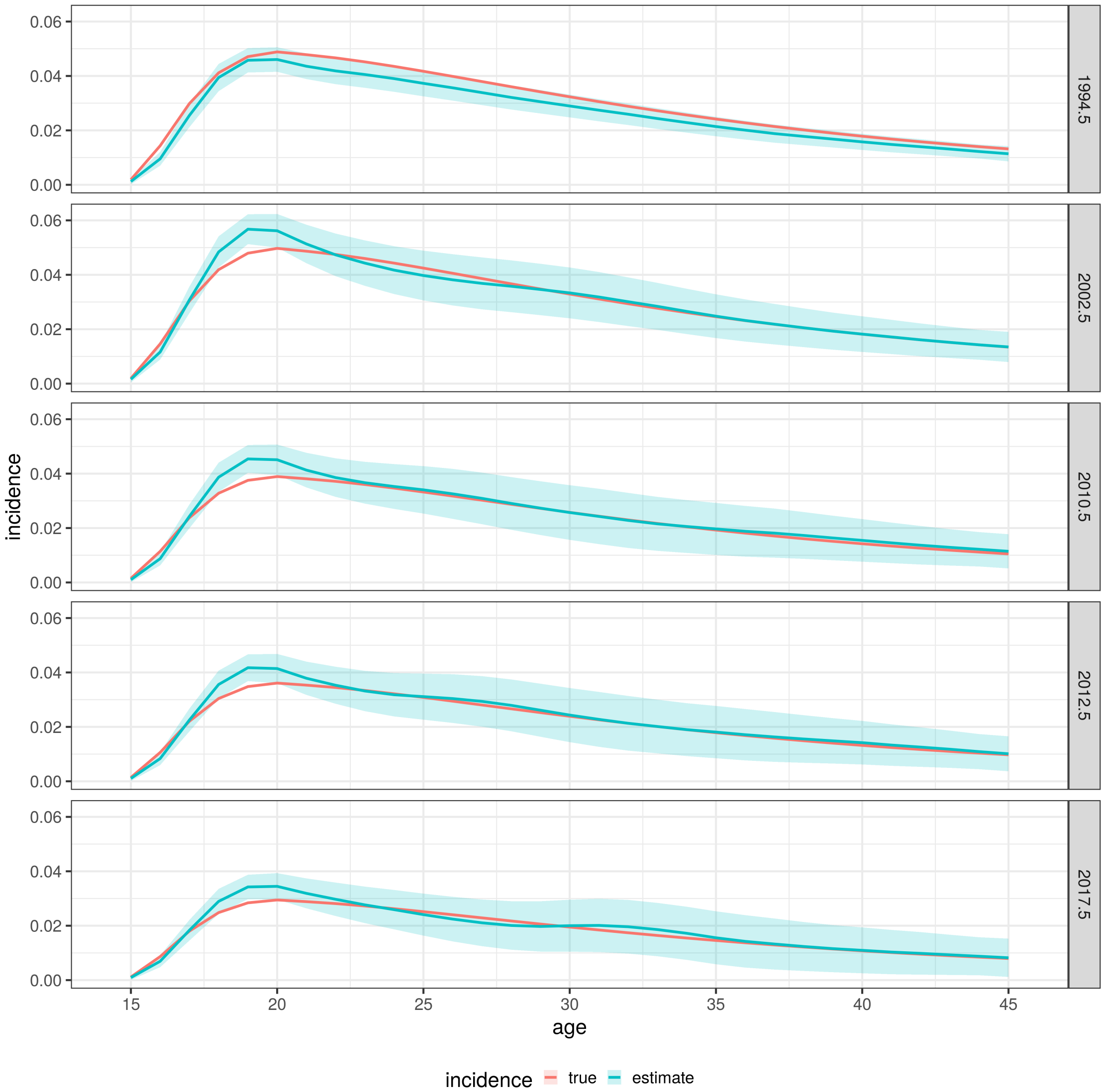

Figure 6.5: Midpoint incidence estimates for selected epidemic stages in (1992, 1997), (2000, 2005), (2008, 2013), (2010, 2015), and (2015, 2020) with inter survey interval 5.

Figure 6.5 show detailed age specific incidence estimates obtained at the canonical time points from two surveys conducted 5 years apart around the indicated time. The solid (blue) line is the expected point estimate, and the shading indicates the 95% confidence interval obtained from a data set which attains the expected value. The central 95% of point estimates generated by simulating many surveys yields much the same image. Estimation is slightly biased at the younger ages, when incidence varies sharply by age, and hence, for the individuals concerned, over time.

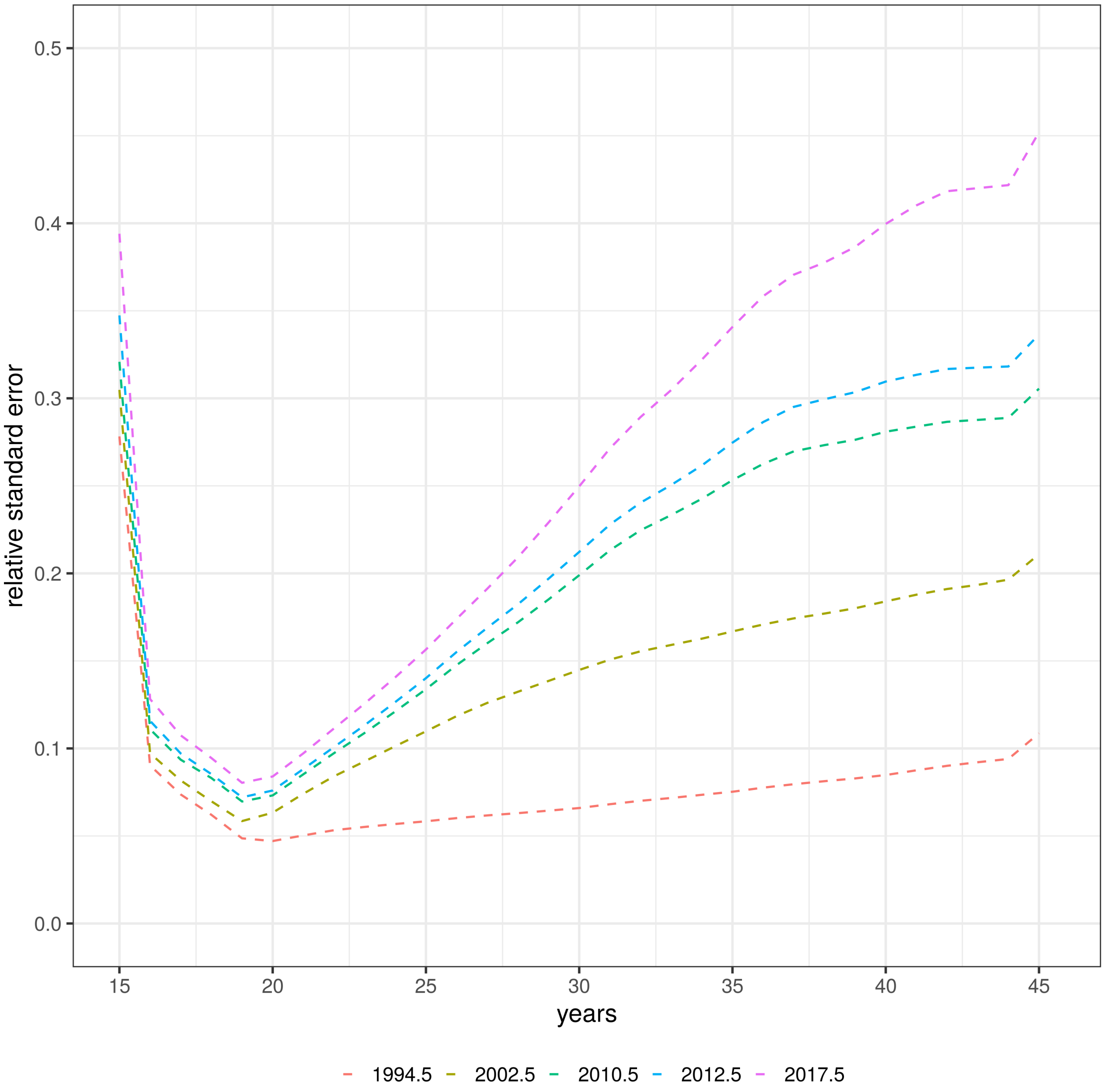

Figure 6.6: *Relative standard error of the midpoint incidence estimates for selected epidemic stages.* The Figure shows relative standard errors for a range of simulated pairs of survey each with a sample of 4000 per 5 year age bin. The surveys are simulated 5 years apart and the corresponding midpoints are 1994.5, 2002.5, 2010.5, 2012.5, and 2017.5.

Figure 6.6 shows much the same information as Figure 6.5, but disregards bias and shows the relative rather than absolute standard error. We see that at the edge of the data range, and at older ages, precision is substantially poorer than at the 'sweet spot' around 20 years, especially for a more mature epidemic, when prevalence is high.

6.4.6 Sensitivity Analysis (Different Epidemic Stages).

As noted earlier, for the core of our demonstrative calculations we have calculated the emergent excess mortality required by the Mahiane estimator, and supplied this number, free of charge, as it were, to our estimation procedure.

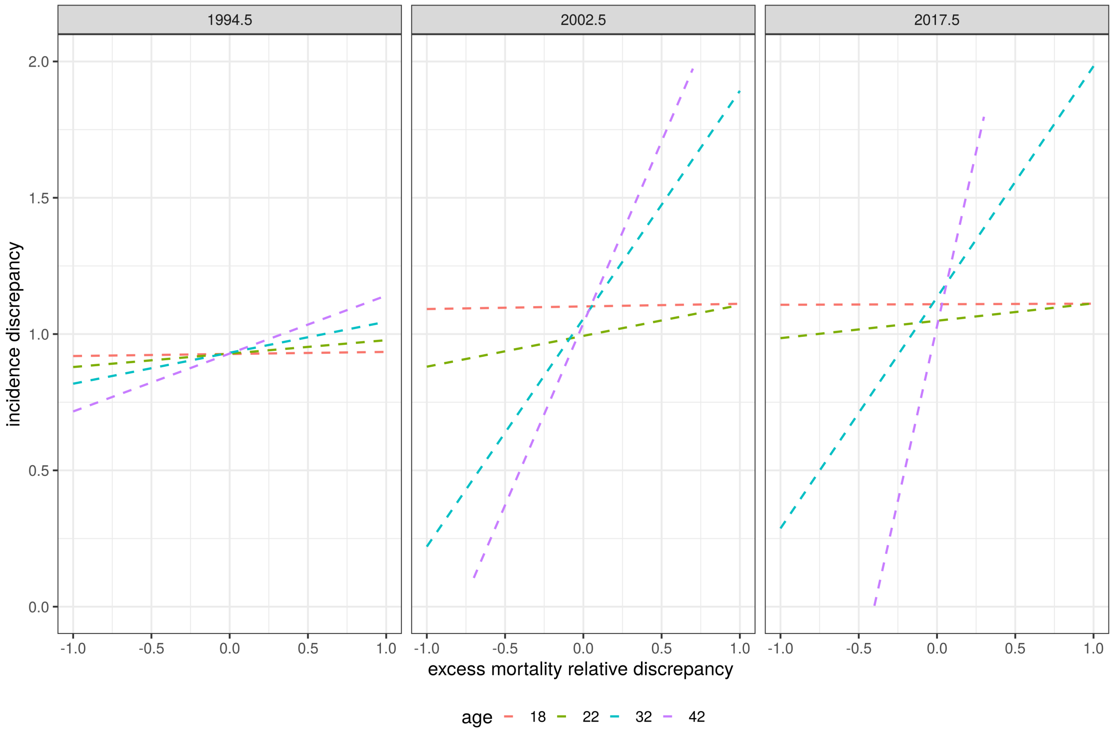

Figure 6.7: *Sensitivity analysis of the excess mortality's discrepancy ratio.*Sensitivity analysis of the excess mortality's discrepancy ratio. The picture depicts the bias of the incidence as a function of the bias in the excess mortality for selected ages at different epidemic stages.

In practice, it may be difficult to obtain a precise and unbiased estimate of this parameter, which summarises significant diversity and complexity. Figure 6.7 demonstrates the impact on the incidence estimate, of having an incorrect estimate of the excess mortality, supplied with a putative zero standard error. To scale the scenarios across the various indicated ages and times, we define the 'discrepancy' in a relative way from -1 to 1, the limits in which the excess mortality is estimated as 0, or twice its actual value, respectively. We see that as prevalence increases in a maturing epidemic, the sensitivity to the estimation of the excess mortality becomes very significant.

In practice, the estimates of excess mortality will have a significant standard error. Considering the same combination of ages and times, Figure 6.8 considers fractional/relative standard errors ranging from 0 to 1, and shows the relative standard errors thus induced on the incidence estimates.

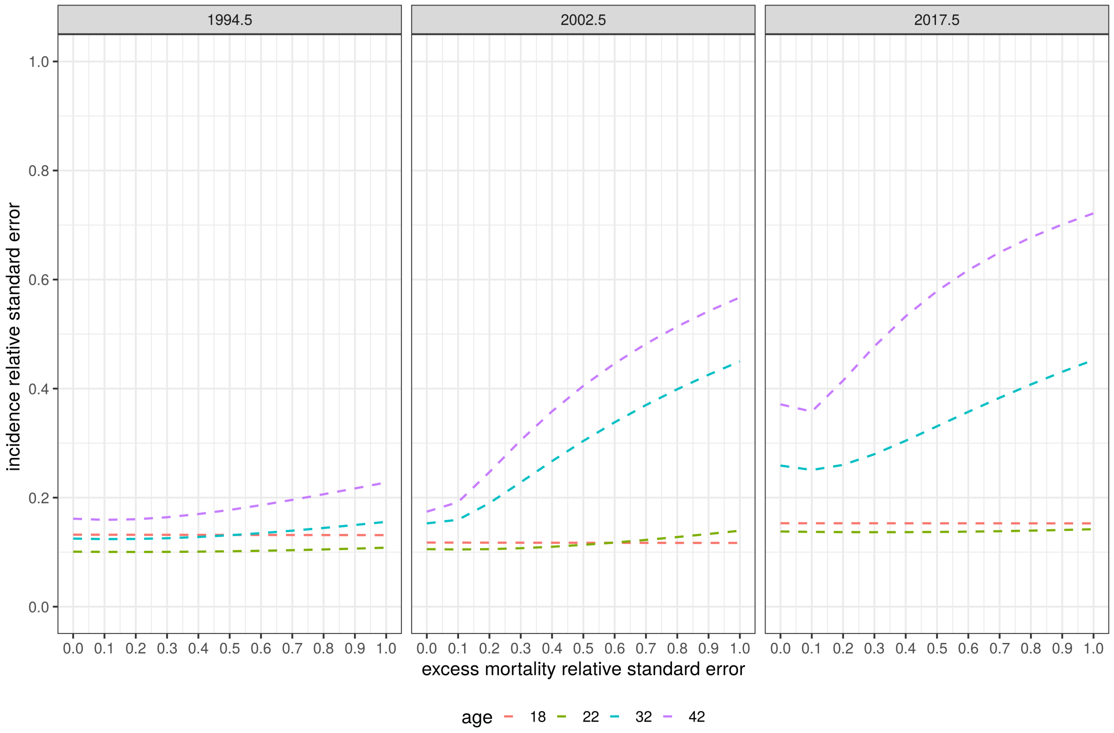

Figure 6.8: *Sensitivity analysis of the excess mortality's standard error.* The Figure depicts the relative standard error of the incidence estimate at selected ages 18, 22, 32, and 42 as functions of the relative standard error of the excess mortality, each age represents a specific epidemic stage for example age 18 represents an early epidemic state, a time when the incidence is rapidly rising and there is not much excess mortality. The analysis is based on a pair of cross sectional surveys simulated in (2015, 2020)

At younger ages, it matters little, but at older ages, the precision of the excess mortality estimate becomes important, as 1) it multiplies the prevalence in the estimator, and 2) the term in the estimator which has the prevalence gradient becomes less important as prevalence saturates.

6.5 Discussion/Conclusions

Previously, Mahiane et al. [3] derived an instantaneous, age-time specific incidence estimator - assuming some knowledge of survival after infection, summarised as differential mortality. That prior work did not critically evaluate techniques for summarising the population level survey data into a prevalence, P, and gradient of prevalence $\left(\frac{dP}{dt}\right)$. Given the ever-growing abundance of survey data with HIV infection status ascertainment, we investigated ways to optimally smooth such data for the purpose of incidence estimation using the Mahiane et al. [3] approach, leading to the following general remarks:

- Serostatus data can be smoothed into prevalence, including robust estimation of the gradient of prevalence, using generalised (binomial) linear regression on age and time, with a moving window for data inclusion (plus minus 5-10 years around an age of interest), and a third or fourth order polynomial fitting function.

- This is essentially the same finding as we made, in a companion piece [71], when investigating the smoothing of survey data where there was no immediate concern for extracting a prevalence gradient, while crucially relying, for incidence estimation, on ‘recent infection’ ascertainment- using the ideas of Kassanjee et al. [4].

- We have not investigated the prospects for consistently obtaining the required estimates for “mean excess attrition/mortality", which must be supplied in order to interpret the prevalence gradient in terms of incidence.We worked in the limit in which the relevant excess mortality is known precisely.

- The general approach demonstrated here can be refined/adapted to particular contexts by simulating scenarios resembling that context- which is not very difficult to do, using the simulation environment developed for the present investigation.

- Beyond minor fine tuning of data inclusion rules, polynomial order of fitting functions, possibly choice of link function in binomial regression, and quantifying excess mortality, there appears to be nothing noteworthy left to be done, that can extract any more incidence related information from large household surveys of the kind which are widely performed in the heavily HIV affected regions such as sub-Saharan Africa.

- As demonstrated in our analysis of recency data based incidence estimation, the base procedure naturally provides a finely age-resolved family of estimates, which can either be taken at face value, or used as the basis for further age averaging that may improve relative precision at the cost of hiding some age structure

Even when handling data in what we suspect is a nearly theoretically optimal way, there are fundamental sobering limitations:

- The synthetic cohort approach cannot avoid reliance on estimates of (net) infection associated excess mortality, which is further complicated in highly mobile populations.
- Even surveys of substantial size do not admit much disaggregation beyond the fundamental age dependence which is crucial in understanding HIV epidemiology.

We have emphasized the analysis of survey data, but also demonstrated that the same ideas and tools developed here can be used to support design decisions by investigating exposure to bias and statistical error, in simulated scenarios that resemble the intended real world application, but where the 'true' (simulated) incidence is known. Such investigations can help decide such key features of surveys as sample size/density, age ranges of interest, and timing.

What crucially remains to be understood, in this vein, is how best to use both the Kassanjee et al. [4] and Mahiane et al. [3] analyses, simultaneously, on data sets to which both are applicable and, hence, to understand the benefit of the additional investment in effort, complexity and expense which is implied by conducting ascertainment of 'recent infection' among confirmed HIV infected respondents. This is the subject of our third piece in this set of three companion pieces.

6.A Appendix: Simple Mahiane et-al Estimator – Proof

We derive the Mahiane et al. [3] estimator for a closed population based on the non-mechanistic SI model, i.e. we are not concerned about the interaction of people but what the dynamical evolution of prevalence (P) based on the interplay of incidence and mortality. Hence if we consider a birth cohort then the prevalence is defined by;

$$P = \frac{I}{S+I} \tag{6.A.1}$$

Where S - is the number of susceptible, I - of infected individuals in the birth cohort and, we aim at deriving a formula for incidence λ at any given age/time. Given the definition of prevalence and the respective population renewal equations for a standard SI model. The change in the prevalence is given by;

$$\frac{dP}{dt} = \frac{\frac{dI}{dt} \cdot (S+I) - (\frac{dS}{dt} + \frac{dI}{dt}) \cdot I}{(S+I)^2} \tag{6.A.2}$$

Where $\frac{dI}{dt}$ - is the rate of change of the infected population with respect to time and $\frac{dS}{dt}$ - is the rate of change of the susceptible population with respect to time and are given by;

$$\frac{dI}{dt} = \lambda S - I \cdot (\mu + M)$$

And,

$$\frac{dS}{dt} = -S \cdot (\mu + \lambda)$$

Note that μ is the base (background) mortality rate, λ is the incidence rate and M is the disease induced mortality (excess mortality). Substituting for $\frac{dS}{dt}$ and $\frac{dI}{dt}$ in Equation 6.A.3 yields,

$$\frac{dP}{dt} = \frac{((\lambda S - I \cdot (\mu + M)) \cdot (S+I) - (-S \cdot (\mu + \lambda) + \lambda S - I \cdot (\mu + M) \cdot I}{(S+I)^2} \tag{6.A.3}$$

Therefore substituting Equation 6.A.1 into Equation 6.A.3 and rearranging we have;

$$\frac{1}{(1-P)} \cdot \frac{dP}{dt} + MP = \lambda \tag{6.A.4}$$

Remark

This applies to a birth cohort and similar derivation for an age/time structured population yields

$$\frac{1}{(1-P)} \cdot \left(\frac{\partial P}{\partial t} + \frac{\partial P}{\partial a} \right) + MP = \lambda \tag{6.A.5}$$

In addition, if the population is viewed as birth cohorts, it follows that the age specific incidence at a given time yields Equation 6.A.4

6.B Appendix: Mahiane et-al Estimator – Quantifying Errors Using Delta Method

The Delta method [72, 73] is used to derive the uncertainty associated with the Mahiane estimator. The Delta method is based on the Taylor series expansion. For a small neighbourhood of μ, then $g(\mu)$ is considered a linear function.

Given a function

$$w = G(x) \tag{6.B.1}$$

Let μ be the mean of x, then based on the delta method, the mean and variance of w are

$$\bar{w} = G(\mu)$$

and

$$\text{variance}(w) = \left(\frac{dG}{dx}\right)^2 \cdot \text{variance}(x)$$

Definition 6.B.1. *If, in some neighbourhood of the point $X = M_x$, $Y = M_y$ the function $F(X,Y)$ is continuous and has continuous derivatives of the first and second order with respect to the arguments X and Y, the random variable $\hat{w} = F(\bar{x},\bar{y})$ is asymptotically normal, the mean and variance of limiting normal distribution being given by:*

$$mean = F(M_x, M_y) \tag{6.B.2}$$

$$var(\hat{w}) = \left[\frac{\partial F}{\partial x}\right]^2 \frac{\sigma_x^2}{n} + \left[\frac{\partial F}{\partial y}\right]^2 \frac{\sigma_y^2}{n} + 2\left[\frac{\partial F}{\partial x}\right]\left[\frac{\partial F}{\partial y}\right]\frac{\sigma_{xy}}{n} \; [72] \tag{6.B.3}$$

Given that the incidence function is a multi varied function of prevalence, the change in prevalence, and the excess mortality as highlighted in Equation 6.A.5 and Equation 6.B.1, we use the definition of delta method to derive the variance estimate. We first find the derivatives of the three variables in the incidence estimator.

Substituting the respective derivatives into the the delta method formula we have;

$$
\begin{aligned}
\operatorname{var}(\lambda) = & \left[\frac{\partial \lambda}{\partial P}\right]^2 \operatorname{var}(P) + \left[\frac{\partial \lambda}{\partial \left(\frac{dP}{dt}\right)}\right]^2 \operatorname{var}\left(\frac{dP}{dt}\right) + \left[\frac{\partial \lambda}{\partial M}\right]^2 \operatorname{var}(M) \\
& + 2 \cdot \left(\frac{\partial \lambda}{\partial P}\right) \cdot \left(\frac{\partial \lambda}{\partial \left(\frac{dP}{dt}\right)}\right) \operatorname{cov}\left(P, \frac{dP}{dt}\right) \\
& + 2 \cdot \left(\frac{\partial \lambda}{\partial P}\right) \cdot \left(\frac{\partial \lambda}{\partial M}\right) \operatorname{cov}(P, M) \\
& + 2 \cdot \left(\frac{\partial \lambda}{\partial M}\right) \cdot \left(\frac{\partial \lambda}{\partial \left(\frac{dP}{dt}\right)}\right) \operatorname{cov}\left(\frac{dP}{dt}, M\right)
\end{aligned}
\tag{6.B.4}
$$

Through substitution we have;

$$
\begin{aligned}
\operatorname{var}(\lambda) = & \left[\frac{1}{(1-p)^2} \cdot \frac{dP}{dt} + M\right]^2 \cdot \sigma_P^2 + \left[\frac{1}{1-p}\right]^2 \cdot \sigma_{\frac{dP}{dt}}^2 + [P]^2 \cdot \sigma_M^2 + \\
& 2 \cdot \left[\frac{1}{(1-p)^2} \cdot \frac{dP}{dt} + M\right] \cdot \left(\frac{1}{1-P}\right) \cdot \sigma_{P(\frac{dP}{dt})} + \\
& 2 \cdot \left[\frac{1}{(1-p)^2} \cdot \frac{dP}{dt} + M\right] \cdot P \cdot \sigma_{PM} + \\
& 2 \cdot \left[\frac{1}{1-p}\right] \cdot P \cdot \sigma_{M\frac{dP}{dt}}
\end{aligned}
\tag{6.B.5}
$$

where σ_{PM} is the covariance of mortality and prevalence, $\sigma_{P\frac{dP}{dt}}$ is the covariance of prevalence and the gradient of prevalence and σ_P is the standard error of prevalence, $\sigma_{\frac{dP}{dt}}$ is the standard error of the gradient of the prevalence, and σ_M is the standard error of excess mortality, and $\sigma_{M\frac{dP}{dt}}$ is the covariance of excess mortality and the gradient of prevalence. Note that the standard error of the incidence is: $se(\lambda) = \sqrt{var(\lambda)}$. Alternatively, one can use bootstrap to estimate the standard error of λ.

6.C Appendix: Functional forms of the Epidemiological/Demographic Functions.

Birth Rate

This is a constant function of time and for the present analyses has no impact on any calculations as the population state is used merely to determine prevalence.

Incidence

The incidence's functional form was selected to resemble the incidence estimates observed in generalised epidemics. Incidence was allowed to be high in younger ages and lower in mature ages.

The incidence rate is given $\lambda(t, a) = f(a) \cdot P(t)$ i.e. the product of a function of age and a function of time. The *age structure function*, $f(a)$ is parameterised using a log normal function, as used in Mahiane et al. [3], rescaled by R so that it lies between 0 and 1 as shown in Figure 6.9 (left). This means that $P(t)$ can be understood as the peak incidence experienced at time t (i.e. the highest incidence experienced by any age at time t):

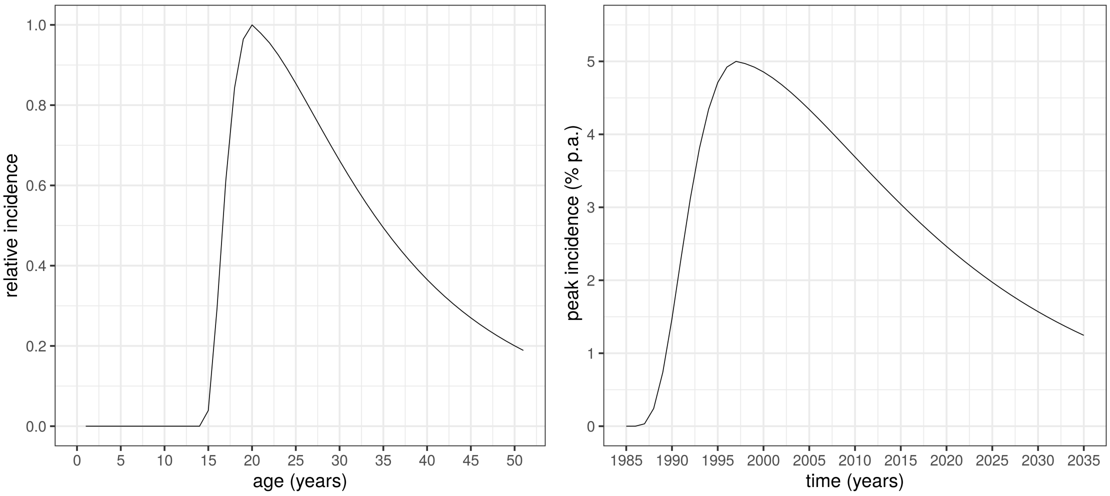

Figure 6.9: *The relative incidence as a function of age (left) and Peak incidence (i.e. incidence experienced by 20 year olds) as a function of time (right).* The incidence always peaks at age 20 and individuals aged 35 experience an incidence rate that is half of what is experienced by the 20 year olds at that given time. The maximum incidence in the simulation is 5% (p.a.) in 1997.

Background Mortality

Mortality among uninfected individuals is 1 % per annum at birth, and climbs linearly with age, reaching 3 percent per annum at age 50. We do not survey the population over age 50.

Excess Mortality and Averaged Excess Mortality (Platform Output)

The excess mortality as a function of age a, time t, and time since infection (τ) is another input parameter in the simulation platform. The excess mortality is parameterised using a Weibull functional. The scale parameter ($\alpha(a, t)$) is a function of age and time and specifies the median survival times. The median survival time is defined by linearly decreasing function of the age at infection between the maximum and minimum median survival times and depends on time (pre-treatment era versus treatment era), that is, the median survival times are defined for each era.

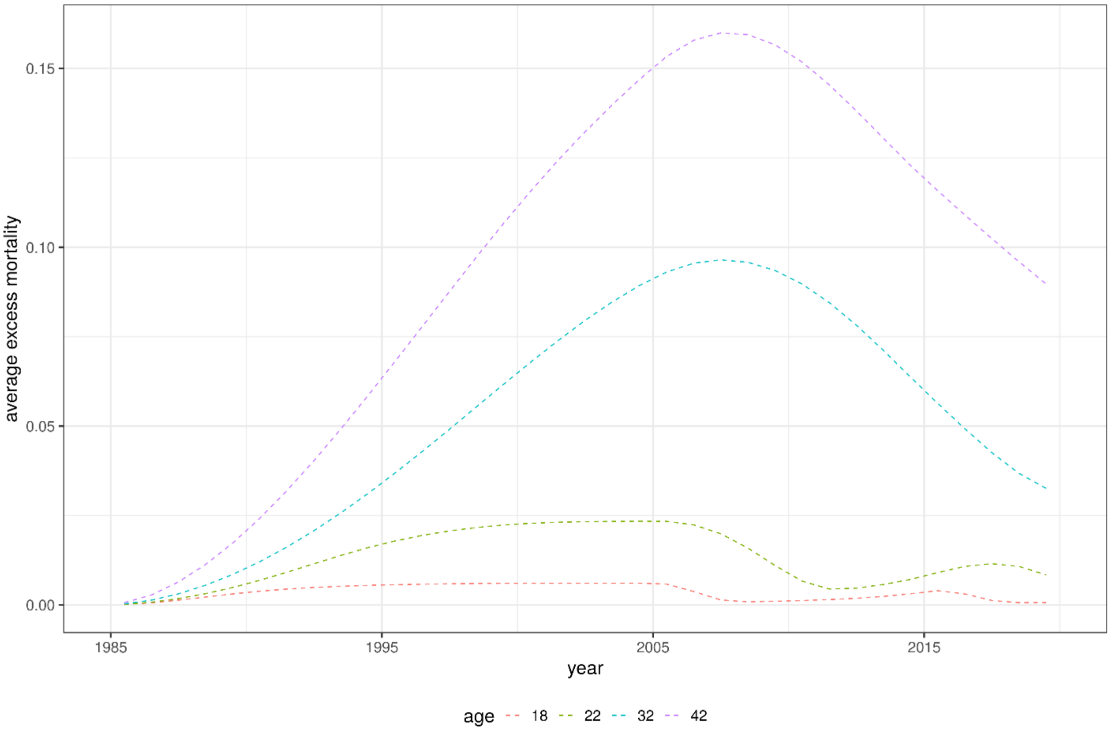

Figure 6.10: The plot describes averaged excess mortality as calculated from the input excess mortality by the simulation platform for selected ages 18, 22, 32, and 42.

Mahiane et al. [3] incidence estimator requires an averaged excess mortality as a function of time and age and hence the platform yields an averaged excess mortality for every age and time simulated. Figure 6.10, shows the averaged excess mortality for times 1985:20 at ages 18, 22, 32, and 42.

6.D Appendix: Emergent Population Prevalence

Figure 6.11 depicts the average prevalence output as a function of time. We chose to estimate incidence at various epidemic stages which are when prevalence is plateauing, steadily increasing and decreasing.

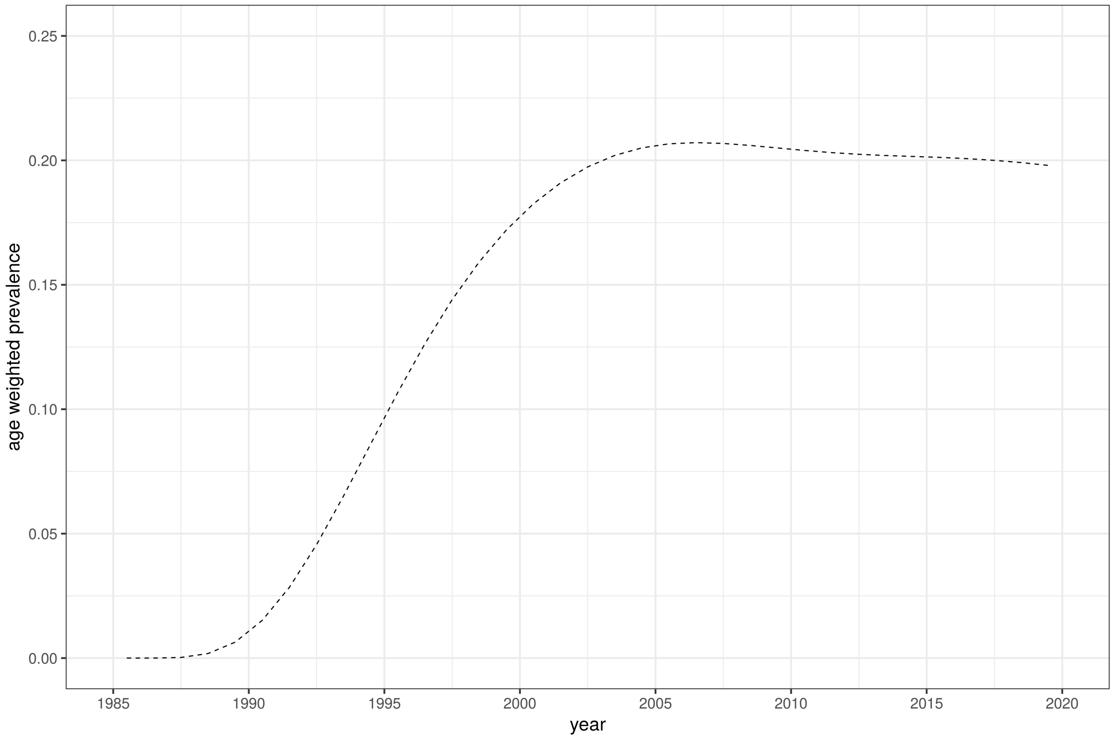

Figure 6.11: The plot highlights the trajectory of prevalence in time and is weighted to the age distribution in the simulated population.

6.E Appendix: Variance Estimate for the Slope-logit Link Function

The prevalence is estimated by using the generalised linear models package in **R** with link function ‘logit’, given by Equation 6.E.1 below;

$$log\left(\frac{P(a,t)}{1-P(a,t)}\right) = Q(a,t) = \beta_{00} + \beta_{01}\cdot(a-a_0) + \beta_{10}\cdot(t-t_0) + \beta_{11}\cdot(t-t_0)\cdot(a-a_0) + \beta_{02}\cdot(a-a_0)^2 + \beta_{20}\cdot(t-t_0)^2 + \ldots \tag{6.E.1}$$

which means that the prevalence is given by;

$$P(a,t) = \frac{1}{1+e^{-Q(a,t)}} \tag{6.E.2}$$

The fitted model yields the required parameters (regression coefficients (β_{00}, $\beta_{01}, \ldots$ in Equation 6.E.1)) of the age/time polynomial fitted. Independent of how complex $Q(a,t)$ is, evaluating the prevalence and its gradient at the point of interest (a_0, t_0), offers a simplification such that Equation 6.E.2 reduces to Equation 6.E.3 below;

$$P(a,t) = \frac{1}{1+e^{-\beta_{00}}} \tag{6.E.3}$$

To estimate the prevalence gradient $\frac{dP}{dt}$ from Equation 6.E.1 we derive the respective partial derivatives with respect to time $\frac{\partial P}{\partial t}$ and age $\frac{\partial P}{\partial t}$, since $\frac{dP}{dt} = \frac{\partial P}{\partial t} + \frac{\partial P}{\partial a}$.

And hence the corresponding gradient to $\frac{dP}{dt}$ is given by Equation 6.E.4 where $Q(a,t)$-is the age/time polynomial function, $Q_a(a,t)$ is the partial derivative of $Q(a,t)$ with respect to age (a) and similarly $Q_t(a,t)$, is the partial derivative of $Q(a,t)$ with respect to time (t).

$$G = \frac{dP(a,t)}{dt} = \frac{(Q_a(a,t)+Q_t(a,t))\cdot e^{-Q(a,t)}}{(1+e^{-Q(a,t)})^2} \tag{6.E.4}$$

By estimating the prevalence and the gradient at the point of interest Equation 6.E.4 reduces to Equation 6.E.5 below;

$$G = \frac{(\beta_{01}+\beta_{10})\cdot e^{-\beta_{00}}}{(1+e^{-\beta_{00}})^2} \tag{6.E.5}$$

The corresponding standard error of the gradient is not part of the accessible parameters and hence the standard error is estimated by either bootstrapping the data or using the delta method for error propagation, which is ideal for computationally intense simulations.

Note that G (Equation 6.E.4) is a function of 3 random variables (β_{00}, β_{01}, andβ_{10}) and hence we use the delta method (error propagation methods) [72, 73] to estimate the variance of G and is approximated by 6.E.6 below;

$$\begin{aligned}\mathrm{Var}\left(G\right) = &\left[\frac{\partial G}{\partial \beta_{00}}\right]^2 \sigma^2_{\beta_{00}} + \cdot \left[\frac{\partial G}{\partial \beta_{01}}\right]^2 \cdot \sigma^2_{\beta_{01}} + \left[\frac{\partial G}{\partial \beta_{10}}\right]^2 \cdot \sigma^2_{\beta_{10}} + \\ &2 \cdot \left[\frac{\partial G}{\partial \beta_{00}}\right] \left[\frac{\partial G}{\partial \beta_{01}}\right] \cdot \sigma_{\beta_{00},\beta_{01}} + 2 \cdot \left[\frac{\partial G}{\partial \beta_{00}}\right] \left[\frac{\partial G}{\partial \beta_{10}}\right] \cdot \sigma_{\beta_{00},\beta_{10}} + \\ &2 \cdot \left[\frac{\partial G}{\partial \beta_{01}}\right] \left[\frac{\partial G}{\partial \beta_{10}}\right] \cdot \sigma_{\beta_{01},\beta_{10}}\end{aligned} \tag{6.E.6}$$

The partial derivatives can be derived from 6.E.5 and standard errors of the fitted parameters ($\sigma_{\beta_{00}}$, $\sigma_{\beta_{01}}$, and $\sigma_{\beta_{10}}$), and covariances $\sigma_{\beta_{00},\beta_{01}}$, $\sigma_{\beta_{00},\beta_{10}}$, and $\sigma_{\beta_{01},\beta_{10}}$ are estimated through the *vcov* function in **R**.

6.F Appendix: Inclusion/Exclusion Distance and Polynomial Order Permutations

In the main article we presented some of the ages to demonstrate the trade off between bias and the standard error when varying the polynomial order and inclusion distance. In this section we present the relative errors of all the ages (15-45) at an epidemic stage simulated in 2017.5 in Figure 6.12- Figure 6.14.

For all the ages we observe a general trend of a decreasing relative bias and standard error as we increase the polynomial order and inclusion distance and as previously stated in the main body of the text a cubic or quartic order polynomial with an inclusion distance ≥ 5 yields accurate and informative incidence estimates.

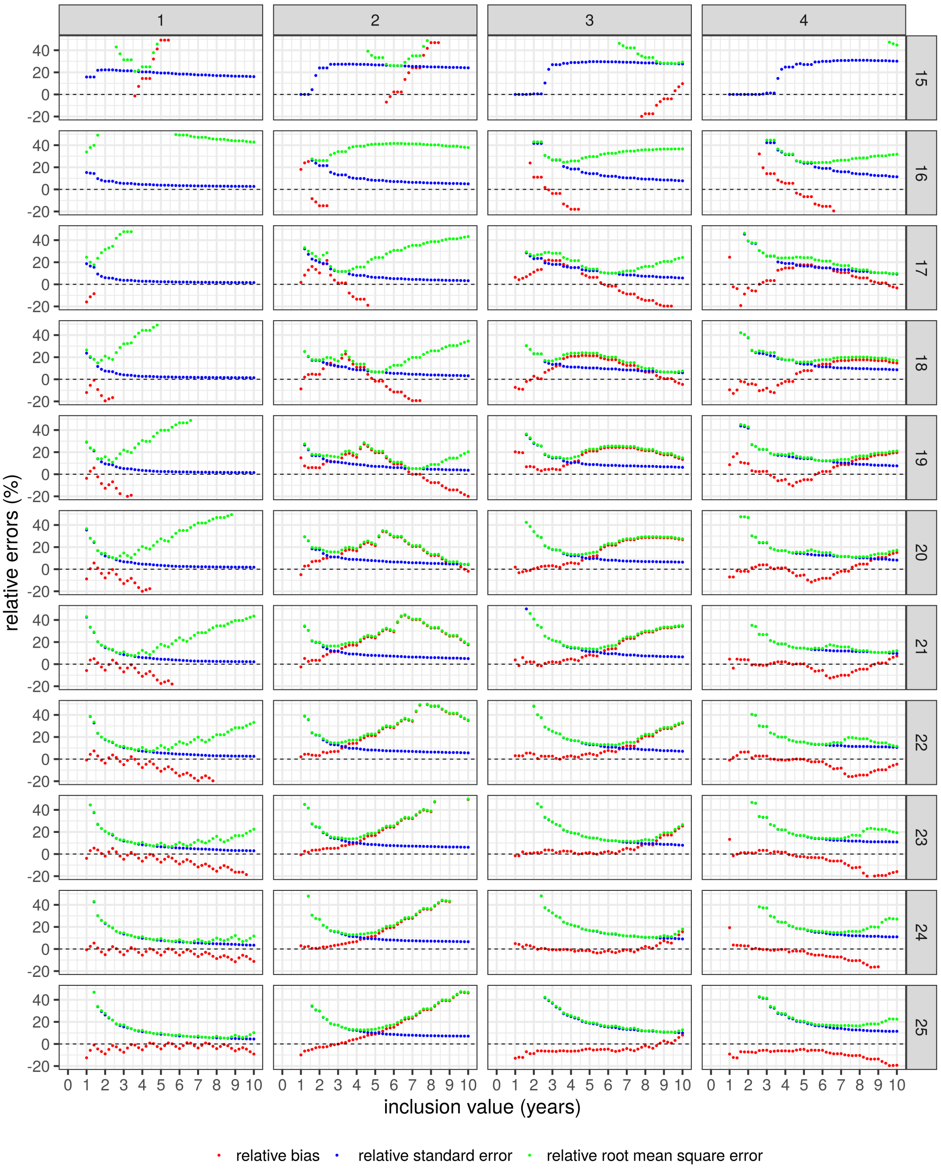

Figure 6.12: The relative errors for an epidemiological scenario simulated with surveys simulated 5 years apart (2015, 2020) and incidence estimated in 2017.5 for ages 18-25. Each survey has a sample size of 4000 per 5 year age bin.

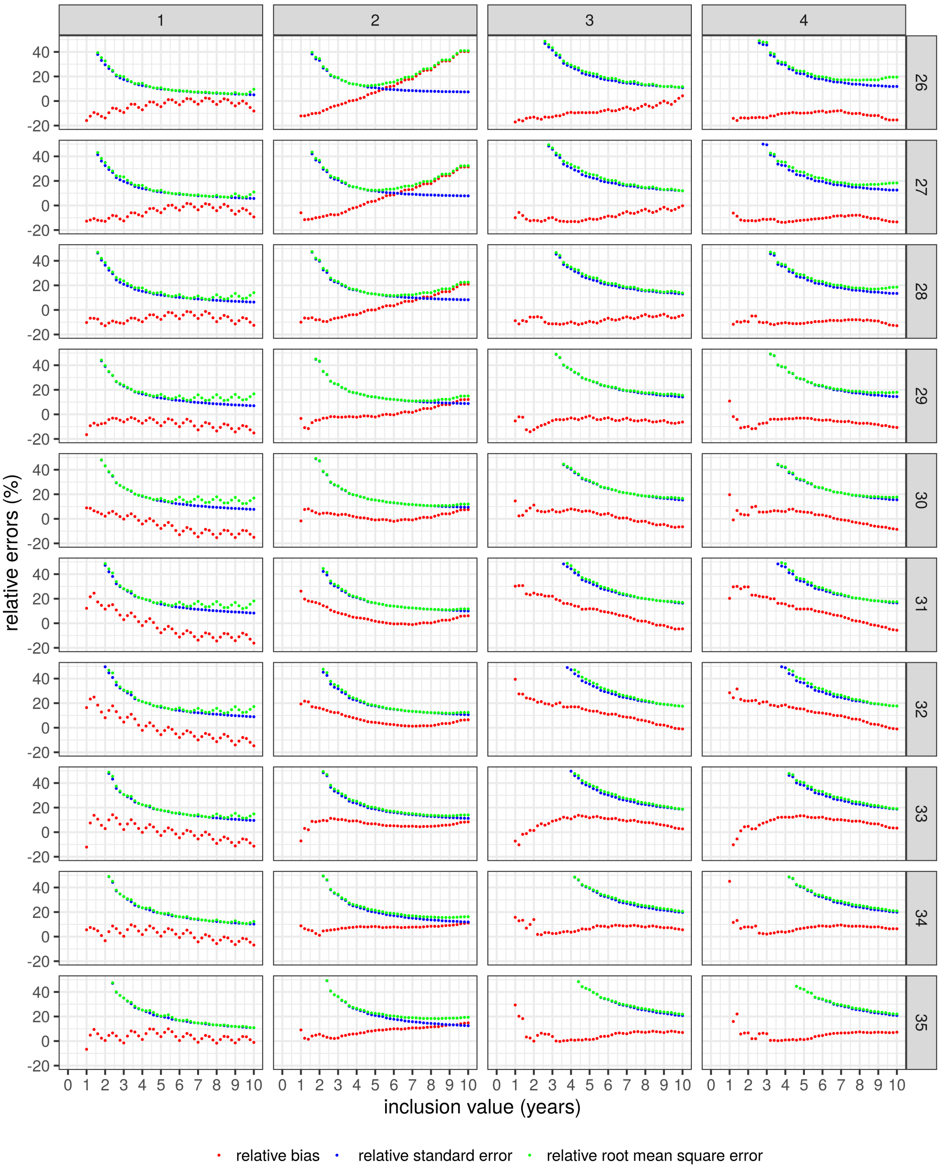

Figure 6.13: The relative errors for an epidemiological scenario simulated with surveys simulated 5 years apart (2015, 2020) and incidence estimated in 2017.5 for ages 26 − 35. Each survey has a sample size of 4000 per 5 year age bin.

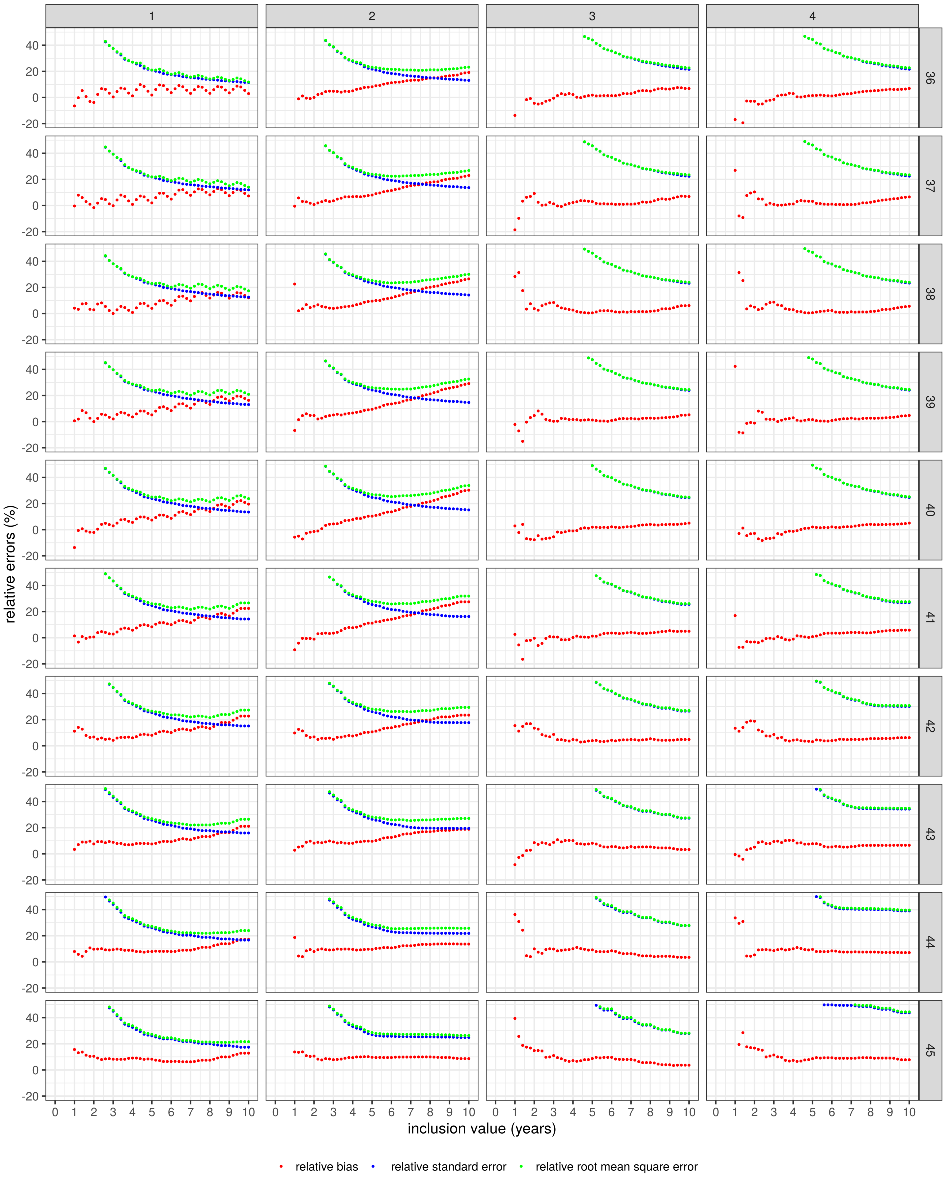

Figure 6.14: The relative errors for an epidemiological scenario simulated with surveys simulated 5 years apart (2015, 2020) and incidence estimated in 2017.5 for ages 36 – 45. Each survey has a sample size of 4000 per 5 year age bin.

Appendix: Relative Error Histograms

We present alternatives to Figure 6.2, presented in the main body of the article as Figure 6.15 to Figure 6.17 and each histograms shows the distributions of the relative errors (bias, standard error and root mean square error) for given polynomial orders and inclusion distances.

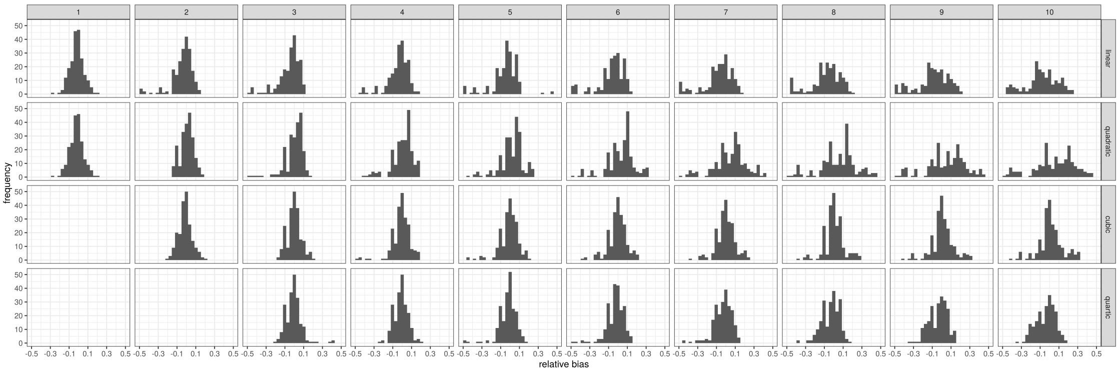

Figure 6.15: The plot shows distributions of the relative bias (random errors) for each permutation of polynomial order (linear-quartic) and inclusion distance (1-10), for sample size 4000 per 5 year age bin and link function 'logit'.

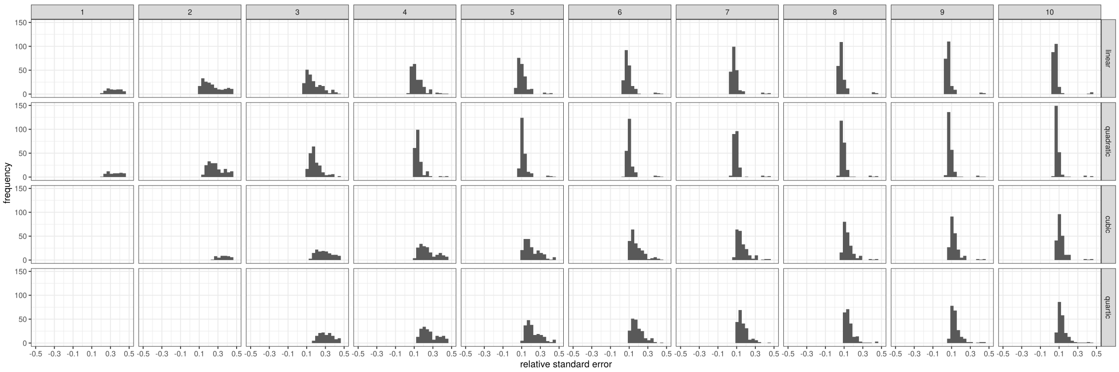

Figure 6.16: The plot shows distributions of the relative standard error (statistical error) for each permutation of polynomial order (linear - quartic) and inclusion distance (1-10), for sample size 4000 per 5 year age bin and link function 'logit'.

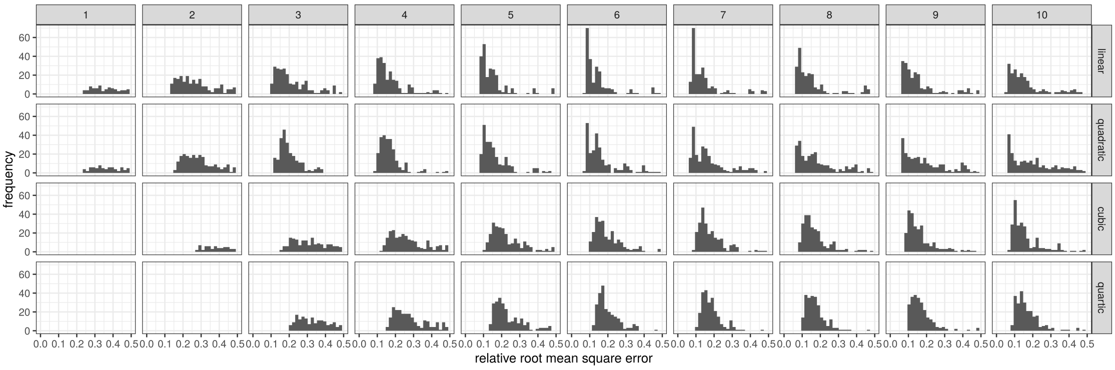

Figure 6.17: The plot shows the overall sum of squares of the relative bias and relative standard error from Figure 6.15 and Figure 6.16. The inter survey interval is set at 5 and the sample size is 4000/5yr age bin. Each facet depicts the distribution of the R. RMSE for a specific inclusion distance and Taylor order permutation.

Chapter 7

The added value of recent-infection testing in population-based HIV surveys

7.1 Abstract

Background

There is no clear consensus on how best to use increasingly available data derived from large population-based surveys featuring HIV infection status ascertainment. In particular, for the purpose of estimating HIV incidence, there is considerable scope for better elucidation of the benefit of adding 'recent infection' ascertainment, which adds considerable additional cost and complexity to surveys which are already costly and complex.

Methods

Using an epidemic/survey simulation tool developed for this and some closely related investigations, we explore the value added by 'recent infection' data from population surveys, to support HIV incidence estimation. This directly piggy-backs on to two companion pieces which have explored, independently, the use of the 'synthetic cohort' paradigm of Mahiane et al (analysing age/time structure of prevalence, in conjunction

This chapter is available as a preprint: authors: Mhlanga, L. Grebe, E. Welte, A., Title: The added value of recent-infection testing in population-based HIV surveys, DOI : 10.21203/rs.3.rs-996585/v2 with estimates of mortality) and the paradigm of Kassanjee et al (focusing on 'recent infection' data).

Results

Our headline findings are that: 1) Recent infection data adds marginal benefit to surveillance focused on the early years after sexual debut, which can reasonably be taken to be a core sentinel group in which surveillance is significantly more efficient than attempts to cover all ages; and 2) by contrast, recent infection data is crucial for the reliable estimation of incidence trends when only two cross sectional surveys are available. We detail numerous components of a general and robust approach to analysing data when both the Mahiane and Kassanjee analyses are in play.

Conclusion

Our main results present non-trivial dilemmas for survey design, as recency data is crucial for stabilising the more timely estimates, but of marginal benefit for the most important sentinel group. We hope that adaptation of our analysis, to simulated scenarios closely aligned to specific contexts facing expensive choices, will support rational investments in, and use of, precious surveillance opportunities and data sets.

7.2 Introduction

A global HIV epidemic has been raging for four decades, and still there is no clear consensus on how best to estimate HIV incidence: i.e. the rate of new infections in a population. Estimating prevalence (the proportion of infected individuals in a population) is relatively straightforward, but not nearly as informative, especially about the recent impacts of interventions, policies, and changing social norms. Incidence estimation for chronic conditions is in general difficult - unlike for transient conditions, for which prevalence and incidence are simply related.

Large scale population-level cross-sectional surveys that include HIV status determination, and in many cases also ascertainment of 'recent infection' as defined by objective laboratory procedures, have been conducted in many Sub-Saharan countries, and have become a/the headline data source for epidemiological assessments at the national and supra-national regional level. Within the last two decades, variations of such surveys have been executed multiple times in numerous countries, leading to rich data sets tracking the prevalence of HIV infection and the 'prevalence' of 'recent infection' among confirmed HIV positive subjects, over time and by age.

- We deployed a comprehensive demography/epidemiology/survey simulation platform which we use again in the present work, and which is separately outlined in more detail (ref to forthcoming)
- We proposed a generic approach to age/time regression in order to use the approach of Kassanjee et al. [4], which crucially relies on ascertainment of 'recent infection' to leverage analysis which is inspired by the simple relationship between incidence and prevalence for transient conditions.
- We demonstrated the applicability of a similar generic regression approach to the estimation of incidence by the approach of Mahiane et al. [3], which crucially relies on the estimation of a 'prevalence gradient', in conjunction with the estimation of a specifically defined 'excess mortality/attrition' for HIV positives (which falls under the broad umbrella of 'synthetic cohort' analysis).

Increasingly, numerous countries, or subnational regions, have data which allows the applications of both the Kassanjee and Mahiane framework. The question which then naturally arises is how best to combine the two methods, which provide nominally separate estimates, but which are correlated in complex ways as they both rely on the same underlying serostatus data which always comprises the bulk of the database. We view this question through the lens of the benefit of the recency data, seen as an add-on to the main prevalence data set. This reflects the points that

- There is no sensible survey design that generates recency data but not prevalence data, and
- At the design stage, before data is available to analyse, one will want to be clear about the benefit of performing the recency ascertainments, which invariably imply substantial increases in both cost and complexity of surveys that are already major undertakings without this requirement.

In outline, the present work has the following high-level components:

1. Simulating ‘realistic’ epidemics and cross-sectional surveys
2. Simulating realistic multiple cross-sectional surveys, where ’recent infection’ is defined by a probability of testing ’recent’ (on some algorithm) which depends explicitly on a function of time-since-infection in a way that is inspired by actual available tests of this kind.
3. Applying various smoothing algorithms to the survey data, in order to infer (age- and time-structured) prevalence of HIV infection and (age-, time-, and time-since-infection-structured) prevalence of ‘recent infection’.
4. Estimating incidence, and incidence differences/trends, from these smoothed functions, using the Kassanjee and Mahiane frameworks, separately and in conjunction.
5. Evaluating the relative merits of the various combinations of approaches, by comparing estimates with the known incidence parameter values in the simulations
6. Proposing guidance on the use and value of ‘recent infection’ ascertainment (for the purpose of HIV incidence estimation)

7.3 Methods.

7.3.1 Optimally Weighted (Midpoint incidence Estimation)

As noted, we are building on work reported in two companion pieces to this one, based primarily on the simulation of a number of cross-sectional surveys in a South-Africa-like epidemic. We have already systematically investigated ways to adapt the methods of Mahiane et al. [3] and Kassanjee et al. [4], to estimate incidence based on survey data from one or more cross sectional surveys, and incidence differences for cases with two/-more cross-sectional surveys.

The functional forms of each of the incidence estimators are

$$I_M = \frac{1}{1-P} \cdot \frac{dP}{dt} + M \cdot P \tag{7.3.1}$$

$$I_K = \frac{P(R-\beta)}{(1-P)\cdot(\Omega-\beta\cdot T)} \tag{7.3.2}$$

Where P is the prevalence of HIV, $\frac{dP}{dt}$ is the gradient of the prevalence as seen from the point of view of a cohort of individuals of identical age, R is the prevalence of 'recent infection' (recency) among the HIV positive subjects, Ω is the Mean Duration of Recent Infection (MDRI), β is the false recency rate (FRR), and T is the time cut-off for being classified as recently infected without being 'falsely' recent. In our simulations, the Mean Duration of Recent Infections (MDRI), false recent rate (FRR), and differential mortality are known exactly, because they are explicitly specified, or emerge from (and are evaluated in) the simulation platform.

To combine the information from the two estimators, we first define a general weighted average of the two estimators:

$$I_{Opt} = W \cdot I_M + (1 - W) \cdot I_K \tag{7.3.3}$$

$$\text{se}\left(I_{Opt}\right) = W^2 \cdot \sigma^2_{I_M} + (1 - W)^2 \cdot \sigma^2_{I_K} + 2 \cdot W \cdot (1 - W) \cdot \text{COV}\left(I_K, I_M\right) \tag{7.3.4}$$

We find the optimal weight by differentiating Equation 7.3.4 with respect to W and setting that to zero:

$$W = \frac{\sigma^2_{I_K} - \rho \cdot \sigma_{I_K} \cdot \sigma_{I_M}}{\sigma^2_{I_k} + \sigma^2_{I_M} - 2 \cdot CoV(I_K, I_M)} \tag{7.3.5}$$

Where σ_{I_k} and σ_{I_M} are the standard errors for I_K and I_M respectively, and ρ is the pearson correlation coefficient of I_K, and I_M COV $\left(I_K, I_M\right)$ is the covariance of I_M, and I_K. According to delta method analysis [72, 73] the $COV(I_M, I_K)$ is given by;

$$COV\left(I_M, I_K\right) = \frac{\partial I_M}{\partial P} \cdot \frac{\partial I_K}{\partial P} \cdot \sigma^2_P \tag{7.3.6}$$

$$\frac{\partial I_M}{\partial P} = \frac{1}{(1 - P)^2} \cdot \frac{dP}{dt} + M$$

and

$$\frac{\partial I_K}{\partial P} = \left(\frac{P\,(R - \beta)}{(1 - P)^2 \cdot (\Omega - \beta \cdot T)}\right) + \left(\frac{(R - \beta)}{(1 - P) \cdot (\Omega - \beta \cdot T)}\right)$$

$$\begin{aligned}\mathrm{Cov}(I_M, I_K) = &\left[\left(\frac{P\,(R-\beta)}{(1-P)^2\cdot(\Omega-\beta\cdot T)}\right)+\left(\frac{(R-\beta)}{(1-P)\cdot(\Omega-\beta\cdot T)}\right)\right]\cdot \\ &\left[\frac{1}{(1-P)^2}\cdot\frac{dP}{dt}+M\right]\cdot\sigma_P^2\end{aligned} \tag{7.3.7}$$

The covariance can be estimated either by equation 7.3.7, or by repeatedly simulating the survey (for example 10,000 times) or resampling from a particular data set (i.e. bootstrapping) and for each iteration estimating I_K and I_M, and hence estimating the $COV(I_M, I_K)$ from the iterates.

Stable approaches to the smoothing of survey data to estimate the prevalence P, the prevalence of recency R, and crucially the gradient of prevalence, $\frac{dP}{dt}$, were discussed in-depth in the two preceding companion papers. In short, a 'one size fits most' approach can be summarised as follows:

- Use generalised linear models (GLM) to fit, in turn, the serostatus and the recency data, with either third or fourth order polynomials in age and time.
- Repeat the fitting procedure for each age and time for which incidence estimates are to be obtained, including data points by a simple proximity rule such as being within some (temporal) 'distance' to the age of interest.
- Use a logit or identity link function for fitting P and a logit or complementary log log link function for R, with some age or age/time inclusion-distance rule.
- By default, we settled on using a cubic order polynomial with an inclusion distance of 6 years and link functions logit for P and complementary log-log for R.
- The prevalence of HIV, prevalence of recent infection among positives, and prevalence gradient $\frac{dP}{dt}\left(=\frac{\partial P}{\partial t}+\frac{\partial P}{\partial a}\right)$ are extracted from the fitted models and inserted into the Mahiane and Kassanjee estimators.

7.3.2 Single Cross-sectional Surveys.

In addition to the usual semi-realistic 'South Africa - like' scenario, we also simulated a stable epidemic with a calendar-time invariant (but age dependent) incidence function, and also used a calendar-time invariant excess mortality (resembling a 'no treatment' scenario).

7.3.3 Two Cross-sectional Surveys.

Realistically, two cross sectional surveys may utilise different 'recency' ascertainment tests, leading to a different values for MDRI and FRR, as these two parameters are context specific [9, 74, 75]. Hence, to avoid this distraction for the present purposes, surveys are simulated with the same recency test.

7.3.4 Incidence Trends.

Incidence trends are a crucial indicator of whether interventions or emergent changes in habits and services are reducing the transmission of HIV. We investigate the prospects for estimation of an incidence trends from two cross sectional surveys. We show how to yield accurate and informative age specific and age range incidence difference estimates and the effect of sample size on the precision of the estimates.

In cases where we attempt to estimate incidence difference from two cross sectional surveys, we estimate age specific incidence at the two survey dates using a shared estimate of $\frac{\partial P}{\partial t}$ in both I_M estimates.

7.4 Results

7.4.1 Single Cross-sectional Surveys

Figure 7.1 shows the incidence estimates from a single cross sectional survey in a scenario in which there is no time dependence to any parameters or prevalences. The key point appears to be that even when the correct value of $\frac{dP}{dt}$ is provided, the highest and most age dependent values of incidence are not being estimated without significant bias by the Mahiane estimator, i.e. when the recent infection data is being ignored. In practice, sample sizes (or sampling density) is likely to be smaller, and the bias shown here may be substantially swamped by poor precision.

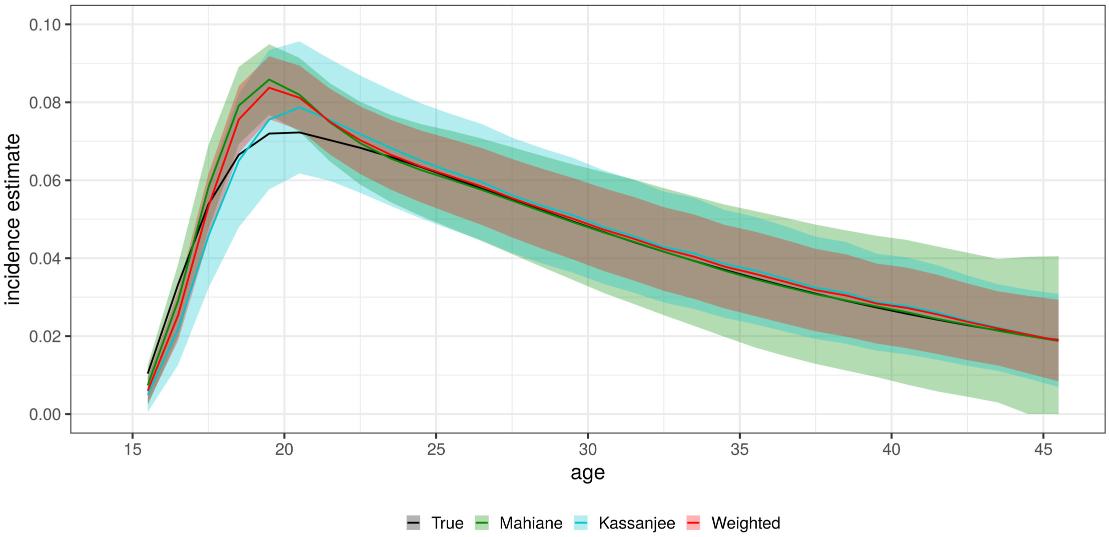

Figure 7.1: The plot compares 3 incidence estimates (Kassanjee et al, Mahiane et al., and optimally weighted estimators) to the true incidence in the platform. Each survey has a sample size of 4000 per 5 year age range with link functions logit for P and clog-log for R using a cubic order polynomial with an inclusion distance of 6. The input incidence function is time invariant and the excess mortality function does not include treatment.

7.4.2 Midpoint Incidence Estimates Comparison (I_K, I_M, and I_{Opt}).

Figure 7.2 and 7.3, shows the incidence estimates at 4 time points corresponding to either an early epidemic (1994.5 and 1999.5) or a mature epidemic stage (2010.5 and 2015.5).

While a logit link function for prevalence provides some stability by automatically constraining the prevalence to values between 0 and 1, it appears that an identity link function may offer superior fitting at various epidemic stages, so this should be explored in simulations adapted to mimic any context in which there has been a major investment in data of this kind.

These results also show the consistent trend that, for young ages, the Mahiane estimator provides most of the information about incidence, and for older ages the Kassanjee estimator provides most of the information.

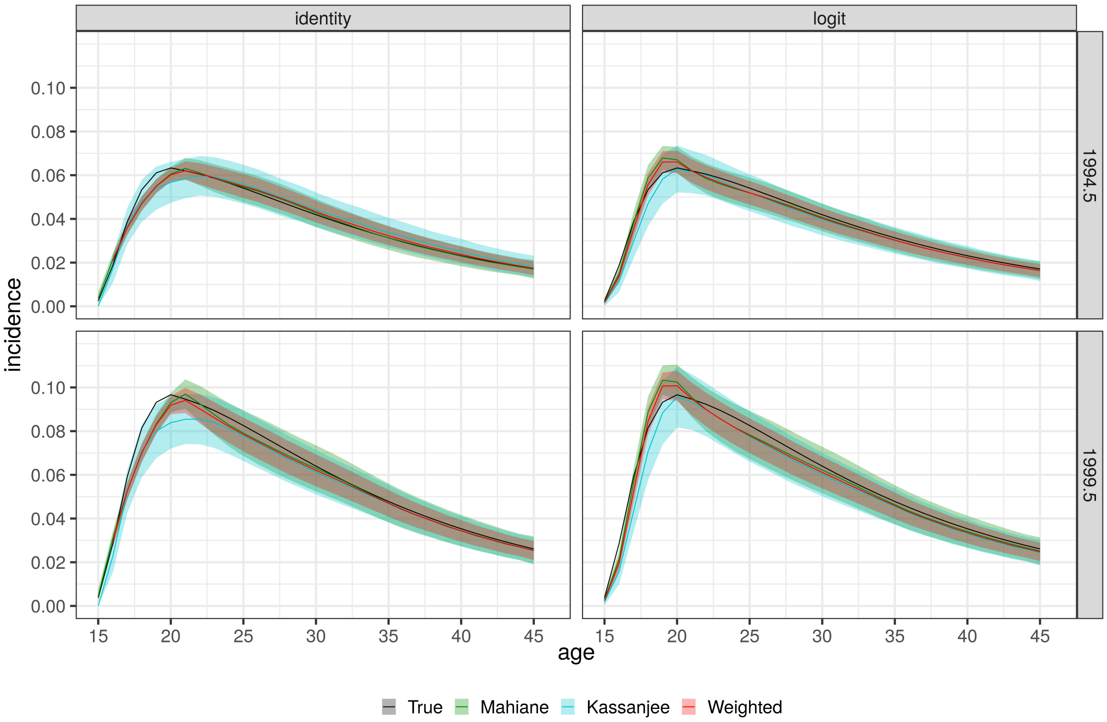

Figure 7.2: Midpoint incidence estimates from a pairs of simulated cross-sectional surveys (1992, 1997), and (1997, 2002)). The incidence estimates are based on 3 approaches, namely Kassanjee et al, Mahiane et al., and optimally weighted incidence estimators. Generous sample sizes of 4000 per 5-year age range were used with link functions identity and logit for P, and c log-log for R using a cubic order polynomial with an inclusion distance of 6.

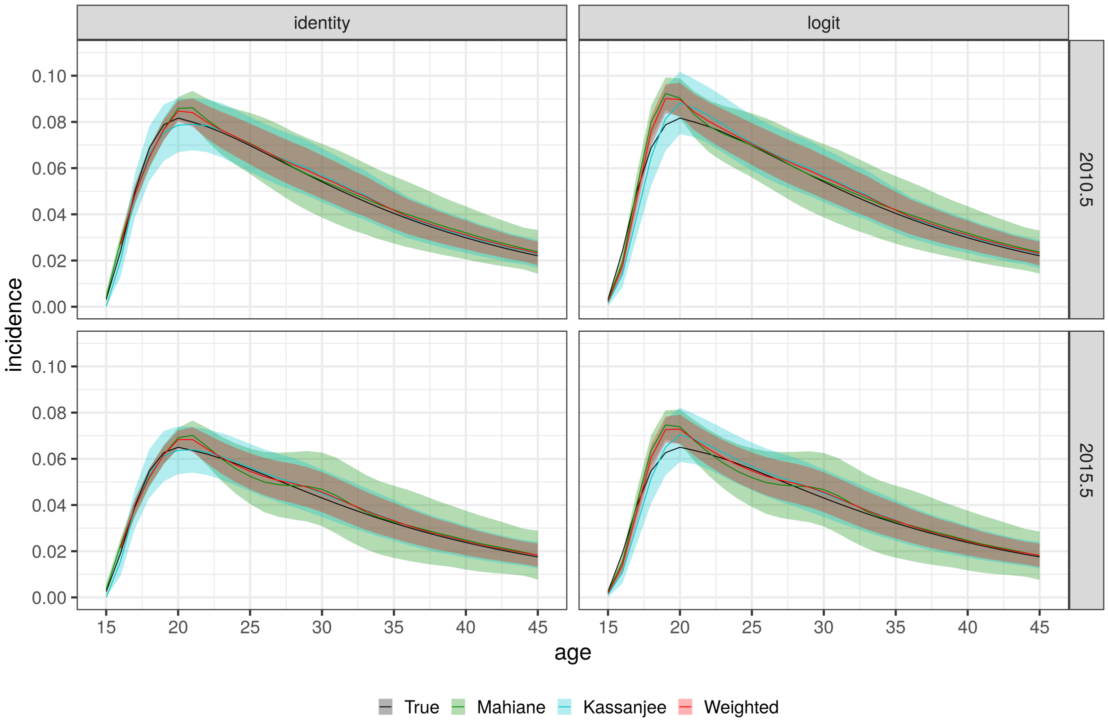

Figure 7.3: *Midpoint incidence estimates from pairs of simulated cross-sectional surveys (2008, 2013), and (2013, 2017)) estimated using the link functions identity (left) and logit (right).* The incidence estimates are based on three approaches, namely Kassanjee et al, Mahiane et al., and optimally weighted incidence estimators. Generous sample sizes of 4000 per 5-year age range were used with link functions identity and logit for P, and c log-log for R using a cubic order polynomial with an inclusion distance of 6.

7.4.3 Comparison of Methods for Estimating the Optimal Weight W (Delta Method vs Bootstrap)

We compared the two approaches of calculating W (an analytical delta method versus the numerical bootstrap approach) and their effect on I_{Opt} and the resulting standard errors. The results are shown in Table 7.1.

There is no substantial (indeed hardly any) difference between the estimates derived from the bootstrap approach and the analytical approach. The concordance of both the standard error and the realised point estimates shows that for computationally intense investigations, the delta method is a good proxy to estimate the standard error. On the other hand, once a major investment has been made in a complex survey, there is no obstacle to implementing an ultimately more robust bootstrap based calculation.

Table 7.1: *I_{Opt} estimates derived from two estimates of W.* W_1 is the Optimal weight calculated from delta method (analytical function) $COV(I_M, I_K)$ versus W_2 is the derived from 10000 bootstrap estimates of I_M and $W_1$to estimate $COV(I_M, I_K)$. To demonstrate this point we use ages 18, 20, 30, and 40, for a single epidemic stage epidemic stage with surveys simulated in 2015 and 2020 and the incidence is estimated at midpoint (2017.5).

	Incidence Point estimate			**Incidence Standard Error**		
Age	**Delta Method**	**Bootstrap**	**Concordance**	**Delta Method**	**Bootstrap**	**Concordance**
18	2.72	2.75	98.9	0.255	0.265	96.1
20	3.33	3.35	99.4	0.267	0.272	98.4
30	1.98	2.00	99.0	0.266	0.271	98.4
40	1.09	2.00	99.1	0.195	0.199	97.9

7.4.4 Sensitivity of the Standard Error I_{Opt} to W (Midpoint)

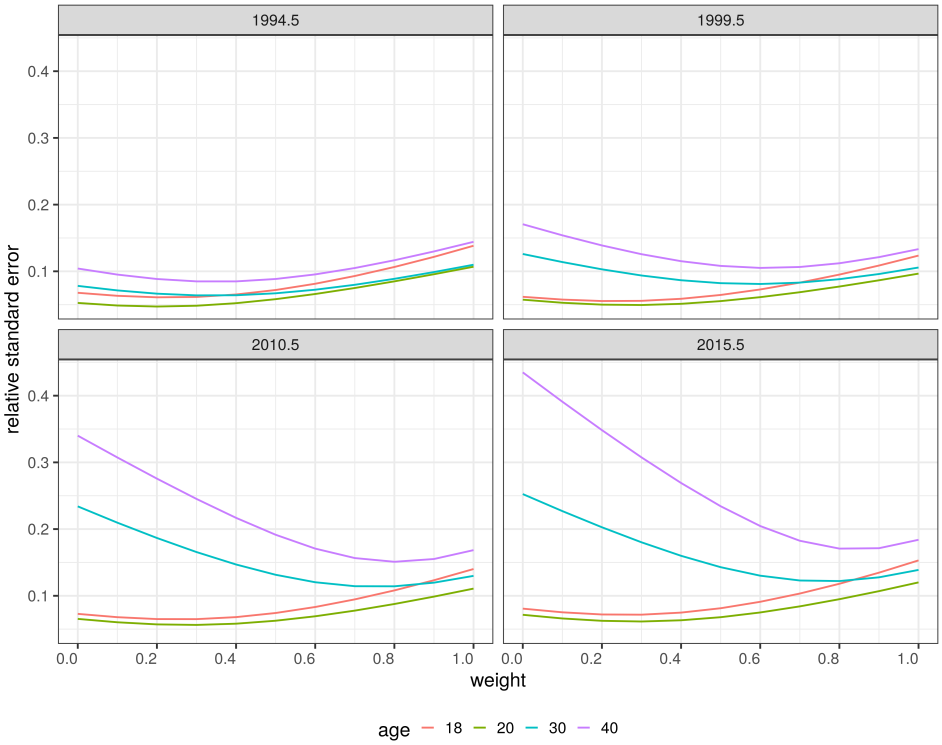

Figure 7.4: *Relative standard error of the optimally weighted incidence estimators as a function weights ranging for 0 to 1 weighted to the 'Recency' estimator.* The plot shows the relative standard errors for ages 18, 20, 30, and 40 epidemic stages 1994.5, 1999.5, 2010.5, 2012.5, and 2015.5.

Figure 7.4 expresses the relative standard error of I_{Opt} as a function of the normalised weight (W). For all 5 epidemic stages, and selected, ages, there is no sharply defined optimal weight required to estimate I_{Opt}. For example, the relative error at age 20 is almost flat for a range of W values (0 to 0.5), and hence any value between 0 and 0.5 yields much the same value of I_{Opt}. The weighting scheme in early epidemics (1992.5) somewhat favours I_M and, as the epidemic matures, and at older ages, the weighting scheme favours I_K.

7.4.5 Incidence Estimates at Survey Times

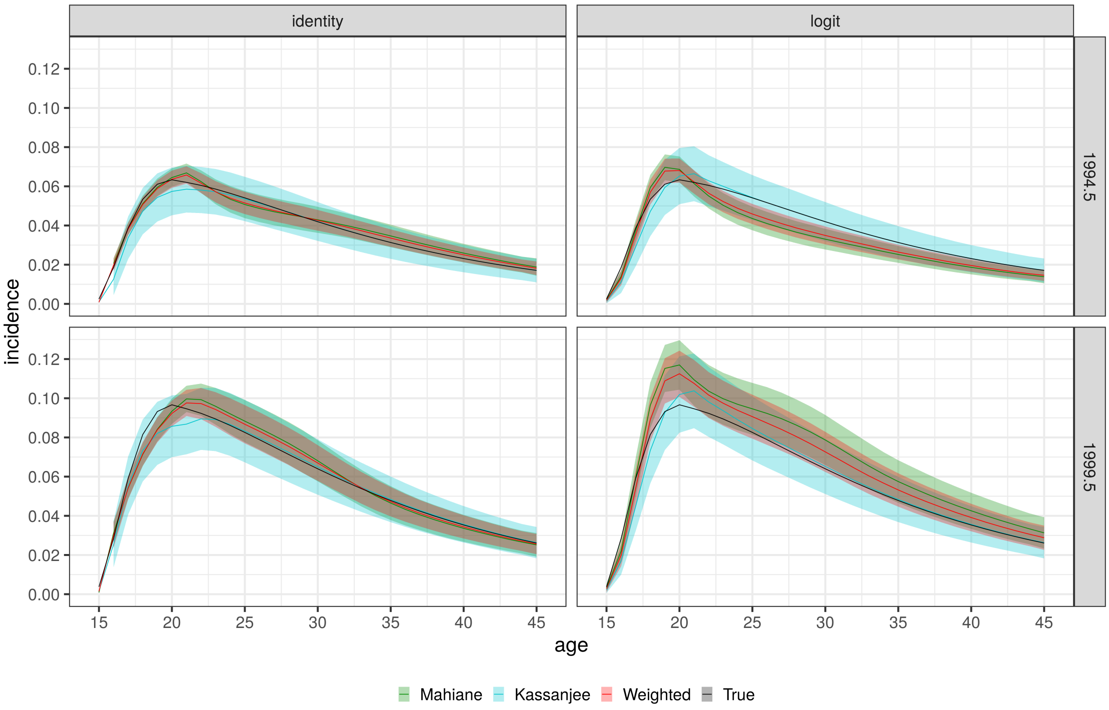

Figure 7.5: *Incidence estimates at the simulated survey dates rapidly rising epidemic (1994.5 and 1999.5).* The incidence estimates are derived from fitting one model to two cross sectional surveys simulated in an epidemic stage with an incidence function that is rapidly rising in time (1994.5, 1999.5). Each survey has a sample size of 4000/5 year age bin. The fit was done using a cubic polynomial with an inclusion distance of 6. Each row depicts the link function (logit vs identity) used for the fitting P. R is fitted using a clog-log link function.

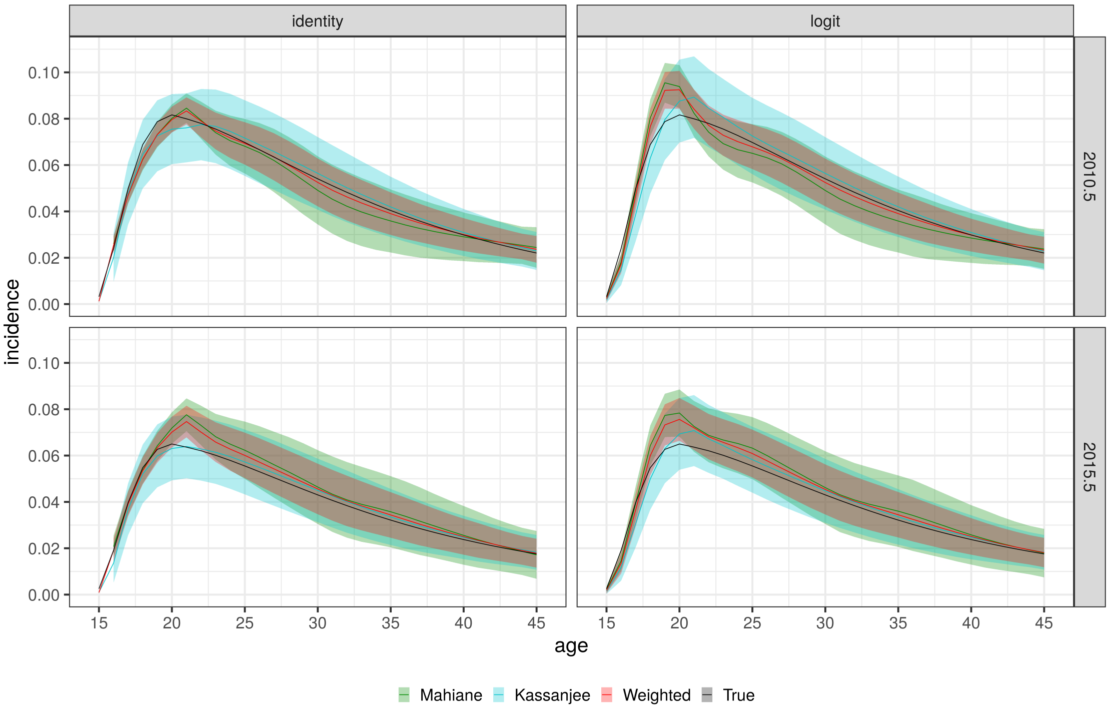

Figure 7.6: *Incidence estimates at the simulated survey dates steadily declining epidemic (2010.5, 2015.5).* The incidence estimates are derived from fitting one model to two cross sectional surveys simulated in an epidemic stage with an incidence function that is steadily declining in time (2010.5, 2015.5). Each survey has a sample size of 4000/5 year age bin. The fit was done using a cubic polynomial with an inclusion distance of 6. Each row depicts the link function (logit vs identity) used for the fitting P. R is fitted using a clog-log link function.

Figure 7.6 and 7.5 show incidence estimates at the cross sectional survey dates, derived from combining two cross sectional surveys. The cross sectional surveys are simulated from particular epidemic stages: either an increasing incidence (between 1994.5 and 1999.5) or a declining incidence (between 2010.5 and 2015.5). For comparison, we once more show the use of both an identity and a logit link function for fitting prevalence.

Incidence estimates (I_M) from the survey dates are more precise compared to the midpoint incidence estimates, in Figure 7.2 and 7.3, probably because incidence is being estimated where the data points are, unlike the midpoint incidence estimates. But this comes at the cost of accuracy - the incidence estimates (I_M) are biased at the cross sectional survey dates due to the challenges of estimating the gradient of prevalence away from the mid time of the data set. Note: just one model is fitted simultaneously to both cross sectional survey datasets (which is not the conventional use for recency data); and I_M is fundamentally designed to estimate the midpoint incidence and not the incidence at the cross sectional survey dates.

7.4.6 Incidence Trends

Two Surveys

Our attempts to estimate incidence trends/difference from two cross sectional surveys, using all 3 approaches I_K, I_M and I_{Opt} are shown in Figure 7.7. Apparently, estimating incidence differences using the Mahiane et al approach requires luck, as it is mostly biased even if they are precise, while incidence difference estimates from I_K are unbiased if not highly informative. It would seem that all the usable information is in the Kassanjee estimate, and a variance minimising I_{Opt} is not necessarily of any additional value, given the exposure to substantial bias.

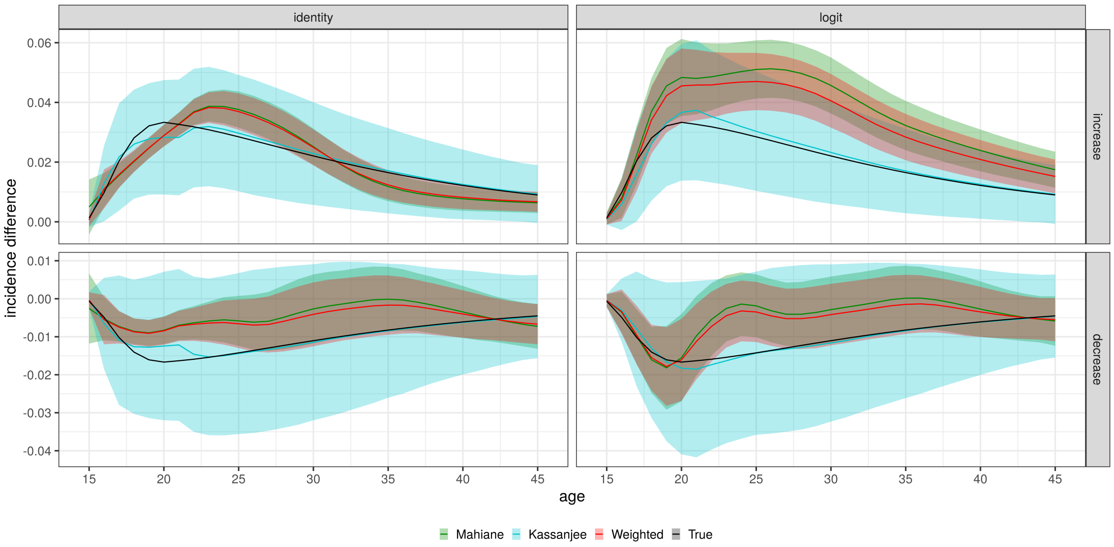

Figure 7.7: *Incidence difference estimate for pairs of surveys simulated (1993, 1998) - a rapid rise and (2010, 2015) - decline.* The plot depicts the incidence difference estimates from two pairs of cross sectional surveys each depicting a particular epidemic stage. Each survey has a sample size of 24000 (4000/5 year age bin) either logit /identity link functions (columns) are used to estimate$P$ and clog log link functions for R.

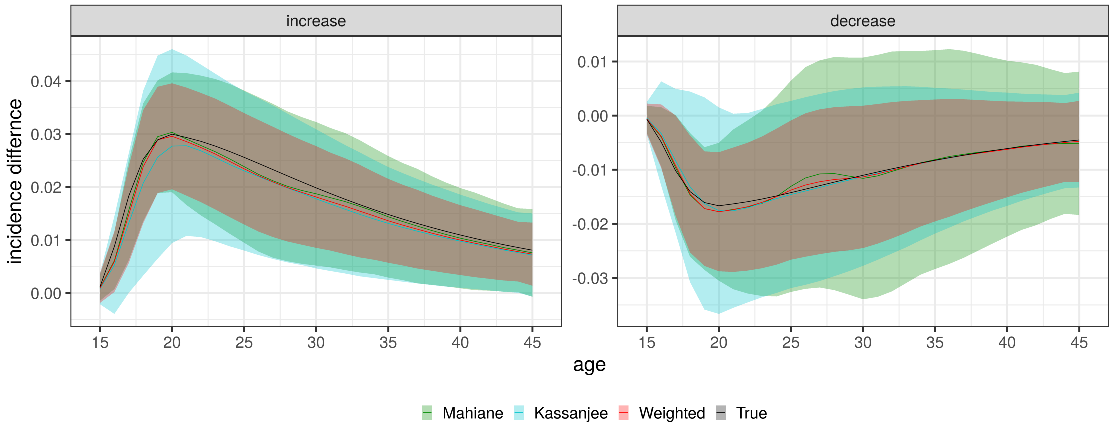

Figure 7.8: *Incidence difference estimates from midpoint incidence estimates of three cross sectional surveys.* The plot shows the incidence difference estimates from two epidemic stages - rapid increase (1993, 1998, and 2003) and rapid decrease (2005, 2010, and 2015). The incidence estimates are calculated from the midpoint of the two consecutive surveys and consequently the difference between the two incidence estimates is calculated. P and R were fitted using a logit and clog log link functions, respectively. The 95% range is estimated through 10000 bootstrap samples

Three Surveys

Figure 7.8 shows incidence difference estimates, based on 3 cross sectional surveys when incidence is steadily rising (1993, 1998, and 2003) and also when incidence is in steady decline (2005, 2010, and 2015). As expected, both the primary approaches (I_K and I_M) yield accurate incidence difference estimates that closely track the incidence difference at all ages, though they are uninformative, in turn, at various ages. Once again, the additional effort of obtaining recency data mainly improves the estimates at older ages.

We can improve the precision of the incidence difference estimates by adding the posthoc age averaging (see Figure 7.9), which we previously introduced in our companion piece [71, 76], based on two cross sectional survey with recency ascertainment. Figure 7.9 compares the post hoc age averaging for selected age groups to the age specific incidence difference of the central age of that age bin. Generally, the incidence difference estimates at the selected age bins are accurate and most importantly the post hoc averaging yields is significantly more informative for all methods, compared to the age specific incidence difference estimates. Note that the age-weighted I_{Opt} is consistently distinguishable from 0, but the less sophisticated estimates are not.

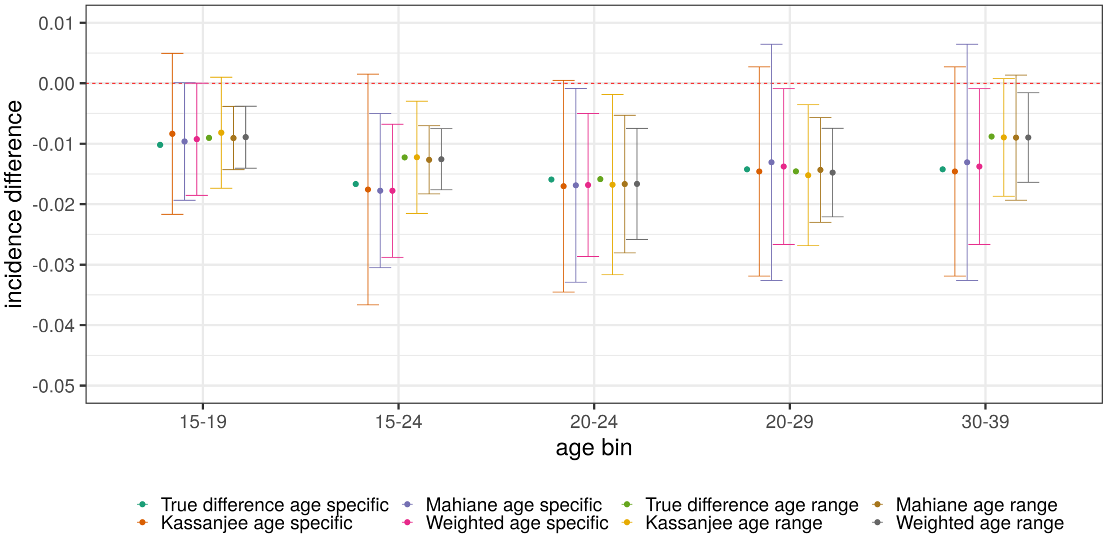

Figure 7.9: *Post hoc average incidence difference estimates from midpoint incidence estimates of three cross sectional surveys (2007.5 and 2012.5).* The plot shows the post hoc age averages from an epidemic stage on a rapid decline (2005, 2010, and 2015). The incidence difference estimates are the weighted averages (total population in the age bin from the platform) of the age specific incidence difference estimates.

7.5 Conclusion

In our preceding companion pieces, we explored the fine points to consider when estimating P, $\frac{dP}{dt}$, and R for use in each of the incidence estimators Mahiane [3] and Kassanjee [4]. This present work explores the benefits of combining I_K and I_M into a (variance) optimised weighted average. We have done this primarily from the point of view of asking what additional benefit is obtained in having the recency data.

With the additional insights gained from the present work, we now regard it as a straightforward matter to implement contextually adapted versions of a well-defined stable approach that consistently yields near-optimal extraction of HIV incidence estimates, based on whatever data is available from substantial population-based surveys of the kind which are being performed on a large scale in the heavily HIV affected countries of sub Saharan Africa.

The question of whether to expend resources on adding recency ascertainment to large population based surveys presents us with a difficult quandary. In general, reliable informative incidence estimation requires very large sample sizes (i.e. very high sampling densities across some age range) and works best when incidence is very high. This, coupled with the epidemiological/sociological importance of incidence among the young, suggests, as we have previously noted [71, 76], that one consider focusing on this group as an informative and important sentinel population, rather than attempting to obtain incidence estimates for all ages, which may simply not be feasible. For these younger ages, recency ascertainment does not really improve single time point estimates. However, we are usually even more interested in incidence differences and trends, than in single estimates, and we have seen that difference estimates based on just two survey rounds are not stable without recency data. By the time one has three rounds of major household surveys, and is in a position to obtain a robust incidence difference estimate without recency data, the better part of a decade will usually have elapsed from the first survey, and the incidence difference estimate will refer to a trend that was applicable to the epidemic some years in the past.

These considerations suggest that before embarking on a multi-year high budget commitment to one or more major surveys with intent to estimate HIV incidence, it is worth investigating the specific situation by means of carefully adapted simulations in which various designs can be simulated, and the specific analysis for burning epidemiological questions can be explored. For example, one may consider surveying just young women (age 15-30, for example) and pursuing the headline estimate of mean incidence in the age group 20-25. Recent infection testing will not yield impressive incidence estimates from one survey round, but without recency testing, there will be very little evidence on incidence changes even after two surveys - at which point the mean incidence estimate over this time will be largely driven by a Mahiane analysis.

There are other detailed loose ends we have not systematically investigated, such as:

- The impact of non-zero values for false recent rate. While it is fashionable among some analysts to presume that FRR is always zero - this is not a safe bet, and there should always at least be a sensitivity analysis on this point.
- When there are multiple surveys which each perform some sort of recent infection testing, it is not obvious that the MDRI and FRR of the test or tests should be taken as having precisely the same value in each survey round. In practice, the best estimates of these test properties may be weakly or strongly correlated, depending on whether the difference is primarily one of choice of assay or epidemic context.

These kinds of additional considerations are not just minor points, and they may warrant very careful investigation in some variation of the analyses we have been describing. Fortunately, the simulation and analysis code we have developed for our present purposes, which is available upon request, can be flexibly and straightforwardly used to adapt the analyses we have presented to many finely specified alternative scenarios.

7.A Appendix: Alternative Derivation for the Covariance of Mahiane and Kassanjee Estimators

Given two normal random variables X and Y such that they can be presented as $X = r_1 \cdot \sigma_X$ and $Y = r_1 \cdot \sigma_1^{(y)} + r_2 \cdot \sigma_2^{(y)}$. By definition the expectation $E(x,y)$ is given by;

$$
\begin{aligned}
\mathrm{Cov}(x,y) &= E(x,y) \\
\mathrm{Cov}(x,y) &= \int_0^\infty \int_0^\infty x \cdot y \cdot P(x) \cdot P(y) dy dx \\
&= \int_0^\infty \int_0^\infty (r_1 \cdot \sigma_1^{(x)}) \cdot (r_1 \cdot \sigma_1^{(y)} + r_2 \cdot \sigma_2^{(y)}) \cdot P(r_1) \cdot P(r_2) dr_1 dr_2 \\
&= \int_0^\infty \int_0^\infty [r_1^2 \sigma_1^{(x)} \sigma_1^{(y)} + r_1 r_2 \sigma_1^{(y)} \sigma_2^{(y)}] \cdot P(r_1) \cdot P(r_2) dr_1 dr_2 \\
&= \int_0^\infty \int_0^\infty r_1^2 \sigma_1^{(x)} \sigma_1^{(y)} \cdot P(r_1) \cdot P(r_2) dr_1 dr_2 + \int_0^\infty \int_0^\infty r_1 r_2 \sigma_1^{(y)} \sigma_2^{(y)}] \cdot P(r_1) \cdot P(r_2) dr_1 dr_2
\end{aligned}
\tag{7.A.1}
$$

Based on the knowledge that r_1 and r_2 are standard normal random variables, it follows that $E(r) = \int_0^\infty r \cdot P(r) dr = 0$ and $E(r^2) = \int_0^\infty r^2 \cdot P(r) dr = 1$ and consequently,

$$
\begin{aligned}
\mathrm{Cov}(x,y) &= \sigma_1^{(x)} \sigma_1^{(y)} \int_0^\infty \int_0^\infty r_1^2 \cdot P(r_1) \cdot P(r_2) dr_1 dr_2 + \underbrace{\int_0^\infty \int_0^\infty r_1 r_2 \sigma_1^{(y)} \sigma_2^{(y)}] \cdot P(r_1) \cdot P(r_2) dr_1 dr_2}_{0} \\
&= \sigma_1^{(x)} \sigma_1^{(y)} \underbrace{\int_0^\infty \int_0^\infty r_1^2 \cdot P(r_1) \cdot P(r_2) dr_1 dr_2}_{1} \\
&= \sigma_1^{(x)} \sigma_1^{(y)}
\end{aligned}
\tag{7.A.2}
$$

Given the functional forms of the Mahiane (I_M) and Kassanjee (I_K) Equation 7.A.3, we seek to find the covariance of the two estimators.

$$I_M = \frac{1}{1-P} \cdot \frac{dP}{dt} + M \cdot P \tag{7.A.3}$$

$$I_K = \frac{P \cdot (R - \beta)}{(1-P) \cdot (\Omega - \beta \cdot T)} \tag{7.A.4}$$

We adapt the exposition in Equation 7.A.1 and Equation 7.A.2 where we show that when two functions x and y share a variable then the $\mathrm{Cov}(x,y)$ is given by $\sigma_1^{(x)}\sigma_1^{(y)}$. We define P, $\frac{dP}{dt}$, and R as a random variables given by

$$P = \mu_P + r_1 \cdot \sigma_P$$

$$\frac{dP}{dt} = \mu_{\frac{dP}{dt}} + r_2 \cdot \sigma_{\frac{dP}{dt}}$$

$$R = \mu_R + r_3 \cdot \sigma_R$$

Substituting P, $\frac{dP}{dt}$, and R into Equations 7.A.3 and Equation 7.A.4 yields Equations 7.A.5 and 7.A.6, below;

$$I_M = (1 - (\mu_P + r_1 \cdot \sigma_P))^{-1} \cdot (\mu_{\frac{dP}{dt}} + r_2 \cdot \sigma_{\frac{dP}{dt}}) + M \cdot (\mu_P + r_1 \cdot \sigma_P) \tag{7.A.5}$$

$$I_K = \frac{(\mu_P + r_1 \cdot \sigma_P)((\mu_R + r_3 \cdot \sigma_R) - \beta)}{(1 - (\mu_P + r_1 \cdot \sigma_P)) \cdot (\Omega - \beta \cdot T)} \tag{7.A.6}$$

Rearranging I_M in 7.A.1 and factoring out $\frac{\mu_{\frac{dP}{dt}}}{1-\mu_P} + M \cdot \mu_P$ which is the mean of I_M, we have,

$$I_M = \left(\frac{\mu_{\frac{dP}{dt}}}{1-\mu_P} + M \cdot \mu_P\right) \cdot \left[\left(1 - \frac{r_1 \cdot \sigma_P}{1-\mu_P}\right)^{-1} \cdot \left(1 + \frac{r_2 \cdot \sigma_{\frac{dP}{dt}}}{\mu_{\frac{dP}{dt}}]}\right) + \left(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}\right)\right] \tag{7.A.7}$$

Using the approximation $(1-x)^{-1} = 1 + x + O(\Delta^2)$ gives us Equation 7.A.8 below;

$$I_M = \left(\frac{\mu_{\frac{dP}{dt}}}{1-\mu_P} + M \cdot \mu_P\right) \cdot \left[\left(1 + \frac{r_1 \cdot \sigma_P}{1-\mu_P} + O(\Delta^2)\right) \cdot \left(1 + \frac{r_2 \cdot \sigma_{\frac{dP}{dt}}}{\mu_{\frac{dP}{dt}}]}\right) + \left(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}\right)\right] \tag{7.A.8}$$

Expanding and simplifying, and ignoring the higher order terms and treating $\frac{dP}{dt}$ as a constant the second term of Equation 7.A.8 yields,

$$\begin{aligned} I_M &= \left(\frac{\mu_{\frac{dP}{dt}}}{1-\mu_P} + M \cdot \mu_P\right) \cdot \left[\left(1 + \frac{r_1 \cdot \sigma_P}{1-\mu_P}\right) + \left(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}\right)\right] \\ &= \frac{\mu_{\frac{dP}{dt}}}{1-\mu_P} \cdot \left(1 + \frac{r_1 \cdot \sigma_P}{1-\mu_P}\right) + M \cdot \mu_P \cdot \left(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}\right) \end{aligned} \tag{7.A.9}$$

Simplification of the last expression in Equation 7.A.9 yields

$$I_M = \underbrace{\frac{\mu_{\frac{dP}{dt}}}{1-\mu_P} + M \cdot \mu_P}_{I_M \text{ estimate}} + \underbrace{\left[\frac{\mu_{\frac{dP}{dt}}}{(1-\mu_P)^2} + M\right] \sigma_P \cdot}_{\text{coefficient of } r_1 \text{ which we equate to } \sigma_1} \quad r_1 \tag{7.A.10}$$

Applying the same sequence of steps we can show that I_K is given by

$$\begin{aligned} I_K &= \frac{P(R-\beta)}{(1-P)\cdot(\Omega-\beta\cdot T)} \\ &= \frac{(\mu_P + r_1 \cdot \sigma_P)((\mu_R + r_3 \cdot \sigma_R) - \beta)}{(1-(\mu_P + r_1 \cdot \sigma_P)) \cdot (\Omega - \beta \cdot T)} \\ &= \frac{(\mu_P \cdot (\mu_R - \beta))(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}) \cdot (1 + \frac{r_3 \sigma_R}{\mu_R})}{(\Omega - \beta \cdot T) \cdot (1-\mu_P)(1 - \frac{r_1 \sigma_P}{1-\mu_P})} \\ &= \left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)(\Omega - \beta \cdot T)}\right] \left[\frac{\left(1 + \frac{r_3 \sigma_R}{\mu_R}\right) \cdot \left(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}\right)}{\left(1 + \frac{r_1 \cdot \sigma_P}{1-\mu_P}\right)}\right] \end{aligned}$$

Using the linear approximation we have;

$$= \left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)(\Omega - \beta \cdot T)}\right] \left[\left(1 + \frac{r_3 \sigma_R}{\mu_R}\right) \cdot \left(1 + \frac{r_1 \cdot \sigma_P}{\mu_P}\right) \left(1 + \frac{r_1 \cdot \sigma_P}{1-\mu_P} + O(\Delta^2)\right)\right]$$

ignoring the higher order terms and simplifying;

$$= \underbrace{\left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)(\Omega - \beta \cdot T)}\right]}_{I_K \text{ estimate}} \left[1 + \frac{r_1 \sigma_P}{\mu_P} + \frac{r_1 \sigma_P}{1-\mu_P} + \ldots\right]$$

$$= \left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)(\Omega - \beta \cdot T)}\right] + \underbrace{\left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)^2 \cdot (\Omega - \beta \cdot T)} + \frac{\mu_R - \beta}{(1-\mu_P)(\Omega - \beta \cdot T)}\right] \cdot \sigma_P \, r_1}_{\text{coefficient of } r_1 \text{ in } I_K \text{ which we equate to } \sigma_{2A}}$$

Based on the exposition that results in Equation 7.A.2 we know that $\mathrm{Cov}(x,y) = \sigma_1^{(x)} \sigma_1^{(y)}$, therefore it follows that the covariance of I_K and I_M is,

$$
\begin{aligned}
\text{Cov}(I_M, I_K) &= \sigma_1 \cdot \sigma_{2A} \\
&= \left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)^2 \cdot (\Omega - \beta \cdot T)} + \frac{\mu_R - \beta}{(1-\mu_P)(\Omega - \beta \cdot T)} \right] \cdot \sigma_P \cdot \left[\frac{\mu_{\frac{dP}{dt}}}{(1-\mu_P)^2} + M \right] \sigma_P \\
&= \left[\frac{\mu_P(\mu_R - \beta)}{(1-\mu_P)^2 \cdot (\Omega - \beta \cdot T)} + \frac{\mu_R - \beta}{(1-\mu_P)(\Omega - \beta \cdot T)} \right] \cdot \left[\frac{\mu_{\frac{dP}{dt}}}{(1-\mu_P)^2} + M \right] \cdot \sigma_P^2
\end{aligned}
$$

This hold under the assumption that the covariance in I_K and I_M is due to the prevalence (P) which is a shared input parameter. The formula derived here is the same as in Equation 7.3.7 derived from the error propagation formula presented in Ku et al. [72] i.e., $\mu_P = P$, $\mu_R = R$, and $\mu_{\frac{dP}{dt}} = \frac{dP}{dt}$.

Chapter 8

Spin off - Covid prevalence smoothing

Introduction

While work was ongoing to develop the analyses which make up the bulk of this thesis, the Covid-19 epidemic presented an opportunity to engage in some prevalence/incidence surveillance applications. We collaborated with SANBS, and used some of the code developed for HIV prevalence smoothing. We investigated the prospects of estimating SARS-CoV-2 infection incidence by using time as a predictor for serostatus - but it turns out that one would need even more data than we had, given the need to stratify by province and race, even if trying to fit a single 'relative prevalence growth rate' parameter. The two primary sections of this chapter reproduce, verbatim, two preprints which were generated from this collaboration with the national blood services. These two preprints Sykes et al. [77], Mhlanga et al. [78] contributed significantly to the national debate on the state of the COVID-19 epidemic, and a unified manuscript with all this work is under preparation and will soon be submitted to Transfusion Journal.

The candidate did the statistical analysis in both the preprints and presented the work at Virtual Conference on Retroviruses and Opportunistic Infections (2021) Virtual CROI (2021).

8.1 Prevalence of anti-SARS-CoV-2 antibodies among blood donors in South Africa during the period January-May 2021.

8.1.1 Abstract

Background

Population-level estimates of the prevalence of anti-SARS-CoV-2 antibody positivity (seroprevalence) are crucial epidemiological indicators for tracking the Covid-19 epidemic. Such data are in short supply, both internationally and in South Africa. The South African blood services (the South African National Blood Service, SANBS and the Western Cape Blood Service, WCBS) are coordinating nationwide surveillance of blood donors.

Methods

Leveraging existing arrangements, SANBS human research ethics committee permission was obtained to test blood donations collected on predefined days (in January and May 2021) for anti-SARS-CoV-2 antibodies, using the Roche Elecsys Anti-SARS-CoV-2 assay on the cobas e411 and e801 platforms currently available in the blood services' donation testing laboratories. Using standard methods, prevalence analysis was done by province, age, time, sex and race.

Results

We report on data from 16762 donations. Prevalence varied substantially across race groups and between provinces, with seroprevalence among Black donors consistently several times higher than among White donors, with the other main population groups (Coloured and Asian) not well represented in all provinces. There is no clear evidence that seroprevalence among donors varies by age or sex. The weighted national estimate of prevalence (in the core age range 15-69 years) is 47.4% (95% CI 46.2-48.6). From January to May, we noted a slight but statistically insignificant increase in seroprevalence in

This section is available as a preprint: authors: Vermeulen M, Mhlanga L, Sykes W, Coleman C, Pietersen N, Cable R, Swanevelder R, Glatt TN, Grebe E, Welte A, van den Berg K., Title: Prevalence of anti-SARS-CoV-2 antibodies among blood donors in South Africa during the period January-May 2021., DOI: 10.21203/rs.3.rs-690372/v2 those provinces (Gauteng and Free State) where sufficient data were available to make such an estimate.

Conclusions

Our study demonstrates substantial differences in dissemination of SARS-CoV-2 infection between different race groups and provinces, in patterns consistent with known differences in historically entrenched socio-economic status and housing conditions. As has been seen in other contexts, even such high seroprevalence does not guarantee population-level immunity against new outbreaks, as evidenced by a substantial third wave that has emerged almost contemporaneously with the end of sampling in this study. The relative importance of various contributions to this resurgence (notably viral evolution, waning of antibody neutralization efficacy, and infection control fatigue) are unclear. Despite its limitations, notably a 'healthy donor' effect and the possible waning of detectable antibodies over the time scale of the COVID-19 pandemic, it seems plausible that these estimates are reasonably generalisable to actual population level antiSARS-CoV-2 seroprevalence. The interpretation of occasional seroprevalence surveys as a proxy for total attack rates, over the ever-lengthening pandemic time scale is likely to become ever more complex. More frequent sampling, including linked repeat observations of frequent donors, could substantially improve the utility of blood donor surveillance.

8.1.2 COVID-19 Seroprevalence in South Africa.

Coronavirus disease 2019 (Covid-19) caused by the virus SARS-CoV-2, manifests in a plethora and range of symptoms, varying from asymptomatic to severe disease which may lead to death. It is this range of severity as well as limited access to health care that makes it difficult to determine how many people have been infected with the virus. After contracting SARS-CoV-2, the majority of people will develop antibodies as part of their immune response. These antibodies last from between 6 and 12 months and can therefore provide an indication of the number of people who have been infected during that time. Given the substantial uncertainties around the true counts of cases of SARS-CoV-2 infection, and prior studies indicating that in many settings the confirmed case count is only a small proportion of all laboratory confirmed infections, it is of ongoing importance to obtain credible estimates of the prevalence of anti-SARS-CoV-2 antibody positivity (seroprevalence), at the community level [79, 80].

8.1.3 Method

The South African National Blood Service (SANBS, serving 8 of 9 provinces in South Africa) and Western Cape Blood Service (WCBS, servicing the Western Cape) obtained ethics clearance from the SANBS Human Research Ethics Committee to perform a SARS-CoV-2 seroprevalence study among South African blood donors. The protocol allowed for the testing of routinely collected donor screening samples on predefined 'collection days' in January, March and May; which were internally communicated to blood centre staff at participating collection sites, but without prior notice to potential donors. All donors underwent routine screening through a self-administered questionnaire, one-on-one assessment and a mini-health screening by blood centre staff. Donors who did not meet the routine donor eligibility criteria were excluded from donation and therefore from the study. Contact with persons infected by COVID-19, unresolved COVID-19 infection or COVID-19-like symptoms in the preceding 14 days resulted in temporary deferral of potential donors

Samples collected at the time of donation were tested for anti-SARS-CoV-2 antibodies, using the Roche Elecsys Anti-SARS-CoV-2 total immunoglobulin nucleocapsid assay on the cobas e411 and e801 platforms already in use at the blood services. This assay, according to the package insert, has diagnostic specificity in excess of 99.5%, and near perfect sensitivity (point estimate of 100%) at 16 days post PCR positivity. It detects only anti-nucleocapsid antibodies, and so does not detect antibodies mounted in response to any of the vaccines in use, which only present (and stimulate production of antibodies against) viral spike proteins. We do not here explore various nuances of how to define and estimate test performance characteristics by distribution of cases (defined primarily by severity of infection and time since infection/symptoms/PCR detection), but we note:

- Sensitivity and specificity 'in our hands' was investigated by testing 618 samples from the pre-COVID-19 era (1 marginal false positive precisely at the diagnostic threshold) and 50 samples confirmed as positive in a COVID-19 convalescent plasma study protocol (with 1 false negative).

- For epidemiological interpretation, we take seroprevalence as a close proxy of the prevalence of having been infected with SARS-CoV-2 at some point. The Elecsys

Anti-SARS-CoV-2 assay appears to have particularly good durability of antibody detection for months post PCR reversion and symptom resolution, with no evidence of antibody waning and seroreversion over more than four months in a US COVID-19 convalescent plasma cohort [81].

- We ignore, for now, the effects of 1) the donor deferral rule that people with confirmed SARS-CoV-2 infection, or COVID-19-like symptoms, are precluded from donation for a period of two weeks after PCR test and/or symptom resolution, and 2) deferral of regular donors who were in quarantine due to a positive contact, and who therefore skipped their routine donation. Given the high rate of asymptomatic infection, this is a relatively minor limitation.

We did not perform structured sampling in the sense of selecting a subset of donation sites or regions within a province. The study merely observed all consenting donors who happened to present themselves at any donation facility on collection days.

Prevalence was estimated by typical categorical and continuous predictors (age, sex, race and province) by standard methods, using the **R** platform for statistical computation. Although we are not aware of any biological basis for expecting racial differences-in South Africa, as elsewhere, race is, for historical reasons, a strong correlate of socio-economic status, living conditions, and social circumstances, and therefore a suspected predictor of prevalence. As freely downloadable data sets from Statistics South Africa do not disaggregate sufficiently for our purposes, our provincial weighted seroprevalence estimates are based on population size estimates from Machemedze et al [82], interpolated to March 2021, and a racial breakdown of provinces as observed in the 2011 census [83]. The level of (dis)aggregation for headline estimates was chosen based on the results of exploratory analysis, as reported below.

Each province was sampled primarily in either January or May, with only Gauteng (GP) and Free State (FS) having a statistically meaningful number of specimens from another month (GP-January, FS-May). To understand the time dimension in our data, we performed a regression in which the data for White and Black donors, from the FS and from GP, was fitted to a model that assigns each of the four subgroups their own prevalence, but with an exponential time dependence that is governed by a single universal rate shared by both provinces and race groups.

8.1.4 Results

The demographic breakdown of the sampled donors is displayed in Figure 8.1. There were slightly more male donors (51.2%). The large majority of donors in our study were White (51.4%) and Black (35.6%) with the remainder distributed mainly between donors self-identifying as Asian (4.2%) and 'Coloured' (8.1%)- a uniquely South African racial label indicating persons with a significant mix of ancestry from, amongst other lineages, South Asia, Indonesia, Southern Africa and Europe [84]. Only 0.8% did not report a racial identification.

Table 8.1: The demographic breakdown of seroprevalence specimens by anti-SARS-CoV-2 antibodies reactivity.

	Reactive	**Non-Reactive**	**Total**	**(%) Reactive**
Sex				
Female	2567 (50.4%)	5611 (48.1%)	8178 (48.8%)	31.4
Male	2527 (49.6%)	6057 (51.9%)	8584 (51.2%)	29.4
Province				
Eastern Cape	569 (11.2%)	896 (7.7%)	1465 (8.7%)	38.8
Free State	289 (5.7%)	793 (6.8%)	1082 (6.5%)	26.7
Gauteng	1988 (39%)	4216 (36.1%)	6204 (37%)	32.0
KwaZulu Natal	663 (13%)	1444 (12.4%)	2107 (12.6%)	31.5
Limpopo	217 (4.3%)	494 (4.2%)	711 (4.2%)	30.5
Mpumalanga	563 (11.1%)	1132 (9.7%)	1695 (10.1%)	33.2
Northern Cape	100 (2%)	367 (3.1%)	467 (2.8%)	21.4
North West	202 (4%)	530 (4.5%)	732 (4.4%)	27.6
Western Cape	503 (9.9%)	1796 (15.4%)	2299 (13.7%)	21.9
Race				
Asian	156 (3.1%)	539 (4.6%)	695 (4.1%)	22.4
Black African	3155 (61.9%)	2810 (24.1%)	5965 (35.6%)	52.9
Coloured	448 (8.8%)	908 (7.8%)	1356 (8.1%)	33.0
White	1288 (25.3%)	7317 (62.7%)	8605 (51.3%)	15.0
Unreported	47 (0.9%)	94 (0.8%)	141 (0.8%)	33.3
Age				
16 - 29	1789 (17.6%)	3270 (14%)	5059 (15.1%)	35.4
30 - 39	1380 (13.5%)	2312 (9.9%)	3692 (11%)	37.4
40 -49	955 (9.4%)	2364 (10.1%)	3319 (9.9%)	28.8
50+	970 (9.5%)	3722 (15.9%)	4692 (14%)	20.7
Total	5 094	11 668	16 762	30.4

Figure 8.1 shows the age distribution of donors included in the present analysis, further decomposed by race and province. The provincial totals are shown in Table 8.2.

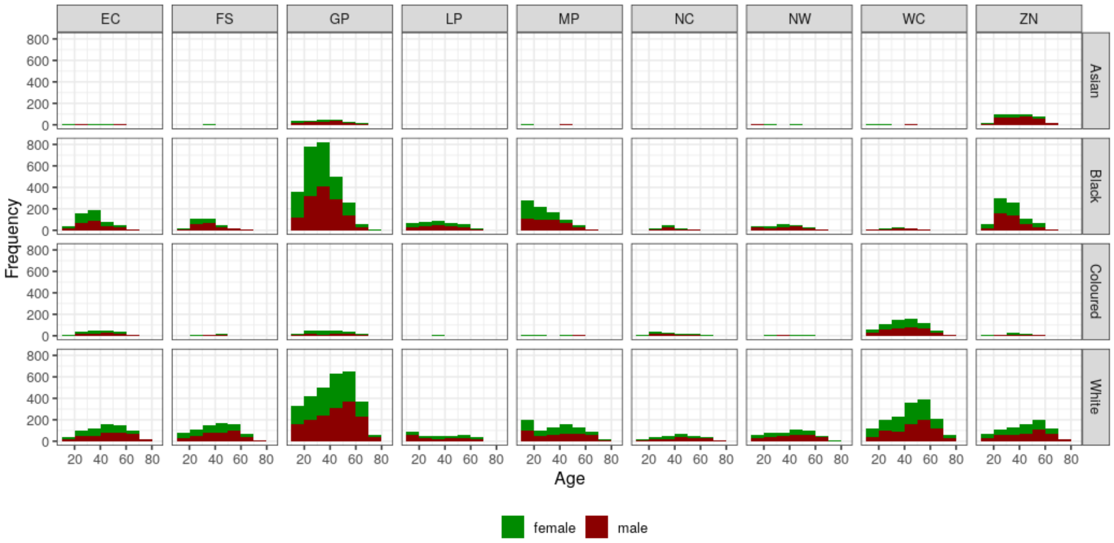

Figure 8.1: *Age and sex structure of surveyed donors*. The Age and sex structure of surveyed donors, broken down by race and province (EC-Eastern Cape, FS-Free State, GP-Gauteng, LP-Limpopo, MP-Mpumalanga, NC-Northern Cape, NW-North West, WC-Western Cape, ZN-KwaZulu Natal)

Table 8.2: *Weighted provincial estimates of prevalence and the implied number of infections.* Weighted provincial estimates of prevalence; the implied number of infections; the number of laboratory confirmed cases; and the (multiplicative) discrepancy between our estimate and the official count.

Province	**Prevalence** (%)	**Estimated Infections**	**Official Dx-Cases**	**Underestimate (Fold)**
Eastern Cape	62.5 (58.8, 65.9)	2,724,350	176,902	15.4
Free State	47.8 (42.8, 53.0)	925,093	81,622	11.3
Gauteng	43.8 (42.3, 45.4)	4,926,044	434,495	11.3
Limpopo	46.3 (41.3, 51.2)	1,687,558	64,966	26.0
Mpumalanga	47.6 (44.5, 50.8)	1,523,296	81,758	18.6
Northern Cape	31.8 (25.7, 38.0)	235,156	25,007	9.4
Northwest	48.5 (42.5, 54.6)	1,302,318	69,328	18.8
Western Cape	37.4 (33.4, 41.4)	1,855,484	294,201	6.3
KwaZulu-Natal	52.1 (49.1, 55.1)	3,950,784	249,703	15.8

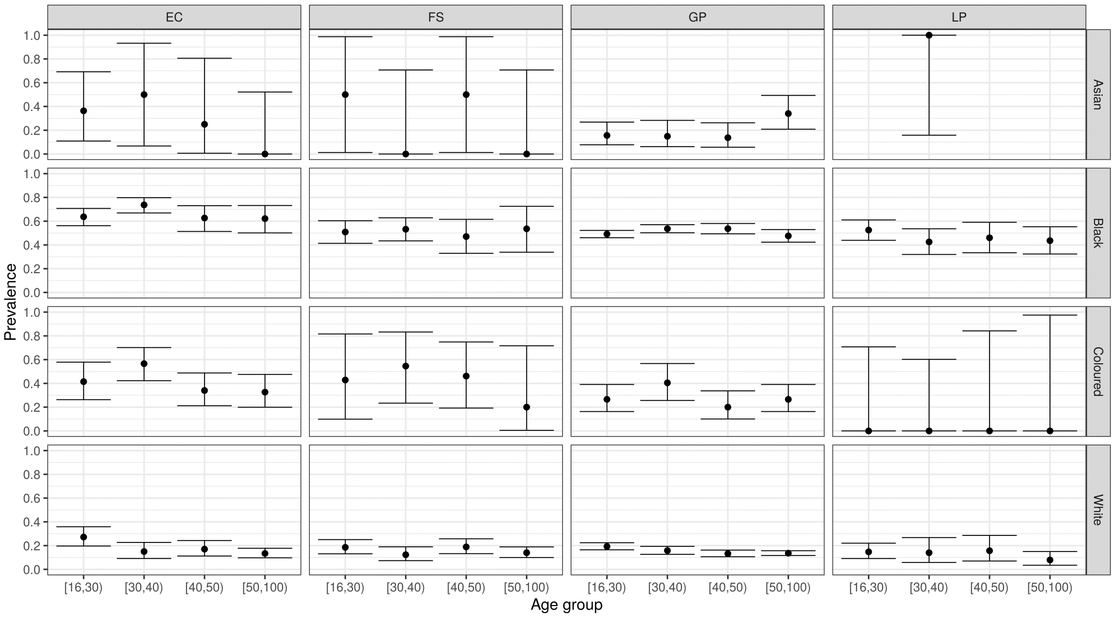

Figure 8.2: Prevalence by age group, broken down by province and race for four provinces (EC-Eastern Cape, FS-Free State, GP-Gauteng, LP-Limpopo)

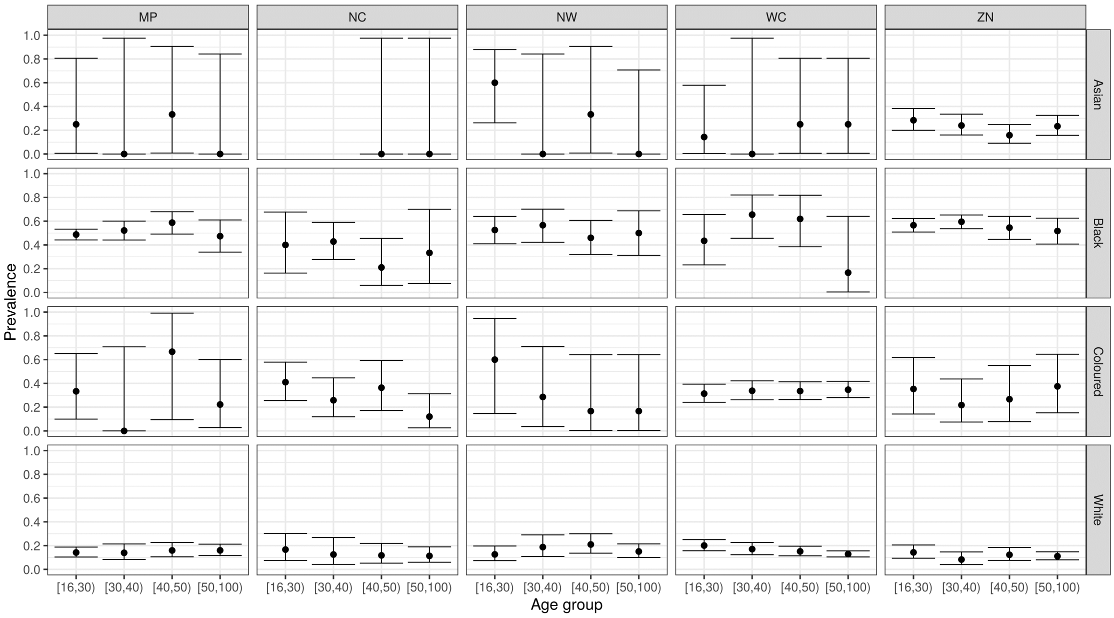

Figure 8.3: Prevalence by age group, broken down by province and race for five provinces (MP-Mpumalanga, NC-Northern Cape, NW-North West, WC-Western Cape, ZN-KwaZulu-Natal)

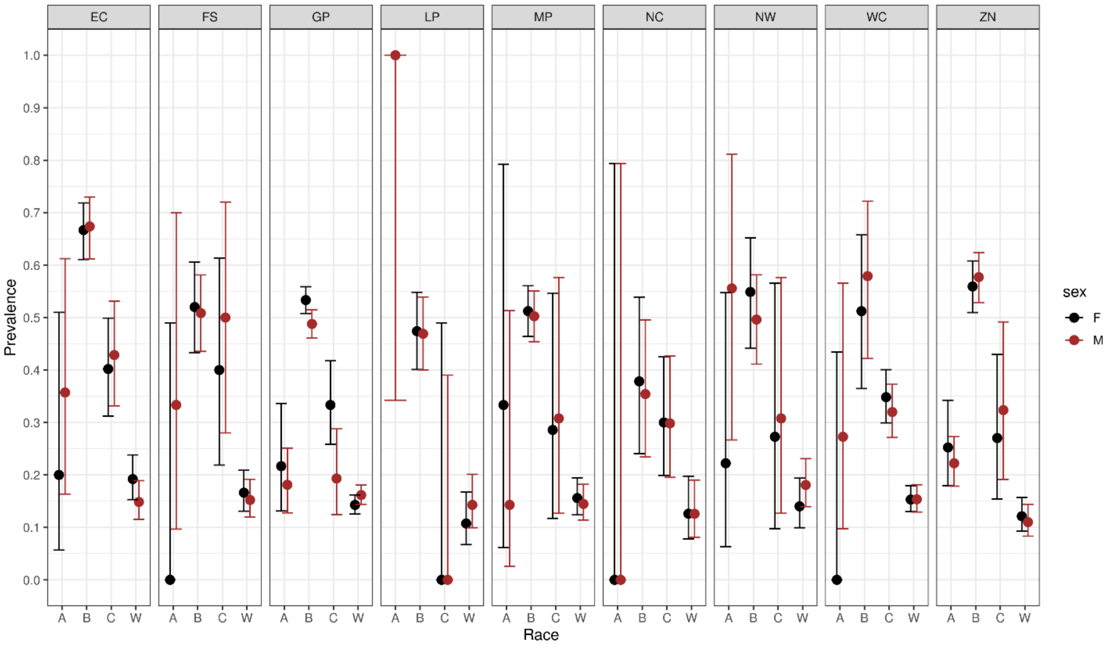

Figure 8.4: Prevalence comparison between sexes, by race and province (EC-Eastern Cape, FS-Free State, GP-Gauteng, LP-Limpopo, MP-Mpumalanga, NC-Northern Cape, NW-North West, WC-Western Cape, ZN-KwaZulu-Natal)

After categorizing by either broad or narrow age bins in all provinces and the major race groups, there was no association between seroprevalence and age (see Figure 8.2 and 8.3 for broad age bins). There was no association between seroprevalence and sex. See Figure 8.4 for disaggregation by sex, race and province. Therefore, for the remaining analysis, we do not disaggregate by either age or sex. The regression of data from GP and FS against time provided an estimate of a (relative) 1.6% per month growth in prevalence, which, at a p value of 0.3, is not statistically significant, but is large enough (and in the right direction) to be consistent with the crude growth in case detections, in the absence of seroreversion.

Figure 8.4 shows the seroprevalence estimates by the remaining meaningful disaggregation-race and province. The large differences, by both race and province, are highly statistically significant as well as epidemiologically meaningful. Note also the race-weighted overall provincial prevalence estimates (which we interpret as provincial 'attack rates'), and the official prevalence of having been diagnosed, based on reporting of positive PCR diagnostic test results, according to the National Institute for Communicable Diseases (NICD) [85, 86] in the dominant month of sampling: January for the Eastern Cape

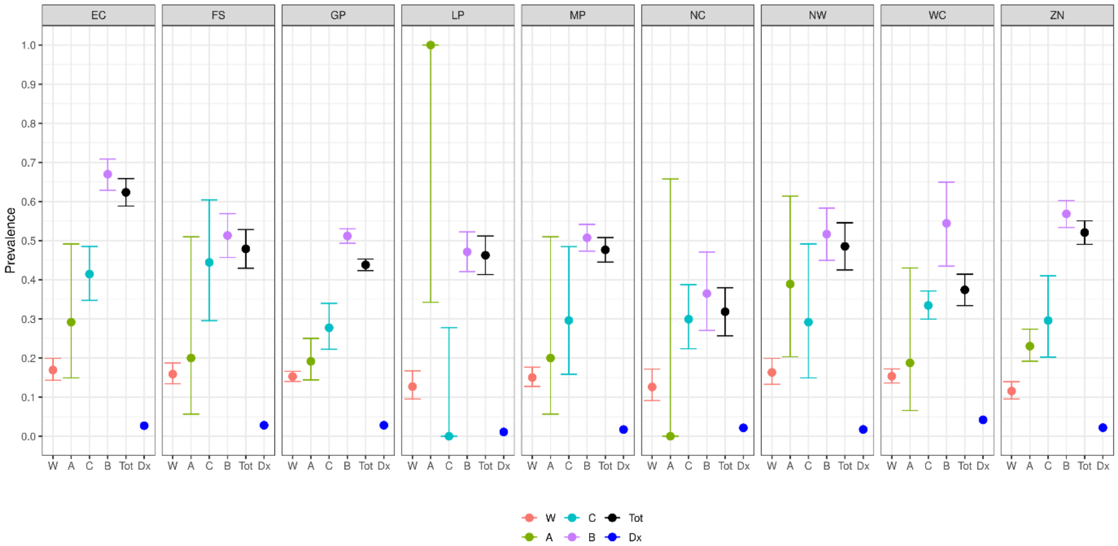

Figure 8.5: Prevalence by race (W-White, A-Asian, C-Coloured, B-Black) and province (EC-Eastern Cape, FS-Free State, GP-Gauteng, LP-Limpopo, MP-Mpumalanga, NC-Northern Cape, NW-North West, WC-Western Cape, ZN-KwaZulu-Natal), showing also the race weighted provincial estimates (Tot), and the prevalence implied by diagnosed cases reported to the National Institute for Communicable Diseases (Dx).

(EC), Free State FS, Northern Cape (NC), KwaZulu Natal (ZN); May for Gauteng (GP), Limpopo (LP), Mpumalanga (MP), Northwest (NW), Western Cape (WC). The NICD reports on testing performed both in the private and public sector.

Table 8.2 shows our provincial estimates of attack rates, as a percentage; the implied number of infections; the number of laboratory confirmed cases according to the NICD [85]; and the (multiplicative) discrepancy between our estimate and the official count. Note that our estimated number of infections is conservatively based on our estimated prevalence being applied only to the age group 15-69, so these factors are not quite as large as implied by Figure 8.5. The estimated seroprevalence ranges from 31.8% in NC to 62.5% in the EC and ranges from 6 (WC) to 26 (LP) fold higher than the official case count.

8.1.5 Discussion

Our study confirms high seroprevalence rates, particularly among Black donors, with little sign of significant population level immunity among other race groups. These substantial differences can most likely be explained by historically based socio-economic factors which hinder the implementation of COVID-19 preventative measures at a community level. The generally high levels of seroprevalence across the whole country are consistent with expectations, given the high burdens experienced on the health care system, and generally low proportion of SARS-CoV-2 infections which present as serious illness.

Previous seroprevalence estimates from South Africa, specifically the WC already found, before the second wave: 1) a very high prevalence (30-40 percent) among pregnant women attending state sector antenatal care, and people living with HIV presenting for routine viral load assessment [87]; and 2) higher prevalence among workers with lower socioeconomic status [88]. A household cohort study performed in a rural setting in Mpumalanga and an urban setting in North West province found a seroprevalence of 7% (95% CrI 5-9%) and 27% (95% CrI 23-31%), after the first wave of infection, and 26% (95% CrI 22-29%) and 41% (95% CrI 37-45%), respectively, after the second wave [89].

For an indication of the meaning of such high seroprevalence values, in a one year old epidemic, consider: a prevalence of 50%, accumulated over 50 weeks, of a condition with a duration of infectiousness of 1 week, implies an average 'prevalence of infectiousness' of 1% of the population, with inevitable significant elevations above this average value during peaks. For people reliant on public transport, or working in public spaces, it will be difficult to limit close encounters to fewer than 100 people on any given day- i.e. it will be difficult to encounter fewer than one infectious person per day.

We do not claim that blood donors are perfectly representative of the South African population. Firstly, Black and White donors each account for roughly half the total participants of this study, though South Africa's population is about 80 percent Black African and only 8 percent White/European [83]. Other population groups are generally insignificantly small except Asian in ZN (about 20%) and Coloured in the WC (about 50%). Of course, our analysis explicitly weights for racial representativeness. The age weighting we adopted to estimate total infections also produces a face value underestimate for population totals, as it assigns no cases in the age range 0-14 years, which accounts for about 30 percent of the population. Furthermore, repeat blood donors (who supply the majority of donations) are pre-selected to have recently been negative for pathogens included in routine blood safety screening. In South Africa this selection for being HIV negative is certainly relevant, given the country's extraordinary HIV prevalence. Communities which are economically stressed, or without ease of access to blood donor centres, will be under-represented among the study population.

Survey dates represented in this analysis are either:

- Barely past South Africa's 'second wave' in COVID-19 incidence- whence deferral rules based on confirmed infection or COVID-19-like symptoms should slightly depress seroprevalence estimates relative to 'true' prevalence; or
- Shortly before the emergence of the 'third wave', whence the interpretation of all these samples, as being from a fairly well-defined epidemiological stage, is not entirely unreasonable.

The Elecsys Anti-SARS-CoV-2 antibody assay appears to have particularly good detection sensitivity for months post PCR reversion [81], though there may be some seroreversion. Therefore, while further investigation of the issue of representativeness will clearly need to be done, our estimates are subject to downward bias by at least some obvious considerations.

With due consideration to both the patent and latent limitations of our study, the key observations we wish to make at this point are:

- The particularly high attack rates in majority Black communities points to the limitations, thus far, of non-pharmaceutical interventions in the context of economic deprivation and high population density, and the urgency of making vaccines available in all communities.
- The high seroprevalence (especially amongst Black donors) also raises interesting and important questions about the level of collective immunity thus far obtained through the two primary infection waves to date - but we caution against simplistic interpretations, given that substantial outbreaks have been seen in cities after the observation of very high seroprevalence [90], and more recent concerns about vaccine efficacy against new variants.
- The low seroprevalence amongst White donors suggests that predominantly White suburban communities lack meaningful collective immunity, and should take infection control measures very seriously for the foreseeable future, especially at the time of writing, when the third wave is presenting many communities with rapidly increasing incidence.

- Given the relatively low marginal cost of leveraging the infrastructure of the blood services, we are keen to further probe the representativeness of blood-donor-based seroprevalence surveys, and to see to what extent surveillance in the blood services can be a valuable and efficient ongoing activity during major infectious disease outbreaks.

Acknowledgements

The authors would like to thank the many blood donors who give of their time and lifeblood for the benefit of others, without material reward, while also consenting to the research applications of their specimens. We thank the many nurses and logistical and laboratory staff without whom there would be no transfusion service, no safety screening, and no research project. Additionally, we thank numerous colleagues who have helped us clarify our thinking about the issues involved, including ethical, operational and analytical aspects. Funding for the operational activities was provided by the South African National Blood Service and the Western Cape Blood Service in part supported by the South African Medical Research Council with funds received from the Department of Science and Innovation, as well as the ELMA South Africa Foundation (Grant number 20-ESA011). The time of investigators in the blood services is supported by the blood services. Karin van den Berg and Marion Vermeulen are also supported by the NIH Fogarty International Center training grant 1D43-TW010345. Alex Welte and Laurette Mhlanga are supported by a Centre of Excellence grant from the South African Department of Science and Innovation via the National Research Foundation. Eduard Grebe is supported by internal funding from Vitalant Research Institute, San Francisco.

8.2 SARS CoV 2 Infection Fatality Rate Estimates for South Africa.

This is a brief report. The intention is simply to communicate findings and thereby trigger discussion and further work - not to provide a substantial contextual, comparative, or interpretive narrative.

It is of course important to ascertain the risk of death that comes with SARS-CoV-2 infection (the 'Infection Fatality Rate/Ratio' or IFR) but it is difficult to observe directly, as the majority of infections go undiagnosed. Using a positive clinical diagnosis as the defining element of a 'case', leads to the related 'Case Fatality Rate', or CFR. While CFR is relatively simple to assess, within a study or a stable clinical record keeping or case reporting system, it is less clear what it means, and its meaning will inevitably vary from place to place, due to structural inequities; and from time to time, as testing systems adapt to the evolving epidemic.

There has been some speculation that African and some other developing nations have been affected less severely than developed countries, in particular seeing fewer than expected deaths from Covid-19. It has also been suggested that the paucity of data, and the very different age structures of the populations in different parts of the world, might largely explain the apparent differences seen in crude CFRs, which mask distributions of (primarily) age and other key risk factors for severe disease from SARS-CoV-2 infection.

The Blood Services in South Africa (SANBS and WCBS) have recently published national SARS-CoV-2 seroprevalence estimates based on a substantial, approximately nationally representative, sample of blood donors [91]. This analysis indicates no dependence of seroprevalence on sex or age (in the sampled age range of 16 to 80), but alas, as is typical for many health and welfare indicators in South Africa, race and province are strong predictors of seroprevalence.

The Medical Research Council has been producing weekly excess deaths estimates for some time. These are not disaggregated by either race or age. As it is well known that age is a very strong, indeed probably the strongest, predictor of severity of Covid-19, it

This section is available as a preprint: authors: Mhlanga L, Vermeulen M, Grebe E, Welte A, Title: SARS CoV 2 Infection Fatality Rate Estimates for South Africa., DOI: 10.21203/rs.3.rs-707813/v2 is largely meaningless to discuss CFR or IFR without paying attention to age.

Given that the just-published donor-based South African seroprevalence estimates, reflecting prior SARS-CoV-2 infection, vary sharply by race, and substantially by province, and that we know fatality depends strongly on age, it would be optimal to have excess deaths reported by race, precise age, and province. We understand that the vital registration system in South Africa does not report deaths by race, and that the MRC only occasionally publishes disaggregation by age. Indeed, we are aware of a single South African report on excess deaths by age [92] and even then only in decade age bands, and not simultaneously by province.

It would be a simple matter to estimate age specific IFR, nationally averaged, if the excess deaths estimates and the prevalence estimates applied to the same point in time. As it is, the published age disaggregated excess deaths estimates are as of the end of 2020, and our prevalence estimates are representative of the period of January to May 2021 - which we will interpret, for the present purposes, as an estimate applicable to late March. Incidence was not very high from January to May, as this was between the second and third wave - but the delay seen with deaths means that deaths more than doubled between December and March.

For the sake of this preliminary estimate, we rescale the December 2020 age specific excess deaths by a factor of 2.12, to obtain a cumulative, national, excess deaths estimate which has the correct total for March 2021. For provincial, age aggregated estimates, we use the provincial cumulative deaths reported by the MRC in March 2021 - not rescaled provincial estimates from December 2020.

We are choosing to interpret the reported estimated excess natural deaths as Covid-19 deaths. As far as we can tell, this could as credibly be argued to be an under- or overestimate. Some have highlighted collateral deaths of various kinds, and others have noted the reduction in other infection related deaths during lockdown periods.

In order to have a well-defined age aggregated IFR, we allocated neither cases nor deaths to the age group <10 years, and we allocated the observed (non-age dependent) prevalence from the blood donor study to all ages from 10 up. These estimates, then, are a population averaged IFRs for persons aged 10 and over.

Table 8.3 indicates our nationally aggregated, age disaggregated, infection fatality rates, which used the population estimates from [82]. These IFR estimates for are broadly comparable with previous estimates of which we are aware, such as from a locale-based study from South Africa [89] and a meta-analysis of estimates from the Global North [93]. For visual representation, we interpret the numbers from Table 8.3 as mid-decade estimates, and fit an exponential curve (see Figure 8.6). It seems reasonable to say that the relationship between age and IFR is heuristically characterisable as a doubling of fatality for every ten years of age.

Table 8.3: Age specific estimates of SARS-CoV-2 Infection Fatality Rates in South Africa, as of March 2021

Age Range	**Population Size**	**Excess Natural Deaths**	**Scaled Excess Natural Deaths**	**SARS-CoV-2 Infections**	**IFR** (%)
1-9	11,217,099	0	0	0	N/A
10-19	10,280,989	332	705	4,873,189	0.014
20-29	9,954,072	1,194	2,535	4,718,230	0.054
30-39	10,333,318	4,213	8,944	4,897,993	0.183
40-49	7,211,051	6,509	13,819	3,418,038	0.404
50-59	5,020,135	13,881	29,470	2,379,544	1.238
60-69	3,327,195	19,724	41,875	1,577,090	2.655
70-79	1,602,572	14,102	29,939	759,619	3.941
80-	725,977	11,010	23,375	344,113	6.793
Total	59,672,408	70,965	150,663	22,967,816	0.656

Table 8.4 shows the provincial age aggregated IFR estimates. The actual estimate, based on the available data, is the column ‘estimated IFR’. To better understand differences in IFR between provinces, we calculated a so called ‘expected IFR’ by province, which is what we would observe if provinces all share the national age dependent IFRs, in each case averaged over the province-specific age distribution. This way, we can compare the actual estimate with the ‘expected’ estimate, and thus not unnecessarily interpret a provincial IFR to be ‘relatively high’ simply because that province has a relatively older population. What is not clear is to what extent these various indicators reflect 1) differences in the relationship between blood donors and the provincial population which bias the provincial seroprevalence estimates in different ways, 2) differences in actual age-specific fatality between provinces, and 3) differences in the quality of death data and excess deaths estimates.

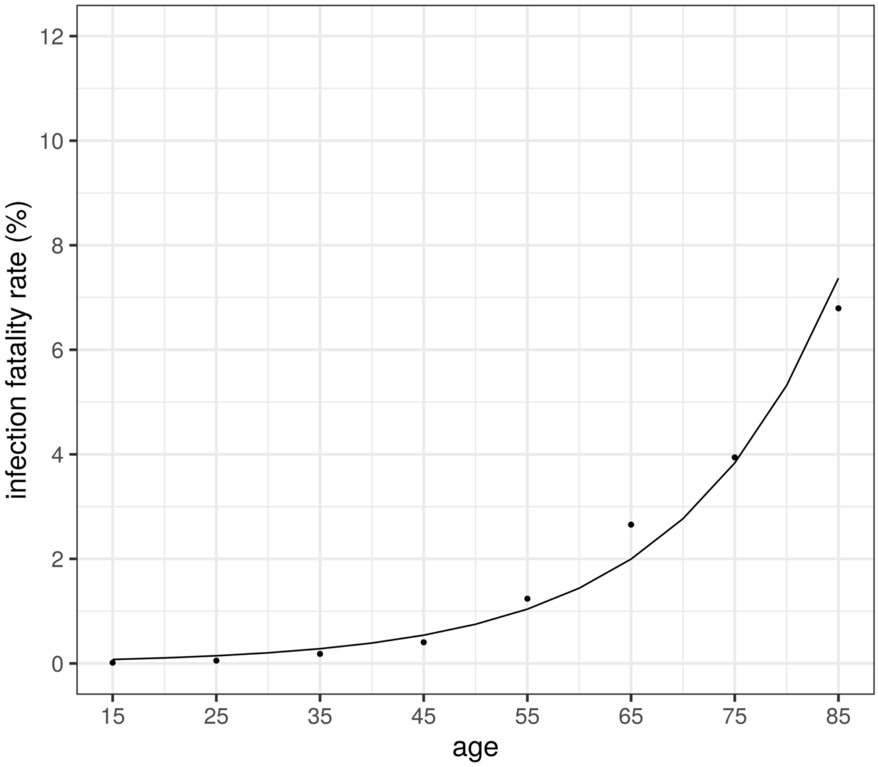

Figure 8.6: Fitted SARS-CoV-2 Infection Fatality Rate as an exponential function of age, in South Africa, as of March 2021.

Table 8.4: *South African Provincial, age aggregated, SARS-CoV-2 Infection Fatality Rates.* The 'expected IFR' (%) column indicates what the provincial IFR (%) would be if the national age specific IFR estimates apply to each province, and are adapted to the province only by age-averaging the IFR using the provincial age distribution.

Province	Population	Excess Natural Deaths	Infections	Estimated IFR	Expected IFR
Eastern Cape	5,430,323	33,900	3,392,727	0.999	0.771
Free State	2,353,101	6,884	1,077,680	0.639	0.654
Gauteng	12,907,289	24,411	5,661,984	0.431	0.615
Limpopo	4,610,507	13,731	2,132,791	0.644	0.736
Mpumalanga	3,874,435	10,617	1,846,066	0.575	0.617
Northern Cape	919,620	3,067	29,253	1.048	0.697
North West	3,255,572	5,212	1,580,355	0.330	0.641
Western Cape	5,882,400	16,179	2,200,998	0.735	0.720
KwaZulu Natal	9,222,061	36,661	4,802,063	0.763	0.583
Total	59,672,408	70,965	150663	22,967,816	0.656

In fact, we are also not sure whether deaths may end up being allocated to provinces differently than the provincial allocation of the deceased persons during life, given significant mobility of people of working age.

Our analysis is not optimal, mainly because the mortality data which one would ideally use, and which forms the basis of the routine MRC reports, is not publicly available. When this limitation is addressed, it is hard to imagine that the fatality rate estimates will change substantially. Since estimates of IFR are clearly important as part of the overall epidemiological assessment, scenario projections, and health system evaluation, we are disseminating these estimates at the present time in the hope that they will stimulate discussion and epidemiological thinking.

Acknowledgements

The authors would like to thank the many blood donors who give of their time and lifeblood for the benefit of others, without material reward, while also consenting to the research applications of their specimens. We thank the many nurses and logistical and laboratory staff without whom there would be no transfusion service, no safety screening, and no research project. Additionally, we thank numerous colleagues who have helped us clarify our thinking about the issues involved, including ethical, operational and analytical aspects. Funding for the operational activities was provided by the South African National Blood Service and the Western Cape Blood Service in part supported by the South African Medical Research Council with funds received from the Department of Science and Innovation, as well as the ELMA South Africa Foundation (Grant number 20-ESA011). The time of investigators in the blood services is supported by the blood services. Karin van den Berg and Marion Vermeulen are also supported by the NIH Fogarty International Center training grant 1D43-TW010345. Alex Welte and Laurette Mhlanga are supported by a Centre of Excellence grant from the South African Department of Science and Innovation via the National Research Foundation. Eduard Grebe is supported by internal funding from Vitalant Research Institute, San Francisco.

8.3 Concluding Remarks / Ongoing Work

The collaboration with the blood services is ongoing. This work involves support for further study design details, analysis of data from forthcoming rounds, and also the production of a manuscript [94] based on the preprints which have been reproduced in this chapter. We do plan an additional analysis, aimed at estimating the prevalence gradient and incidence (using Mahiane et al. [3] with excess mortality $= 0$).

Chapter 9

Discussion, Recommendations, and conclusions

9.1 Discussion

The main aim of the thesis is to optimize and benchmark population-level HIV incidence estimation methods. Our exploration and exposition was limited to the Mahiane et al. [3] and Kassanjee et al. [4] frameworks and further explored the optimally weighted incidence estimator (a hybrid of Mahiane and Kassanjee estimators).

We chose the Mahiane et al. [3] and Kassanjee et al. [4], estimators because they do not make unnecessary epidemiological/demographic assumptions, are easily adaptable to yield age-specific incidence estimates and have proved their robustness to give accurate and relatively informative incidence estimates. Below we broadly recap and summarise the high-level points and recommendations made in each chapter.

We used a custom simulation platform to simulate an SA-like epidemic under varied conditions to investigate method validity, applicability, and robustness. The thesis presented a custom-designed simulation platform for these and other purposes and is adaptable to similar benchmarking exercises. We gave detailed reports on our investigations, including validated conclusions and sound recommendations to smoothing population-level survey data to yield incidence estimates.

In the case of the recency incidence estimator (Kassanjee et al. [4]), we show that there are subtle points to consider if one seeks to estimate age-specific incidence;

1. A “one size fits most" is a moving window approach with higher-order polynomials and a wide enough inclusion distance (> 5) to estimate the prevalence and prevalence of recency and consequently incidence.

2. Continuous fitting of the age structure of the prevalence data yields more informative age-specific incidence estimates versus the binning approach which yields accurate incidence estimates but are uninformative.

3. Post hoc age averaging is one way of improving incidence and incidence trends estimates, a caveat is that it hides the age structure in the incidence estimate which is more epidemiologically interesting.

4. In cases where incidence is truly high, we show that one can derive informative incidence trends and statistically significant age-specific incidence trends estimates from just two cross-sectional surveys. The incidence around the crucial ages are informative and further benefit from post hoc age averaging.

Our explorations focused on using simulated datasets to address some missed opportunities available via Mahiane et al. [3] incidence estimator. We investigated the applicability of the estimator in a world where HIV surveillance data (population-level survey and excess mortality) is abundant. We made some methodological improvements and showed that;

1. Similar to the recency approach summarising population survey data into prevalence and gradient of prevalence to yield age-specific incidence estimates requires using a moving window approach. Our analysis suggest that, an investigator should consider using higher-order polynomial (cubic/quartic) and a wide inclusion distance (>5) to estimate the prevalence and its derivative.

2. Similar to Mahiane et al. [3], Grebe et al. [9] the synthetic cohort yields informative incidence estimates at younger compared to the older ages, which suggest focusing resources on younger ages or key populations for surveys meant for incidence estimation.

3. Mahiane framework is most reliable in estimating incidence at the midpoint where the gradient of the prevalence is accurate versus the actual survey dates. Unless one believes prevalence was time-invariant at the survey dates in that time leading to the survey then incidence from the date will be an optimal estimate.

4. Incidence trends from three cross-sectional surveys are reliable (accurate and informative) compared to the incidence trends from two cross-sectional surveys.

Chapter 7 adopts the methodological improvements made in the previous chapters (**??**) as optimal approaches to handle the survey data, and based on these improvements we simultaneously used both the Kassanjee and Mahiane analyses on data sets to which both are applicable, and demonstrated that;

1. It is straightforward to implement the "one size fits most" polynomial order and inclusion distance permutation to extract near-optimal incidence estimates from the Kassanjee and Mahiane methods. Based on the Kassanjee and Mahiane methods one can apply the inverse variance method to yield an estimate that is accurate and informative than either of the two.
2. Recency ascertainment does not really improve the time incidence estimates in young ages i.e., most of the information is in the synthetic cohort (given accurate and precise information on excess mortality).
3. Incidence difference estimates from just two cross-sectional surveys are unstable without recency information. If the aim is to estimate incidence differences then the recency approach alone yields accurate, but uninformative estimates unless the sample sizes are relatively huge. In such cases one should use the recency estimator in conjunction with post hoc age averaging to yield an age-range incidence trend with reduced uncertainty.
4. Three surveys are optimal in estimating incidence differences, but unfortunately, the time elapsed may be too long to derive any exciting epidemiological inferences.
5. There is no exact weight required for the optimally weighted estimator, but any value in a stipulated range of values does yield sub-optimal incidence estimates.

9.2 Conclusion

We showed how the simulation platform can be used systematically to compare incidence estimators that are either of "synthetic cohort" or "recency" form. Most importantly the simulation platform is not only meant for HIV epidemics any chronic condition is likely to benefit from the use of this platform and not for predictive/mechanistic purposes but for method development and validation. We recommend the use of the simulation platform to compare the various HIV incidence estimators and determine the optimal approach, which may lead to a consensus on how best to estimate HIV incidence.

Estimating HIV incidence and incidence trends will remain a challenge if we continuously view this problem through the same lens. Our emphasis has been on the perspective of an analyst in possession of substantial survey data, like PHIA [12], HSRC [13], KAIS [14], and DHS [22]. For such datasets, we demonstrated that the same ideas and tools developed here are applicable at the onset of the survey process, for example, by investigating the impact of key design elements like survey intervals, age ranges, and sampling density.

As far as we know this is the first study that systematically compares the performance of two incidence estimators from different frameworks and we show that the methods are comparable. The method gave similar results and confidence intervals, especially at young ages.

The following details that were not systematically investigated, and were exhaustively discussed in the body of the thesis and we again recap them as limitations below:

1. *Excess mortality is known precisely*: With real data this is not the case and it is crucial that for purposes of the Mahiane and optimally weighted incidence estimators accurate and precise “mean excess attrition/mortality" should be supplied to reduce bias and improve precision.

2. *The impact of non-zero values for false recent rate*: This is a context specific measure and in our investigations we tailored the $P_R(t)$ functions so that it yields an $FRR = 0$ and hence we advise researchers to use the correct FRR which can be determined from Kassanjee et al. [74].

3. *Multiple estimates of recency test properties*: The survey simulations for incidence estimation at midpoint, using the Kassanjee framework, where designed to have the same ‘recency’ test properties which in most real surveys is not the case. And hence investigations on interpolation and/or averaging methods of the MDRI and FRR is necessary to address this conundrum and forms an interesting question to investigate.

9.3 Limitations

We did not investigate the effect of complex-sampling strategies on the incidence. But if an experimenter has population-level survey data (with a complex sampling strategy) we recommend using bootstrapping approaches that reproduce the complex sampling strategy, to estimate the standard errors.

The Mahiane approach requires accurate and precise excess mortality measures and there is need for further research to address the question of excess mortality. But again in an era were ART is available the second term reduces considerable. Currently some studies Grebe et al. [9] rely on excess mortality estimates from the Thembisa model [57]. Also it is crucial with real data for one to consider a sensitivity analysis of the incidence estimator to misspecification or uncertainty about the excess mortality component. The thesis did not engage with the uncertainty of the MDRI, FRR, and excess mortality, which we consider as potential future work. Approaches to estimating incidence from population-level surveys have emerged, for example, Fellows et al. [95].

Arguably, the sequence of steps to yield an optimal incidence estimator are cumbersome therefore, an alternative approach would be to express a single likelihood for the trinomial survey observation recent infection, long-term infection, HIV negative, which depends on the prevalence, incidence, excess mortality, MDRI and FRR, note similar approaches were implemented in Eaton et al. [52], and Le Bao and Hallett [96].

As previously highlighted these challenges are easily addressed by using the code developed in this work. The code is readily available on request and will make way into Inctools [60]. Currently, work is on going to apply these methodological improvements to real life data, and we are aiming to have our simulation platform package (on GIT [68]) hosted on CRAN soon.

List of references

[1] Joint United Nations Programme on HIV/AIDS (UNAIDS). Global hiv and aids statistics - fact sheet, 2021. URL https://www.unaids.org/en/resources/fact-sheet.

[2] Timothy B Hallett, Basia Zaba, Jim Todd, Ben Lopman, Wambura Mwita, Sam Biraro, Simon Gregson, J Ties Boerma, Alpha Network, et al. Estimating incidence from prevalence in generalised hiv epidemics: methods and validation. *PLoS medicine*, 5(4):e80, 2008.

[3] Guy Severin Mahiane, Rachid Ouifki, Hilmarie Brand, Wim Delva, and Alex Welte. A general hiv incidence inference scheme based on likelihood of individual level data and a population renewal equation. *PloS one*, 7(9):e44377, 2012.

[4] Reshma Kassanjee, Thomas A McWalter, Till Bärnighausen, and Alex Welte. A new general biomarker-based incidence estimator. *Epidemiology (Cambridge, Mass.)*, 23 (5):721, 2012.

[5] Mary Mahy, Martina Penazzato, Andrea Ciaranello, Lynne Mofenson, Constantin T Yianoutsos, Mary-Ann Davies, and John Stover. Improving estimates of children living with hiv from the spectrum aids impact model. *AIDS (London, England)*, 31 (Suppl 1):S13–S22, 2017.

[6] John Stover, Tim Brown, Robert Puckett, and Wiwat Peerapatanapokin. Updates to the spectrum/estimations and projections package model for estimating trends and current values for key hiv indicators. *AIDS*, 31(1):S5–S11, 2017.

[7] Jeffrey W Eaton, Tim Brown, Robert Puckett, Robert Glaubius, Kennedy Mutai, Le Bao, Joshua A Salomon, John Stover, Mary Mahy, and Timothy B Hallett. The estimation and projection package age-sex model and the r-hybrid model: new tools for estimating hiv incidence trends in sub-saharan africa. *AIDS (London, England)*, 33(Suppl 3):S235, 2019.

[8] John Stover, Robert Glaubius, Lynne Mofenson, Caitlin M Dugdale, Mary-Ann Davies, Gabriela Patten, and Constantin Yiannoutsos. Updates to the spectrum/aim model for estimating key hiv indicators at national and subnational levels. *AIDS (London, England)*, 33(Suppl 3):S227, 2019.

[9] Eduard Grebe, Alex Welte, Leigh F Johnson, Gilles Van Cutsem, Adrian Puren, Tom Ellman, Jean-François Etard, Consortium for the Evaluation, Performance of HIV Incidence Assays (CEPHIA), and Helena Huerga. Population-level hiv incidence estimates using a combination of synthetic cohort and recency biomarker approaches in kwazulu-natal, south africa. *PloS one*, 13(9):e0203638, 2018.

[10] Brian G Williams and Eleanor Gouws. The epidemiology of human immunodeficiency virus in south africa. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1411):1077–1086, 2001.

[11] Demographic Health Surveys. dhsprogram.com, 2018. URL https://dhsprogram.com/what-we-do/survey-Types/dHs.cfm.

[12] Population based HIV Impact Assessment. The phia project, 2018. URL https://phia.icap.columbia.edu/.

[13] Human Sciences Research Council. South African National HIV Prevalence HSRC. Hiv incidence, behaviour and communication survey (sabssm) 2012: Combined - all provinces. [data set]. sabssm 2012 combined. version 2.0. pretoria south africa: Human sciences research council [producer] 2012, 2018. URL http://curation.hsrc.ac.za/doi-10.14749-1517402043.

[14] National AIDS Control Council KAIS. Kais 2012 final report, 2018. URL https://nacc.or.ke/kais-2012-final-report/.

[15] Elizabeth Pisani, Stefano Lazzari, Neff Walker, and Bernhard Schwartländer. Hiv surveillance: a global perspective. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 32:S3–S11, 2003.

[16] Robert Magnani, Keith Sabin, Tobi Saidel, and Douglas Heckathorn. Review of sampling hard-to-reach and hidden populations for hiv surveillance. *Aids*, 19:S67–S72, 2005.

[17] Ron Brookmeyer, Jacob Konikoff, Oliver Laeyendecker, and Susan H Eshleman. Estimation of hiv incidence using multiple biomarkers. *American journal of epidemiology*, 177(3):264–272, 2013.

[18] Basia Zaba and Simon Gregson. Measuring the impact of hiv on fertility in africa. *AIDS (London, England)*, 12:S41–50, 1998.

[19] Olive Shisana, Thomas Rehle, LCea Simbayi, Khangelani Zuma, S Jooste, Pillay-Van Wyk, Ntombizodwa Mbelle, J Van Zyl, W Parker, NP Zungu, et al. South african national hiv prevalence, incidence, behaviour and communication survey, 2008: a turning tide among teenagers? 2009.

[20] Thomas M Rehle, Timothy B Hallett, Olive Shisana, Victoria Pillay-van Wyk, Khangelani Zuma, Henri Carrara, and Sean Jooste. A decline in new hiv infections in south africa: estimating hiv incidence from three national hiv surveys in 2002, 2005 and 2008. *PloS one*, 5(6):e11094, 2010.

[21] Thomas Rehle, Leigh Johnson, Timothy Hallett, Mary Mahy, Andrea Kim, Helen Odido, Dorina Onoya, Sean Jooste, Olive Shisana, Adrian Puren, et al. A comparison of south african national hiv incidence estimates: A critical appraisal of different methods. *PloS one*, 10(7):e0133255, 2015.

[22] The Demographic Health Survey Program DHS. Hiv prevalence, 2018. URL https://dhsprogram.com/topics/HIV-Corner/hiv-prev/index.cfm.

[23] Population based HIV Impact Assessment (PHIA). Eswatini, 2018. URL https://phia.icap.columbia.edu/countries/swaziland//.

[24] Jessica Justman, Jason B Reed, George Bicego, Deborah Donnell, Keala Li, Naomi Bock, Alison Koler, Neena M Philip, Charmaine K Mlambo, Bharat S Parekh, et al. Swaziland hiv incidence measurement survey (shims): a prospective national cohort study. *The Lancet HIV*, 4(2):e83–e92, 2017.

[25] Wanjiru Waruiru, Carol Ngare, Victor Ssempijja, Thomas Gachuki, Inviolata Njoroge, Patricia Oluoch, Davies O Kimanga, William K Maina, Rex Mpazanje, Andrea A Kim, et al. The status of hiv testing and counseling in kenya: results from a nationally representative population-based survey. *Journal of acquired immune deficiency syndromes (1999)*, 66(Suppl 1):S27, 2014.

[26] William K Maina, Andrea A Kim, George W Rutherford, Malayah Harper, Boniface O K'Oyugi, Shahnaaz Sharif, George Kichamu, Nicholas M Muraguri, Willis Akhwale, and Kevin M De Cock. Kenya aids indicator surveys 2007 and 2012: implications for public health policies for hiv prevention and treatment. *Journal of acquired immune deficiency syndromes (1999)*, 66(Suppl 1):S130, 2014.

[27] Joint United Nations Programme on HIV/AIDS. Global report. *UNAIDS report on the global AIDS epidemic*, 364:2010, 2010.

[28] Alain Pinsonneault and Kenneth Kraemer. Survey research methodology in management information systems: an assessment. *Journal of management information systems*, 10(2):75–105, 1993.

[29] Frank Tanser, Victoria Hosegood, Till Bärnighausen, Kobus Herbst, Makandwe Nyirenda, William Muhwava, Colin Newell, Johannes Viljoen, Tinofa Mutevedzi, and Marie-Louise Newell. Cohort profile: Africa centre demographic information system (acdis) and population-based hiv survey. *International journal of epidemiology*, 37(5):956–962, 2008.

[30] Paul J Feldblum, Mary H Latka, Johann Lombaard, Candice Chetty, Pai-Lien Chen, Connie Sexton, and Shelly Fischer. Hiv incidence and prevalence among cohorts of women with higher risk behaviour in bloemfontein and rustenburg, south africa: a prospective study. *BMJ open*, 2(1):e000626, 2012.

[31] Junjie Xu, Minghui An, Xiaoxu Han, Manhong Jia, Yanling Ma, Min Zhang, Qinghai Hu, Zhenxing Chu, Jing Zhang, Yongjun Jiang, et al. Prospective cohort study of hiv incidence and molecular characteristics of hiv among men who have sex with men (msm) in yunnan province, china. *BMC infectious diseases*, 13(1):3, 2013.

[32] Paul J Feldblum, Sónia Enosse, Karine Dubé, Paulo Arnaldo, Chadreque Muluana, Reginaldo Banze, Aristides Nhanala, Joana Cunaca, Pai-Lien Chen, Merlin L Robb, et al. Hiv prevalence and incidence in a cohort of women at higher risk for hiv acquisition in chokwe, southern mozambique. *PLoS One*, 9(5):e97547, 2014.

[33] H Irene Hall, Ruiguang Song, Philip Rhodes, Joseph Prejean, Qian An, Lisa M Lee, John Karon, Ron Brookmeyer, Edward H Kaplan, Matthew T McKenna, et al. Estimation of hiv incidence in the united states. *Jama*, 300(5):520–529, 2008.

[34] Timothy D Mastro, Andrea A Kim, Timothy Hallett, Thomas Rehle, Alex Welte, Oliver Laeyendecker, Tom Oluoch, and Jesus M Garcia-Calleja. Estimating hiv incidence in populations using tests for recent infection: issues, challenges and the way forward. *Journal of HIV AIDS surveillance & epidemiology*, 2(1):1, 2010.

[35] Ron Brookmeyer and Jacob Konikoff. Statistical considerations in determining hiv incidence from changes in hiv prevalence. *Statistical Communications in Infectious Diseases*, 3(1), 2011.

[36] WB Kannel and DL McGee. Diabetes and glucose tolerance as risk factors for cardiovascular disease: the framingham study. *Diabetes care*, 2(2):120–126, 1979.

[37] Timothy B Hallett, John Stover, Vinod Mishra, Peter D Ghys, Simon Gregson, and Ties Boerma. Estimates of hiv incidence from household-based prevalence surveys. *AIDS (London, England)*, 24(1):147, 2010.

[38] Ron Brookmeyer and Thomas C Quinn. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. *American journal of epidemiology*, 141(2):166–172, 1995.

[39] Marvin J Podgor and M Cristina Leske. Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in medicine*, 5(6):573–578, 1986.

[40] Robert C Brunet and Claudio J Struchiner. Rate estimation from prevalence information on a simple epidemiologic model for health interventions. *theoretical population biology*, 50(3):209–226, 1996.

[41] Robert C Brunet and Claudio J Struchiner. A non-parametric method for the reconstruction of age-and time-dependent incidence from the prevalence data of irreversible diseases with differential mortality. *Theoretical population biology*, 56(1): 76–90, 1999.

[42] Brian Williams, Eleanor Gouws, David Wilkinson, and Salim Abdool Karim. Estimating hiv incidence rates from age prevalence data in epidemic situations. *Statistics in medicine*, 20(13):2003–2016, 2001.

[43] Centers for Disease Control, Prevention (CDC, et al. Hiv prevalence estimates–united states, 2006. *MMWR. Morbidity and mortality weekly report*, 57(39):1073–1076, 2008.

[44] Philip S Rosenberg. Backcalculation models of age-specific hiv incidence rates. *Statistics in medicine*, 13(19-20):1975–1990, 1994.

[45] Rino Bellocco and Ian C Marschner. Joint analysis of hiv and aids surveillance data in back-calculation. *Statistics in medicine*, 19(3):297–311, 2000.

[46] Ron Brookmeyer and Mitchell H Gail. A method for obtaining short-term projections and lower bounds on the size of the aids epidemic. *Journal of the American Statistical Association*, 83(402):301–308, 1988.

[47] Francesco Brizzi, Paul J Birrell, Martyn T Plummer, Peter Kirwan, Alison E Brown, Valerie C Delpech, O Noel Gill, and Daniela De Angelis. Extending bayesian backcalculation to estimate age and time specific hiv incidence. *Lifetime data analysis*, 25 (4):757–780, 2019.

[48] Peter Bacchetti, Mark R Segal, and Nicholas P Jewell. Backcalculation of hiv infection rates. *Statistical Science*, pages 82–101, 1993.

[49] Robert S Janssen, Glen A Satten, Susan L Stramer, Bhupat D Rawal, Thomas R O'brien, Barbara J Weiblen, Frederick M Hecht, Noreen Jack, Farley R Cleghorn, James O Kahn, et al. New testing strategy to detect early hiv-1 infection for use in incidence estimates and for clinical and prevention purposes. *Jama*, 280(1):42–48, 1998.

[50] John Hargrove, Cari Van Schalkwyk, and Hayden Eastwood. Bed estimates of hiv incidence: resolving the differences, making things simpler. *PLoS One*, 7(1):e29736, 2012.

[51] Leigh Johnson, Rob Dorrington, Thomas Rehle, Sean Jooste, Linda-Gail Bekker, Melissa Wallace, Landon Myer, and Andrew Boulle. Thembisa version 1.0: A model for evaluating the impact of hiv/aids in south africa. *Centre for Infectious Disease Epidemiology and Research Working Paper February*, 2014.

[52] Jeffrey W Eaton, Laura Dwyer-Lindgren, Steve Gutreuter, Megan O'Driscoll, Oliver Stevens, Sumali Bajaj, Rob Ashton, Alexandra Hill, Emma Russell, Rachel Esra, et al. Naomi: a new modelling tool for estimating hiv epidemic indicators at the district level in sub-saharan africa. *Journal of the International AIDS Society*, 24:e25788, 2021.

[53] Severin G Mahiane, Jeffrey W Eaton, Robert Glaubius, Kelsey K Case, Keith M Sabin, and Kimberly Marsh. Updates to spectrum's case surveillance and vital registration tool for hiv estimates and projections. *Journal of the International AIDS Society*, 24:e25777, 2021.

[54] John Stover, Tim Brown, and Milly Marston. Updates to the spectrum/estimation and projection package (epp) model to estimate hiv trends for adults and children. *Sex Transm Infect*, 88(Suppl 2):i11–i16, 2012.

[55] Le Bao, Xiaoyue Niu, Mary Mahy, and Peter D Ghys. Estimating hiv epidemics for sub-national areas. *arXiv preprint arXiv:1508.06618*, 2015.

[56] Leontine Alkema, Adrian E Raftery, and Tim Brown. Bayesian melding for estimating uncertainty in national hiv prevalence estimates. *Sexually transmitted infections*, 84(Suppl 1):i11–i16, 2008.

[57] Leigh Johnson and Rob Dorrington. Thembisa version 4.3: A model for evaluating the impact of hiv/aids in south africa. *Centre for Infectious Disease Epidemiology and Research Working Paper February*, 2020.

[58] Institute for Health Metrics and Evaluation. Hiv/aids, 2020. URL http://www.healthdata.org/hiv-aids,urldate={2019-12-31}.

[59] Christopher JL Murray, Katrina F Ortblad, Caterina Guinovart, Stephen S Lim, Timothy M Wolock, D Allen Roberts, Emily A Dansereau, Nicholas Graetz, Ryan M Barber, Jonathan C Brown, et al. Global, regional, and national incidence and mortality for hiv, tuberculosis, and malaria during 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 384(9947):1005–1070, 2014.

[60] Eduard Grebe, Alex Welte, Avery McIntosh, Petra Baumler, and Stefano Ongarello. *inctools: Incidence Estimation Tools*, 2019. URL https://CRAN.R-project.org/package=inctools. R package version 1.0.15.

[61] George TH Ellison. 'population profiling' and public health risk: when and how should we use race/ethnicity? 2005.

[62] Anthon du P Heyns, Richard J Benjamin, JP Ronel Swanevelder, Megan E Laycock, Brandee L Pappalardo, Robert L Crookes, David J Wright, and Michael P Busch. Prevalence of hiv-1 in blood donations following implementation of a structured blood safety policy in south africa. *Jama*, 295(5):519–526, 2006.

[63] J Steven McDougal, Bharat S Parekh, Michael L Peterson, Bernard M Branson, Trudy Dobbs, Marta Ackers, and Marc Gurwith. 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 Research & Human Retroviruses*, 22(10):945–952, 2006.

[64] Thomas A McWalter and Alex Welte. A comparison of biomarker based incidence estimators. *PloS one*, 4(10):e7368, 2009.

[65] Reshma Kassanjee. Characterisation and application of tests for recent infection for hiv incidence surveillance. *Johannesburg: University of the Witwatersrand*, 2014.

[66] Reshma Kassanjee, Daniela De Angelis, Marian Farah, Debra Hanson, Jan Phillipus Lourens Labuschagne, Oliver Laeyendecker, Stéphane Le Vu, Brian Tom, Rui Wang, and Alex Welte. Cross-sectional hiv incidence surveillance: A benchmarking of approaches for estimating the 'mean duration of recent infection'. *Statistical communications in infectious diseases*, 9(1), 2017.

[67] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

[68] Mhlanga Laurette, Grebe Eduard, and Welte Alex. *Population Simulation*, 2021. URL https://rdrr.io/github/laurettemhlanga/PopulationSimulation/. [Internet].

[69] Mhlanga Laurette, Grebe Eduard, and Welte Alex. Optimising hiv incidence estimation for two/more cross-sectional surveys without recency. forthcoming.

[70] Mhlanga Laurette, Grebe Eduard, and Welte Alex. Recent-infection testing in population-based hiv surveys: What it can give us, and how to get it? forthcoming.

[71] Laurette Mhlanga, Grebe Eduard, and Alex Welte. Optimal accounting for age and time structure of hiv incidence estimates based on cross-sectional survey data with ascertainment of 'recent infection'. 2021.

[72] Harry H Ku et al. Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards*, 70(4), 1966.

[73] Gary W Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

[74] Reshma Kassanjee, Christopher D Pilcher, Michael P Busch, Gary Murphy, Shelley N Facente, Sheila M Keating, Elaine Mckinney, Kara Marson, Matthew A Price, Jeffrey N Martin, et al. Viral load criteria and threshold optimization to improve hiv incidence assay characteristics-a cephia analysis. *AIDS (London, England)*, 30 (15):2361, 2016.

[75] World Health Organization et al. Who working group on hiv incidence measurement and data use: 3-4 march 2018, boston, ma, usa: meeting report. Technical report, World Health Organization, 2018.

[76] Laurette Mhlanga, Grebe Eduard, and Alex Welte. Smoothing age/time structure of hiv prevalence, for optimal use in synthetic cohort based incidence estimation. 2021.

[77] Wendy Sykes, Laurette Mhlanga, Ronel Swanevelder, Tanya Nadia Glatt, Eduard Grebe, Charl Coleman, Nadia Pieterson, Russell Cable, Alex Welte, Karin van den Berg, et al. Prevalence of anti-sars-cov-2 antibodies among blood donors in northern cape, kwazulu-natal, eastern cape, and free state provinces of south africa in january 2021. 2021.

[78] Laurette Mhlanga, Marion Vermeulen, Eduard Grebe, and Alex Welte. Sars cov 2 infection fatality rate estimates for south africa. 2021.

[79] Jaap Goudsmit. The paramount importance of serological surveys of sars-cov-2 infection and immunity. *European journal of epidemiology*, 35(4):331–333, 2020.

[80] Chih-Cheng Lai, Jui-Hsiang Wang, and Po-Ren Hsueh. Population-based seroprevalence surveys of anti-sars-cov-2 antibody: An up-to-date review. *International Journal of Infectious Diseases*, 2020.

[81] Micheal Busch. Private Communication, 2021.

[82] Takwanisa Machemedze, Andrew Kerr, Rob Dorrington, et al. *South African population projection and household survey sample weight recalibration*. United Nations University World Institute for Development Economics Research, 2020.

[83] Statista Research Department. Total population of south africa 2018, by ethnic groups. 2021. URL https://www.statista.com/statistics/1116076/total-population-of-south-africa-by-population-group/.

[84] Michelle Daya, Lize Van Der Merwe, Ushma Galal, Marlo Möller, Muneeb Salie, Emile R Chimusa, Joshua M Galanter, Paul D Van Helden, Brenna M Henn, Chris R Gignoux, et al. A panel of ancestry informative markers for the complex five-way admixed south african coloured population. *PloS one*, 8(12):e82224, 2013.

[85] National Institute for Communicable Diseases. Covid-19 weekly epidemiology brief, 2021. URL https://www.nicd.ac.za/wp-content/uploads/2021/01/COVID-19-Weekly-Epidemiology-Briefweek-2-2021.pdf.

[86] National Institute for Communicable Diseases. Covid-19 weekly epidemiology brief., 2021. URL https://www.nicd.ac.za/wp-content/uploads/2021/01/COVID-19-Weekly-Epidemiology-Briefweek-2-2021.pdf.

[87] M Hsiao, MA Davies, E Kalk, et al. Sars-cov-2 seroprevalence in the cape town metropolitan sub-districts after the peak of infections. *NICD COVID-19 Special Public Health Surveill Bull*, 18:1–9, 2020.

[88] Jane Alexandra Shaw, Maynard Meiring, Tracy Cummins, Novel N Chegou, Conita Claassen, Nelita Du Plessis, Marika Flinn, Andriette Hiemstra, Léanie Kleynhans, Vinzeigh Leukes, et al. Higher sars-cov-2 seroprevalence in workers with lower socioeconomic status in cape town, south africa. *Plos one*, 16(2):e0247852, 2021.

[89] Jackie Kleynhans, Stefano Tempia, Nicole Wolter, Anne von Gottberg, Jinal N Bhiman, Amelia Buys, Jocelyn Moyes, Meredith L McMorrow, Kathleen Kahn, F Xavier Gómez-Olivé, et al. Longitudinal sars-cov-2 seroprevalence in a rural and urban community household cohort in south africa, during the first and second waves july 2020-march 2021. *medRxiv*, 2021.

[90] Ester C Sabino, Lewis F Buss, Maria PS Carvalho, Carlos A Prete, Myuki AE Crispim, Nelson A Fraiji, Rafael HM Pereira, Kris V Parag, Pedro da Silva Peixoto, Moritz UG Kraemer, et al. Resurgence of covid-19 in manaus, brazil, despite high seroprevalence. *The Lancet*, 397(10273):452–455, 2021.

[91] Marion Vermeulen, Laurette Mhlanga, Wendy Sykes, Charl Coleman, Nadia Pietersen, Russell Cable, Ronel Swanevelder, Tanya Nadia Glatt, Eduard Grebe, Alex Welte, et al. Prevalence of anti-sars-cov-2 antibodies among blood donors in south africa during the period january-may 2021. 2021.

[92] D Bradshaw, RE Dorrington, R Laubscher, TA Moultrie, and P Groenewald. Tracking mortality in near to real time provides essential information about the impact of the covid-19 pandemic in south africa in 2020. *South African Medical Journal*, 2021.

[93] Andrew T Levin, William P Hanage, Nana Owusu-Boaitey, Kensington B Cochran, Seamus P Walsh, and Gideon Meyerowitz-Katz. Assessing the age specificity of infection fatality rates for covid-19: systematic review, meta-analysis, and public policy implications. *European journal of epidemiology*, pages 1–16, 2020.

[94] S Wendy, L Mhlanga, R Swanevelder, T.G Glatt, E Grebe, C Coleman, N Pieterson, R Cable, A Welte, K van den Berg, and M Vermeulen. Prevalence of anti-sars-cov-2 antibodies among blood donors in all provinces of south africa during two time points (january and may) 2021. forthcoming.

[95] Ian E Fellows, Ray W Shiraishi, Peter Cherutich, Thomas Achia, Peter W Young, and Andrea A Kim. A new method for estimating hiv incidence from a single cross-sectional survey. *Plos one*, 15(8):e0237221, 2020.

[96] Jingyi Ye Le Bao and Timothy B Hallett. Incorporating incidence information within the unaids estimation and projection package framework: a study based on simulated incidence assay data. *AIDS (London, England)*, 28(4):S515, 2014.