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

Laurette Mhlanga (laurette@aims.ac.tz)
South African DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University https://orcid.org/0000-0002-7805-4231

Grebe Eduard
Vitalant Research Institute https://orcid.org/0000-0001-7046-7245

Alex Welte
South African DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University https://orcid.org/0000-0001-7139-7509

Research Article

Keywords: HIV Incidence estimation, Incidence, Prevalence, Population-level surveys, Cross sectional surveys

Posted Date: September 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-871044/v1

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

Laurette Mhlanga, Eduard Grebe, Alex Welte

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 this 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.

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.

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 'recent 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(1–5) have been proposed. We will use the framework of Kassanjee et al.(5), which, we would argue, is the formally correct approach.

We envisage a "recent" 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 (5) and for initial efforts to investigate age structure, we suggest looking at Grebe et al (6).

A typical RITA (applied only on sensitively and specifically classified HIV infected respondents) defines 'recent 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 (This addresses the fact that consistent viral suppression typically rolls back the naïve infection-time clock)

Details of thresholds are a subtle matter involving challenges in the developing 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 (8),

The "recent 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 (7).
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 (6–8).

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, 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%. 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 matter 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
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.

Methods

Computational Environment

All computations were performed in the R system for statistical computation (10), and required only an ordinary laptop/pc hardware platform. Core evaluation of the Kassanjee estimator and it variance was largely performed by functionality in the R package *inctools*, available on CRAN.

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
- 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 (11) (Mhlanga, Welte, forthcoming), 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 1.

Estimating Incidence from One Cross-sectional Survey

A single cross-sectional survey with recency data can be used to estimate incidence using the Kassanjee estimator (5):

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

Where, P is the HIV prevalence, R is the prevalence of recency among the positive, Ω is the Mean duration of recent infections (MDRI), β is the False Recent Rate (FRR), and T is the-recency 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.

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.

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.(6). 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.

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.

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.

Ignoring age structure

Figure 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.

To efficiently produce plots of this kind, the code provides the option to have the realised sample prevalences (of HIV, and of recent infection amongst HIV positives) correspond precisely (within available rounding error) to the underlying simulated population prevalences. One can then calculate 95% uncertainty ranges by propagating binomial-exact standard errors through the Kassanjee estimator via the delta method. We have checked, over various time points in our simulated epidemic that this consistently yields results that are visually indistinguishable from indicating the mean and central 95% range of values obtained by simulating a large number of independent surveys

Age structure in 5-year bins

Figure 2 demonstrates the application of I_K to five-year age bins. We see that the incidence estimates among 15-19 year olds (in all epidemic stages) and the estimates in 1990 (all ages) are underestimates of the 'true' incidence (age weighted to the susceptible population within each bin). The other estimates are over-estimates relative to the 'true' incidence. A close reading of the intrinsic limitations of I_K, as outlined in the seminal paper (5), shows that this is not unexpected, as:

- The estimator I_K , by design, estimates a time-weighted incidence, which for heuristic purposes can be described as an average, over a time window of size similar to MDRI (Ω), preceding the survey. When incidence is rapidly rising, either population wide, over time, or in the life histories of the younger age range, then this 'retarded' estimate will tend to be an underestimate.
- Unless there is substantial in-migration, the size of the susceptible population generically decreases over time within any 'cohort'. Hence, the susceptible population at the time of the survey is typically slightly smaller than the mean susceptible population over the time-averaging window implicit in I_K. As the susceptible population occurs in the denominator of the estimator, this leads to a slight upward bias in incidence estimates.

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.

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.

Figure 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 4 shows just the relative root mean square errors, but in addition to the errors in *Prevalence* and *Prevalence of Recency among positive*s, also shows the error in the estimate I_K, for the same ages presented in Figure 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 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.

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 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.

In Figure 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 1 presents a comparison of age specific incidence estimates derived from our proposed regression with the naïve 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 7, is a graphical representation of Table 1, and compares the age specific incidence estimates derived from the binomial exact versus the regression approach for ages 20.5 to 29.5.

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 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.

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.

Figure 9 and Figure 10, present the age specific incidence difference (true and estimates) for the *central age* in the age range indicated on the x-axis, and the *age range averaged* incidence difference (true and estimates). Both the age specific and average age range incidence difference yield accurate incidence difference estimates, but the age specific estimates are less precise.

Mean incidence between survey rounds:

Various attempts will inevitably be made to estimate mean or midpoint incidence between major surveys, based on such ideas as 'synthetic cohort' analysis (12,13). 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 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.

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 (5) 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 on 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* (14) 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:

I. Similar prevalence smoothing criteria (15) – 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 (13).
II. The optimal use of both the Kassanjee and Mahiane analyses on data sets to which both are applicable (16).

Acknowledgements:

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.

The authors acknowledge the support of the South African DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis of this research. Opinions expressed and conclusions arrived at, are those of the authors and do not represent the official views of SACEMA.

References

1. Brookmeyer R, Quinn T. Estimation of current human immunodeficiency virus incidence rates from a cross-sectional survey using early diagnostic tests. Am J Epidemiol. 1995;141(2):166–72.

2. McDougal JS, Parekh BS, Peterson ML, Branson BM, Dobbs T, Ackers M, et al. Comparison of HIV Type 1 Incidence Observed during Longitudinal Follow-Up with Incidence Estimated by Cross-Sectional Analysis Using the BED Capture Enzyme Immunoassay. AIDS Res Hum Retroviruses. 2006 Oct;22(10):945–52.

3. Hargrove JW, Humphrey JH, Mutasa K, Parekh BS, McDougal JS, Ntozini R, et al. Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. AIDS. 2008 Feb;22(4):511–8.

4. McWalter TA, Welte A. A Comparison of Biomarker Based Incidence Estimators. Myer L, editor. PLoS One. 2009 Oct 7;4(10):e7368.

5. Kassanjee R, Mcwalter TA, Bärnighausen T, Welte A. A new general biomarker-based incidence estimator. Epidemiology. 2012;23(5):721–8.

6. Grebe E, Welte A, Johnson LF, Cutsem G Van, Ellman T, Puren A, et al. Population-level HIV incidence estimates using a combination of synthetic cohort and recency biomarker approaches in. 2018;1–16.

7. Kassanjee R, Angelis D De, Farah M, Hanson D, Labuschagne JPL, Laeyendecker O, et al. Cross-Sectional HIV Incidence Surveillance: A Benchmarking of Approaches for Estimating the 'Mean Duration of Recent Infection.' Stat Commun Infect Dis [Internet]. 2017 Jan 1 [cited 2021 Aug 31];9(1). Available from: https://www.degruyter.com/document/doi/10.1515/scid-2016-0002/html

8. Kassanjee R. Characterisation and Application of Tests for Recent Infection for HIV Incidence Surveillance. Vol. PhD, School of Computational and Applied Mathematics. University of the Witwatersrand; 2014.

9. Kassanjee R, Pilcher CD, Busch MP, Murphy G, Facente SN, Keating SM, et al. Viral load criteria and threshold optimization to improve HIV incidence assay characteristics. AIDS. 2016;30(15):2361–71.

10. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2018. Available from: https://www.r-project.org/

11. Mhlanga L, Grebe E, Welte A. Population Simulation [Internet]. 2021. Available from: https://rdrr.io/github/laurettemhlanga/PopulationSimulation/

12. Hallett TB, Zaba B, Todd J, Lopman B, Mwita W, Biraro S, et al. Estimating incidence from prevalence in generalised HIV epidemics: Methods and validation. PLoS Med. 2008;5(4):0611–22.

13. Mahiane GS, Ouifki R, Brand H, Delva W, Welte A. A General HIV Incidence Inference Scheme Based on Likelihood of Individual Level Data and a Population Renewal Equation. Nishiura H, editor. PLoS One [Internet]. 2012 Sep 12;7(9):e44377. Available from: http://dx.plos.org/10.1371/journal.pone.0044377

14. Grebe E, Bäumler P, McIntosh A, Ongarello S, Welte A. Incidence Estimation Tools (inctools) [Internet]. 2018. Available from: https://doi.org/10.5281/zenodo.1493401

15. Mhlanga L, Grebe E, Welte A. Optimising HIV incidence estimation for two/more cross-sectional surveys without Recency data. (*forthcoming*)

16. Mhlanga L, Grebe E, Welte A. Recent-infection testing in population-based HIV surveys: What it can give us, and how to get it? (*forthcoming*)

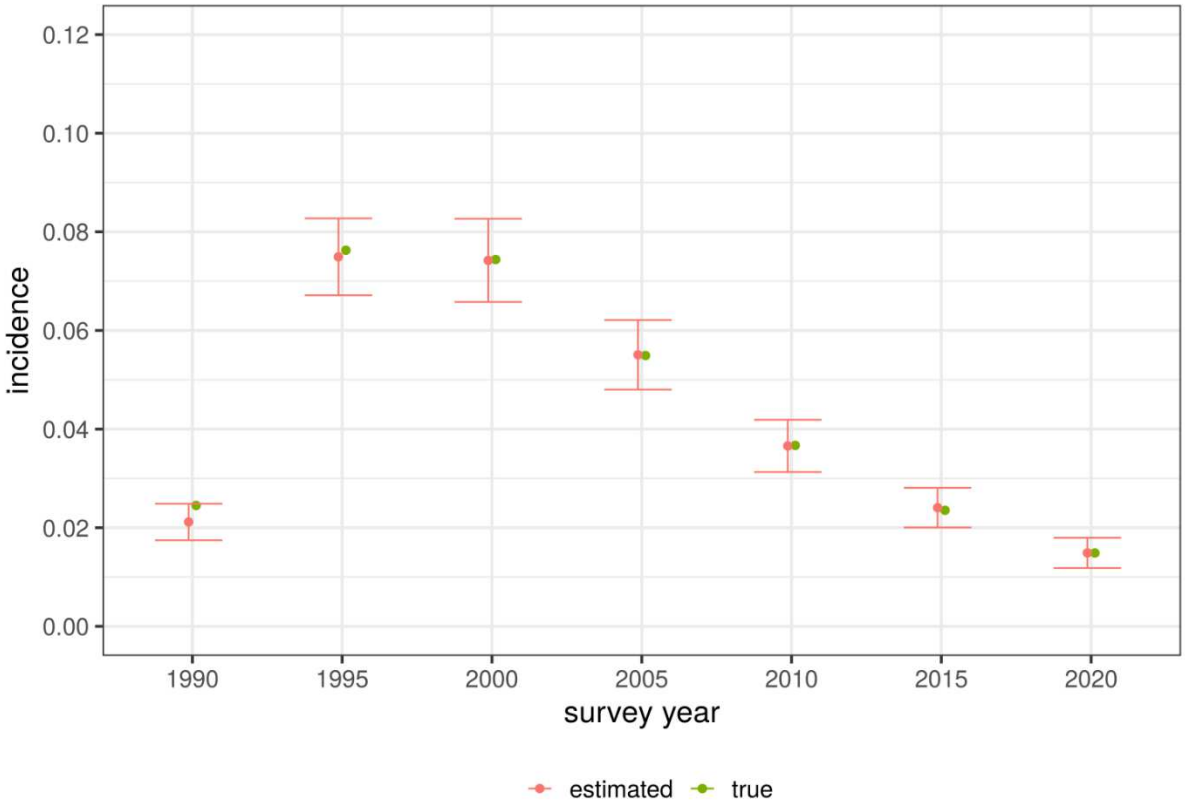

Figure 1: *Representative incidence estimates derived by treating an entire survey data set as one large age bin (red) shown alongside the 'true' incidence of the surveyed population i.e. age-weighted to the susceptible population (green)***.**

From each five-year age bin (15-19- 20-24- … - 40-44) 4 000 individuals were sampled with equal probability (total sample size of 24 000).

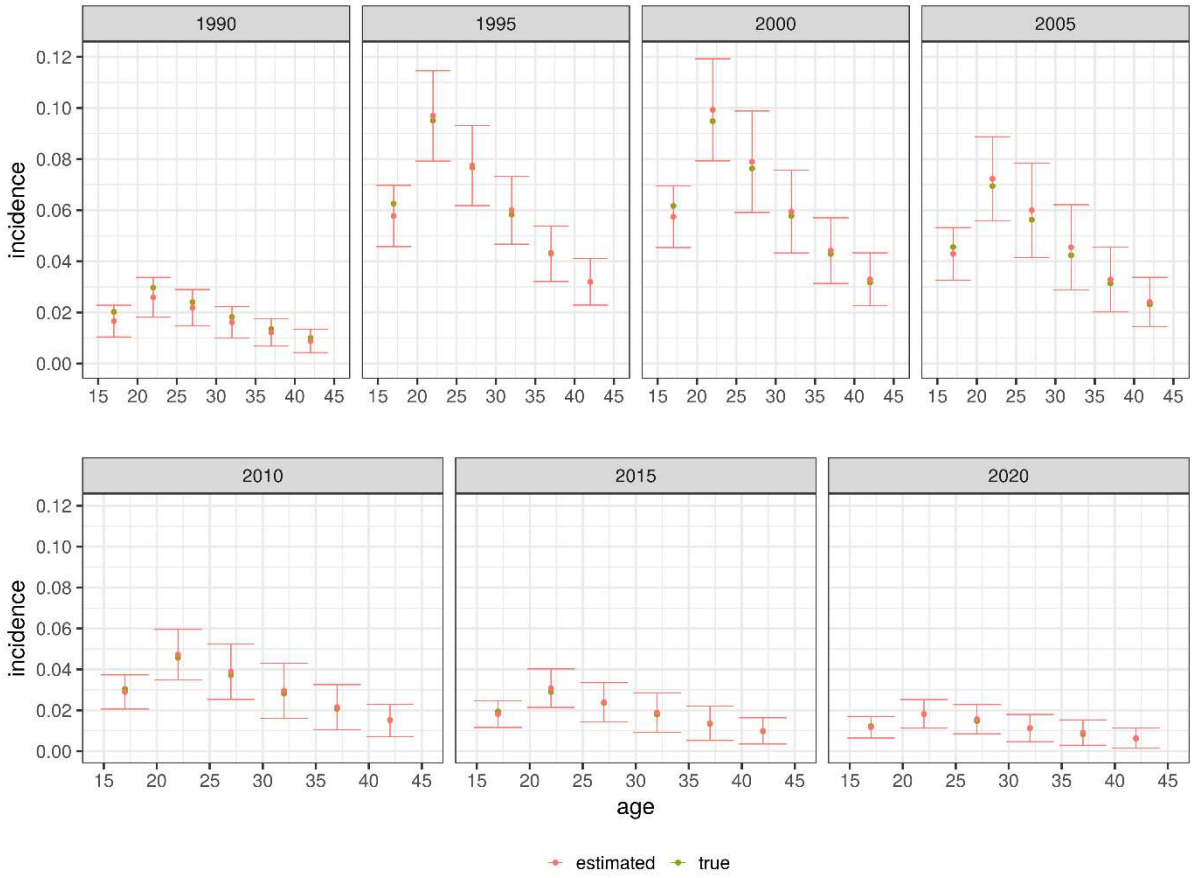

Figure 2: *Incidence estimates derived by decomposition of each survey data set into 5-year age bins (red) shown alongside the 'true' incidence within each age bin i.e. age-weighted to the susceptible population within each bin (green).*

From each five-year age bin (15-19- 20-24- …- 40-44), labelled on the figure axis by its lower bound, 4 000 individuals were sampled with equal probability (total sample size of 24 000).

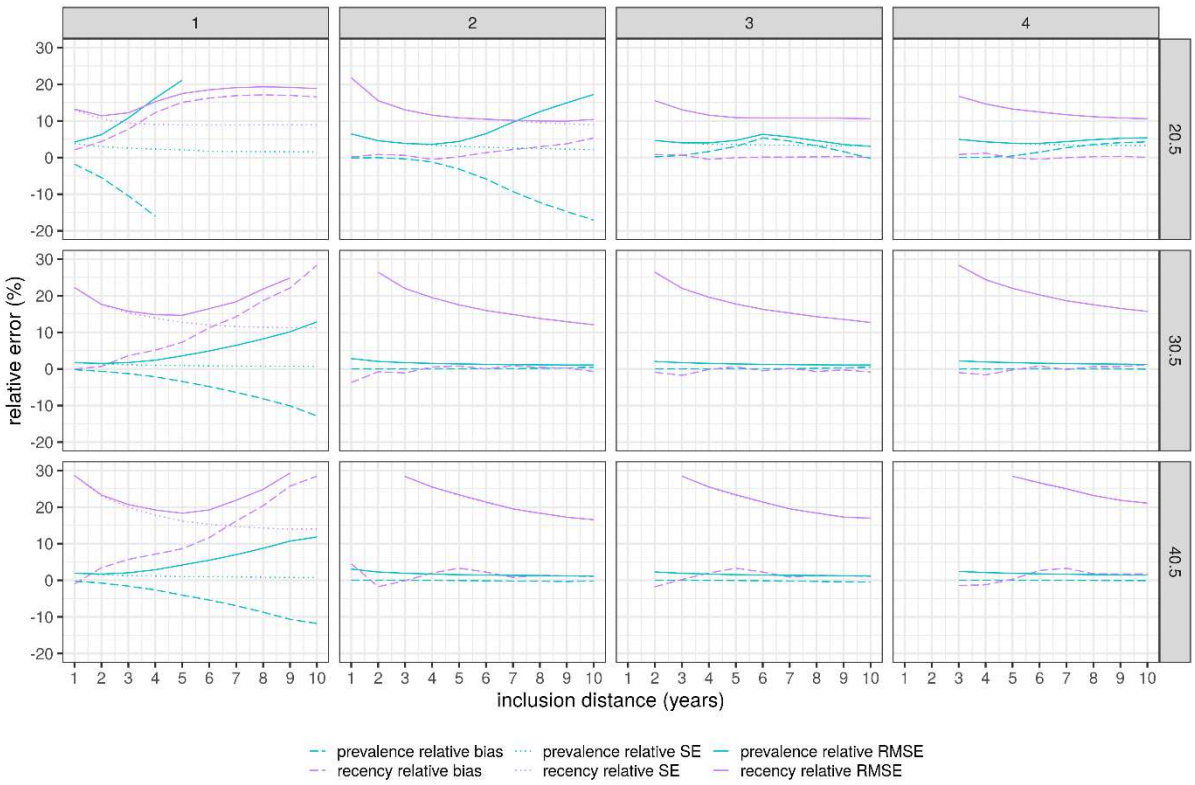

Figure 3: *Relative bias- relative standard error- and relative root mean square error- at age 20.5- 30.5- and 40.5 (rows)- using polynomial orders 1-4 (columns)- in each case as a function of data inclusion radius (x-axes).*

These errors are based on the cross-sectional survey simulated in 2015- with a sampling density of 4000 per 5-year age range.

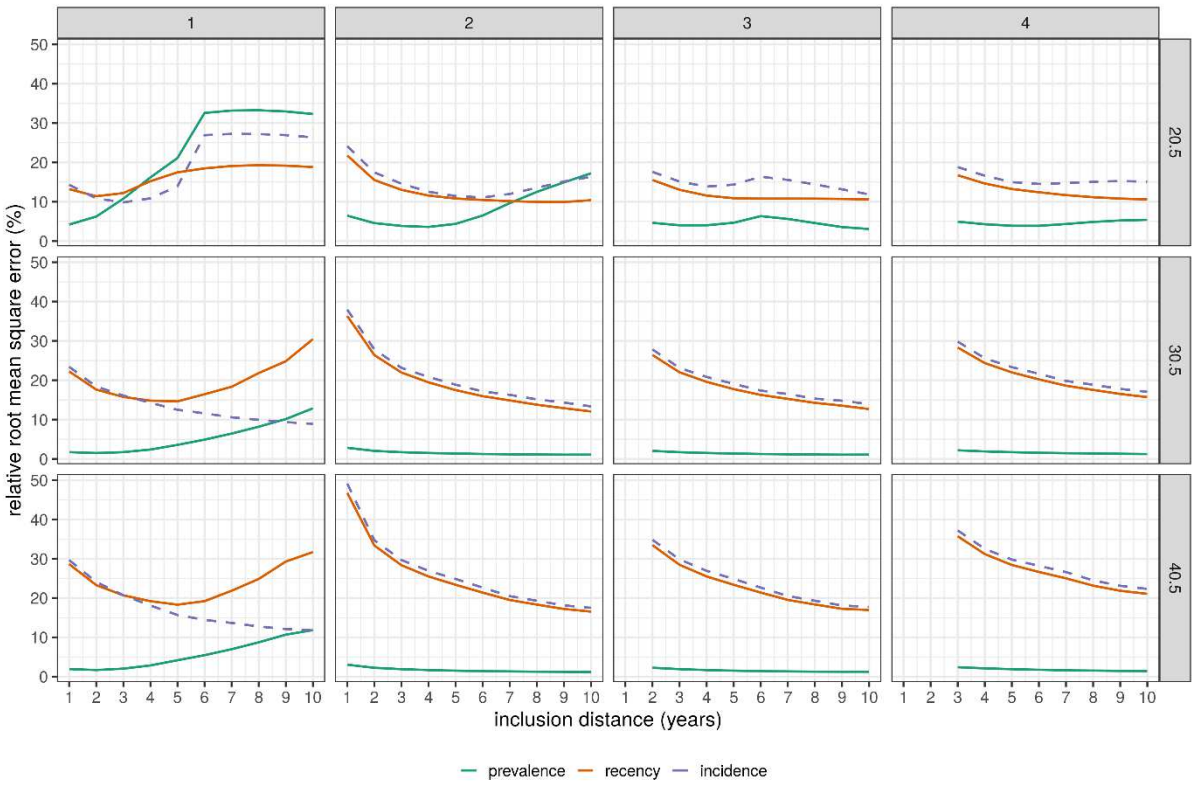

Figure 4: *Relative root mean square error for estimated prevalence, recency, and incidence*.

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.

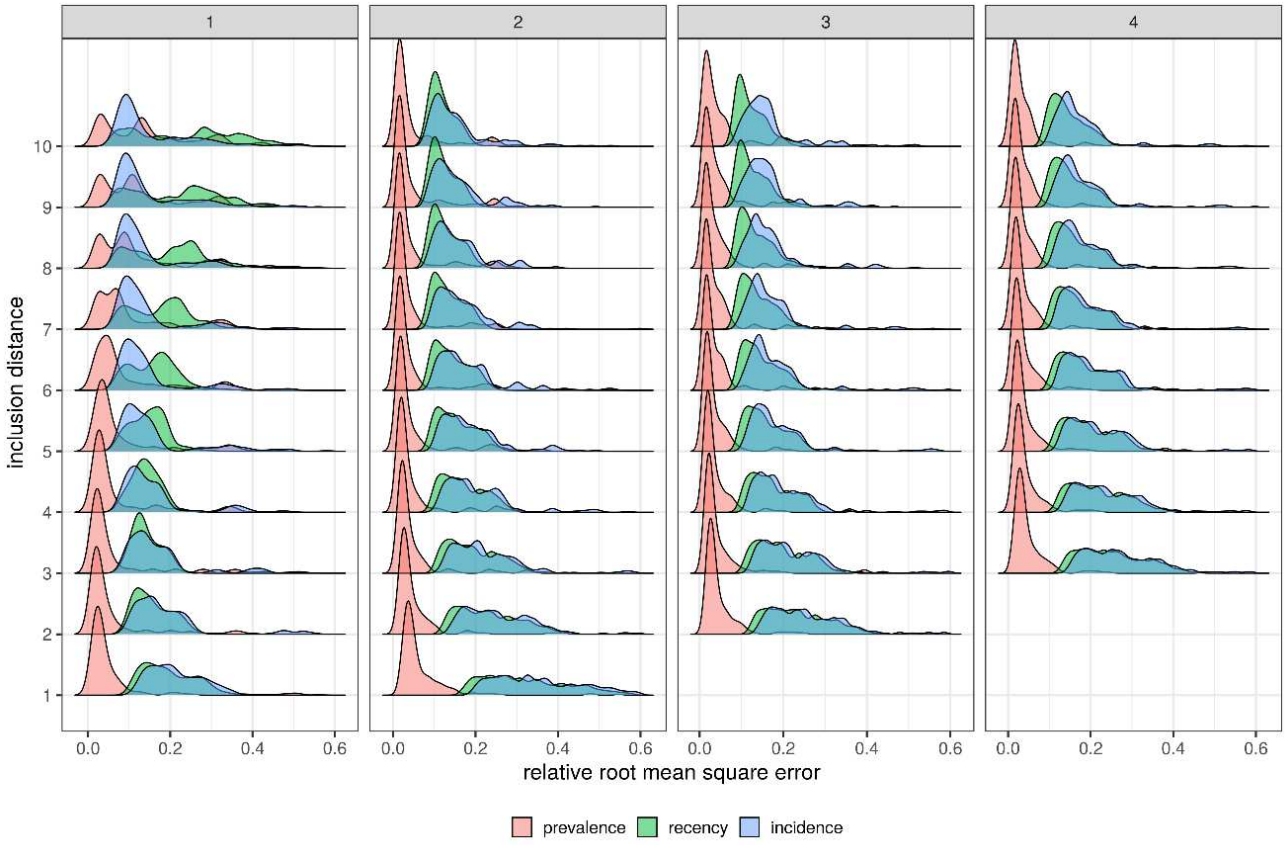

Figure 5: *Distributions of the relative root mean square errors of Prevalence- Recency- and Incidence- over the canonical permutation of survey dates and ages- shown separately for each choice of polynomial order of regression formula (columns 1 – 4) and choice of data inclusion distance in the age 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.

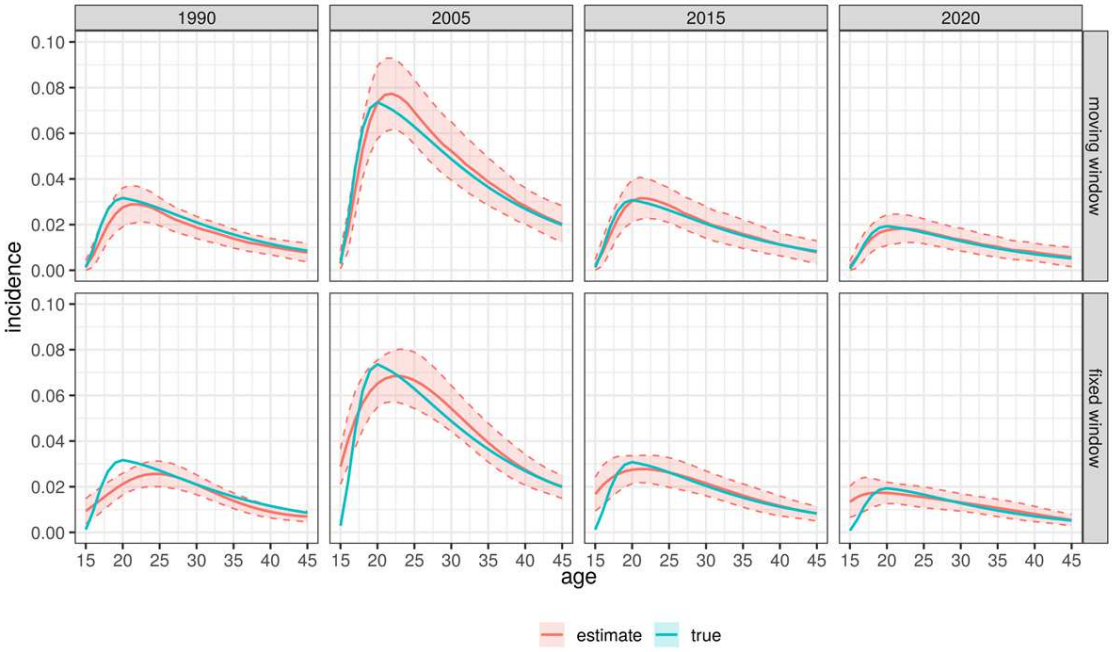

Figure 6: *Comparison of the incidence estimates to the true incidence at epidemic stages simulated in 1990- 2000- 2005- 2015- and 2020 (moving window vs fixed window).*

Table 1: *Age specific incidence estimates (in % p.a.) derived from the regression versus naïve approach.*

Age	True	Regression Estimate	Naïve Estimate	SE. Ratio $\left(\frac{Naïve}{Regression}\right)$
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

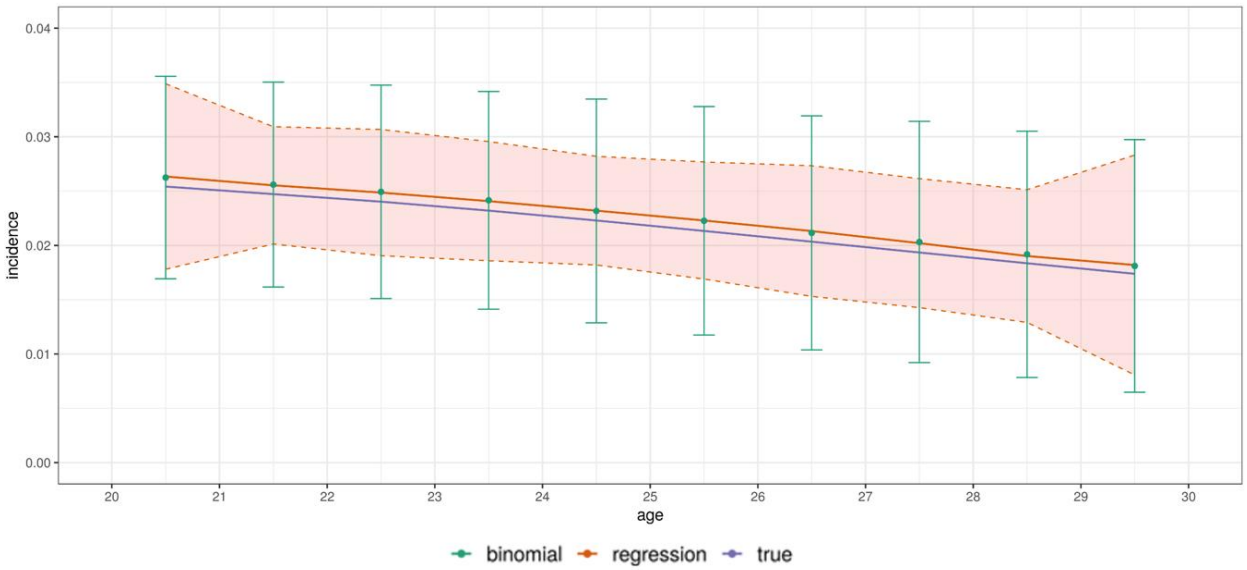

Figure 7: *Representative incidence estimate by age from focused survey of 24000 individuals in the age range 20-30.*

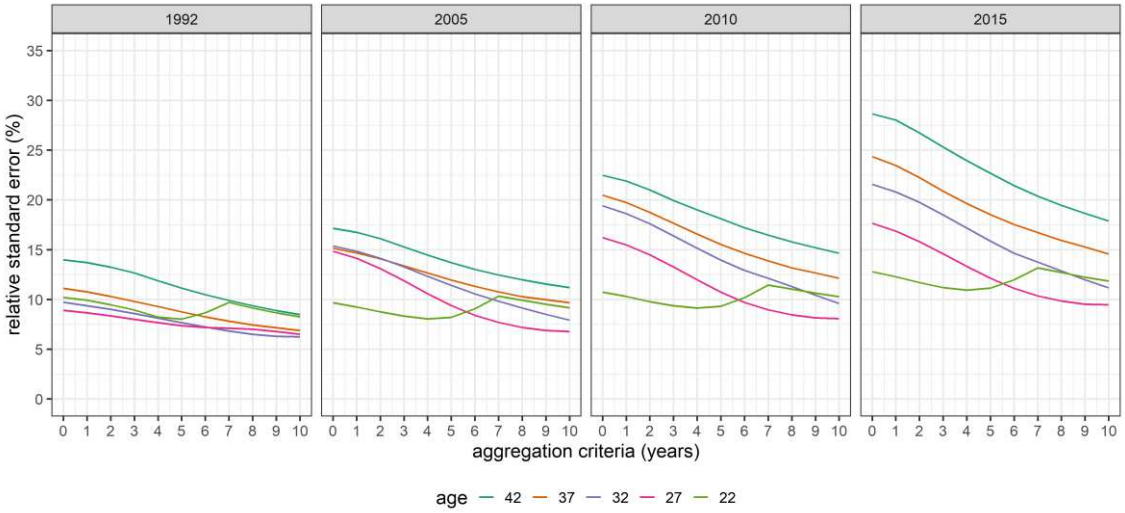

Figure 8: Incidence's relative standard error as a function of the binning strategy for simulated times 1995- 2005- 2010- 2015- and 2020 and ages 22- 27- 32- 37- and 42.

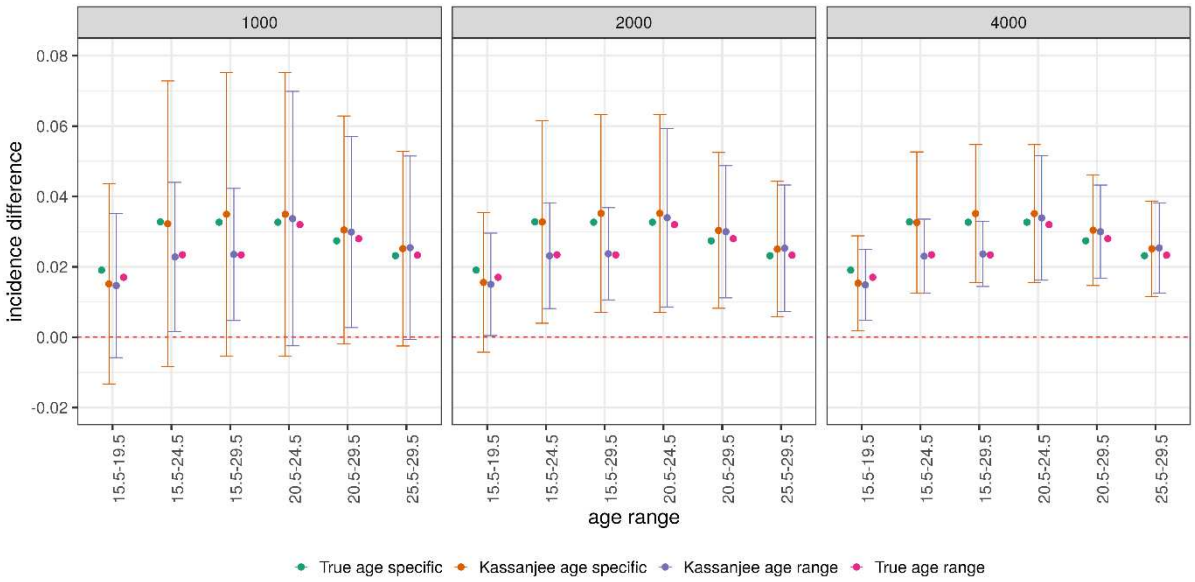

Figure 9: *Comparison of age specific and age range incidence differences*. Estimated from 2 cross sectional surveys five years apart (1993- and 1998) and simulated when incidence was increasing- the sample size is 1000- 2000- and 4000 per 5-year age range (total sample sizes are 8000- 12000- and 24000 respectively).

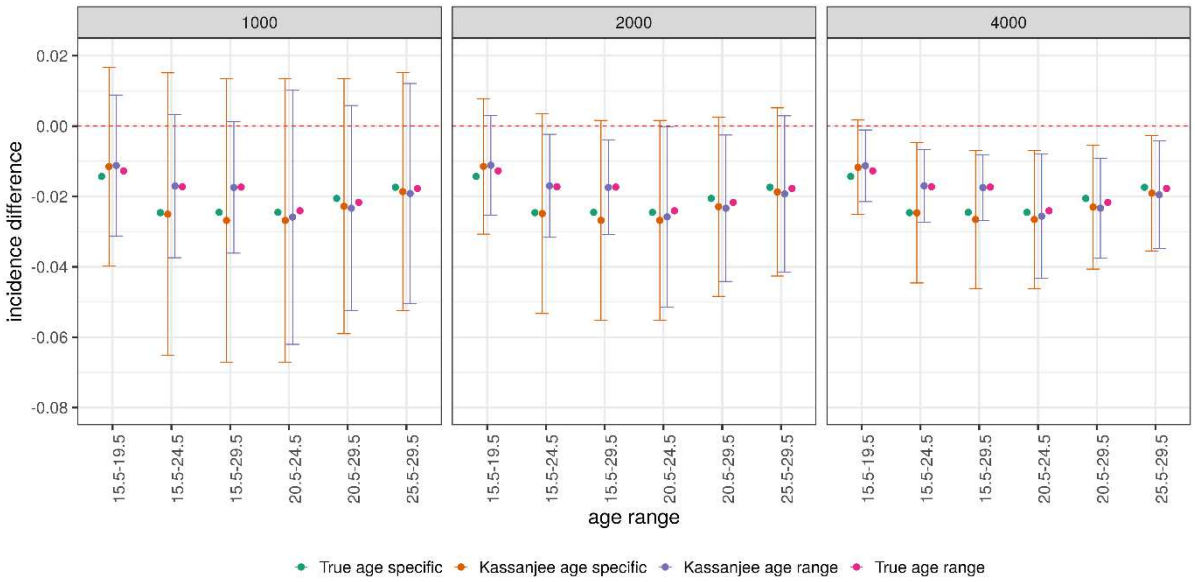

Figure 10: *Comparison of age specific and age range incidence differences***.** Estimated from 2 cross sectional surveys five years apart (2010- and 2015) and simulated when incidence was steadily decreasing- with sample size is 1000- 2000- and 5000 per 5-year age range (total sample sizes are 8000- 12000- and 24000 respectively).

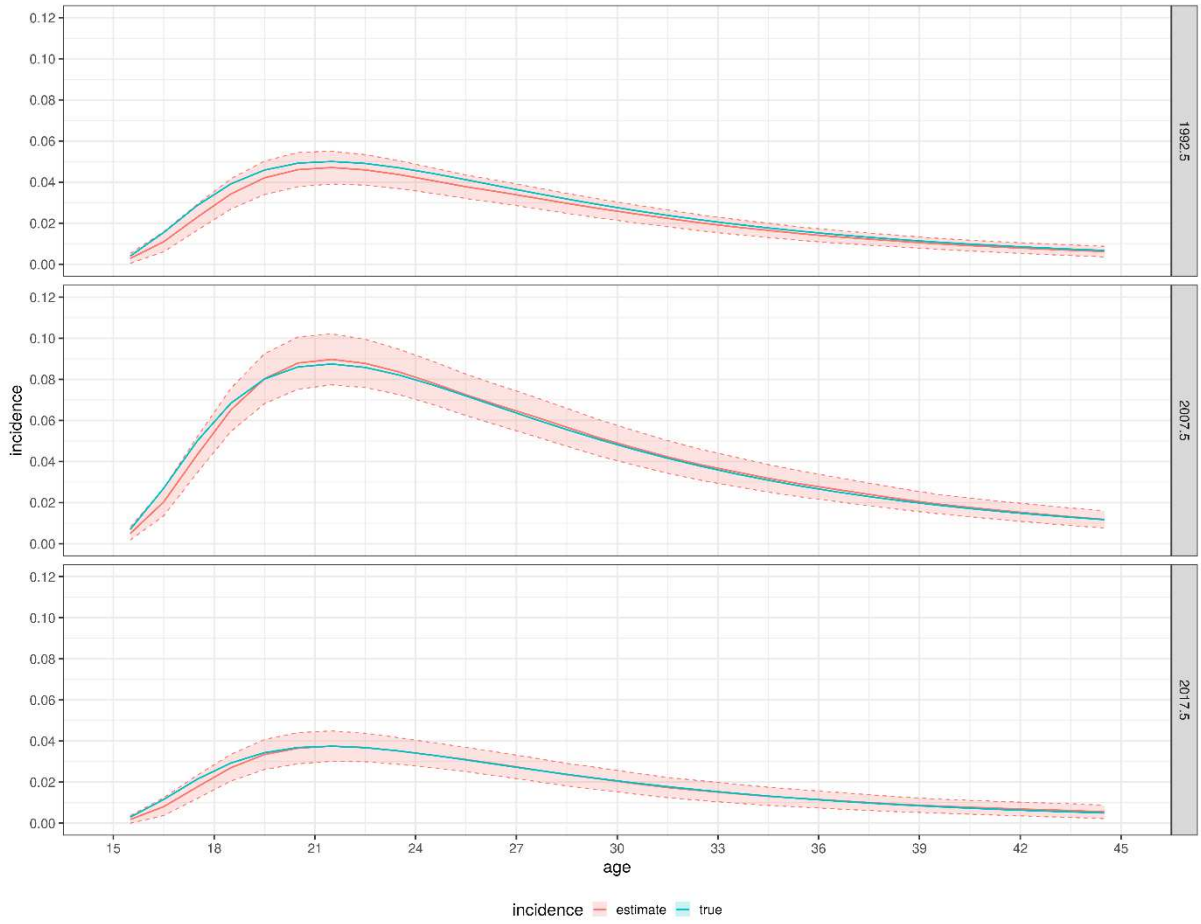

Figure 11: *Midpoint (between surveys) incidence estimates from three time points in our canonical South-Africa-like epidemiological scenario- alongside the true incidence at the corresponding time.*

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixSmoothingKassanjee.pdf