Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Laurette Mhlanga (laurette@aims.ca.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

Method Article

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

Posted Date: October 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-959136/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

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

(Running Head: Smoothing Prevalence for Incidence)

Laurette Mhlanga, Eduard Grebe, Alex Welte

October 2021

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 mortality – 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.

Introduction

Population-level cross-sectional surveys, including HIV status determination, are conducted routinely in many Sub-Saharan countries (1–5). Within the last two decades, variations of such surveys have been executed multiple times in numerous countries (6), making it possible to explicitly track 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 (7–11). 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., (11) 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:

- An estimate of prevalence for the population at an age and time of interest
- 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
- 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:

- Simulating 'realistic' epidemics and cross-sectional surveys
- Applying various smoothing algorithms to the survey data, in order to infer (age- and time-structured) prevalence of HIV infection.
- Estimating incidence, from these smoothed functions, using the Mahiane (11) framework.
- Evaluating the relative merits of various smoothing and averaging schemes, by comparing estimated with the known incidence parameter values in the simulations
- Proposing a generic 'one size 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 (12), 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.

Methods

Using a customised simulation platform (13), developed for this and some closely related investigations, and described in detail separately (Mhlanga/Welte, forthcoming - (14)), we simulated variations on 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 the Appendix) described by an age, time, and time since infection dependent population (density)). The canonical epidemiological scenario is run from 1935 to 2025. Each birth cohort is simulated to age 50.

We simulated cross-sectional surveys in 1992, 1995, 1997, 1998, 2000, 2003, 2005, 2008, 2010, 2013, 2015, and 2020. Sampling density was varied from 2000 to 16000 persons per 5-year age bracket. Incidence estimation is based on the estimator of Mahiane et al (11), given either as

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

or,

$$I_M = \frac{1}{1 - P} \cdot \frac{dP}{dt} + \mathrm{M} \cdot P \qquad 2$$

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:

$$var(\lambda) = \left[\frac{1}{(1 - P)^2} \cdot \frac{dP}{dt} + M\right]^2 \cdot \sigma_P^2 + \left[\frac{1}{1 - P}\right]^2 \cdot \sigma_{\frac{dP}{dt}}^2 + [P]^2 \cdot \sigma_M^2$$

$$+2 \cdot \left[\frac{1}{(1 - P)}\right] \cdot \left[\frac{1}{(1 - P)^2} \cdot \frac{dP}{dt} + M\right] \cdot \sigma_{P,\frac{dP}{dt}} \qquad 3$$

$$+2 \cdot \left[\frac{1}{(1 - P)^2} \cdot \frac{dP}{dt} + M\right] \cdot P \cdot \sigma_{P,M}$$

$$+2 \cdot \left[\frac{1}{1 - P}\right] \cdot P \cdot \sigma_{\frac{dP}{dt},M}$$

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 σs with double subscripts are the indicated covariances and therefore the incidence's standard error is given by $se(\lambda) = \sqrt{var(\lambda)}$.

Our investigation is very similar in inspiration to that reported in our companion article (12), 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 (15)) 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 (11) requires an estimate of 'excess mortality' associated with infection. We defer discussion of how best to estimate this, but note:

- Data from household surveys is generally not an appropriate source of mortality estimates.
- Hence, appropriate estimation of the contextually applicable excess mortality is in practice an open ended problem
- 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.
- 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 provides additional details on how uncertainty in fitting parameters is propagated into uncertainty of incidence estimates, given the potentially nonlinear relation between these parameters and the prevalence and prevalence gradient which are required in the Mahiane 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 also investigate the consequences of fitting procedure on the estimate of the 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.

Results/Discussion

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 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

Rather than considering many more individual combinations of age, time, polynomial order and inclusion distance, we show, in Figure 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 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

Effect of Sample Size

For an inter-survey interval of 5, we investigated the effects of sample size on the overall errors and present the results as distributions over age/time points in Figure 3. 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.

Inter-survey Interval

Figure 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. 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.

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 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.

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 4 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 5 shows much the same information as Figure 4, 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.

Sensitivity analysis on ‘excess mortality’.

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. In practice, it may be difficult to obtain a precise and unbiased estimate of parameter, which summarises significant diversity and complexity. Figure 6 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 7 considers fractional/relative standard errors ranging from 0 to 1, and shows the relative standard errors thus induced on the incidence estimates. 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.

Conclusion

Previously, Mahiane et al. (11), derived an instantaneous, age/time specific HIV incidence estimator, to be used with population level HIV survey data - assuming some knowledge of survival after infection, summarised as differential mortality. That prior work did not critically evaluate techniques for summarising 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 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 to 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 (12) 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 (15).
- We have not investigated the challenges of consistently obtaining the required estimates for "mean excess attrition/mortality", which must be supplied in order to interpret the prevalence and its gradient as an 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. Using the simulation environment developed for the present investigation, this is not very difficult to do.
- Beyond contextual fine tuning of data inclusion rules, polynomial order of fitting functions, and possibly choice of link function in binomial regression, there appears to be little scope for extracting 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 also demonstrated in our separately described analysis of recency data based incidence estimation (12), the base procedure naturally provides a finely age-resolved family of estimates. This can either be taken at face value, or used as the basis for further 'post hoc' 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 (such as by sex, social/economic grouping, locale) beyond the fundamental age dependence which is crucial in understanding HIV epidemiology.

Our emphasis has been on the perspective of an analysist in possession of substantial survey data, but we have also demonstrated that the same ideas and tools developed here can be used at the outset of the survey process, by investigating the impact of key design elements like survey intervals, age ranges, and sampling density.

What crucially remains to be understood, in this vein, is how best to simultaneously use both the Kassanjee and Mahiane analyses 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 (forthcoming, or updated (16)).

References

1. DHS. DHS Methodology [Internet]. 2017 [cited 2021 Oct 7]. Available from: https://dhsprogram.com/What-We-Do/Survey-Types/DHS-Methodology.cfm
2. SABSSM. 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 [Internet]. 2018 [cited 2021 Oct 7]. Available from: http://curation.hsrc.ac.za/doi-10.14749-1517402043
3. PHIA. PHIA Project [Internet]. 2017 [cited 2021 Oct 7]. Available from: http://phia.icap.columbia.edu/about/
4. Swaziland HIV Incidence Measurement Surveys. SHIMS Study protocol [Internet]. 2012 [cited 2021 Oct 7]. Available from: http://shims.icap.columbia.edu/publications/detail/shims-study-protocol-1-june-2012
5. KAIS. KAIS 2012 FINAL REPORT [Internet]. 2018 [cited 2021 Oct 7]. Available from: https://nacc.or.ke/kais-2012-final-report/
6. Rehle TM, Hallett TB, Shisana O, Pillay-van Wyk V, Zuma K, Carrara H, et al. A decline in new HIV infections in South Africa: estimating HIV incidence from three national HIV surveys in 2002, 2005 and 2008. PLoS One. 2010;5(6):e11094.
7. Brunet RC, Struchiner CJ. Rate estimation from prevalence information on a simple epidemiologic model for health interventions. Theor Popul Biol [Internet]. 1996;50(3):209–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9000488
8. Brunet RC, Struchiner CJ. A Non-parametric Method for the Reconstruction of Age- and Time-Dependent Incidence from the Prevalence Data of Irreversible Diseases with Differential Mortality. Theor Popul Biol [Internet]. 1999;56(1):76–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10438670%5Cnhttp://linkinghub.elsevier.com/retrieve/pii/S0040580999914156
9. 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.
10. Konikoff J, Brookmeyer R. Sample Size Methods for Estimating HIV Incidence from Cross-Sectional Surveys. 2015;18(3):386–92.
11. 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
12. 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.” 2021 [cited 2021 Sep 17]; Available from: https://www.researchsquare.com/article/rs-871044/latest.pdf
13. Mhlanga L, Grebe E, Welte A. Population Simulation [Internet]. 2021. Available from: https://rdrr.io/github/laurettemhlanga/PopulationSimulation/
14. Mhlanga L, Grebe E, Welte A. Notes on the age/time structured population simulations., forthcoming
15. Kassanjee R, Mcwalter TA, Bärnighausen T, Welte A. A new general biomarker-based incidence estimator. Epidemiology. 2012;23(5):721–8.
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

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.

Conflict of Interest Statement

The authors declare no competing interests.

Figures

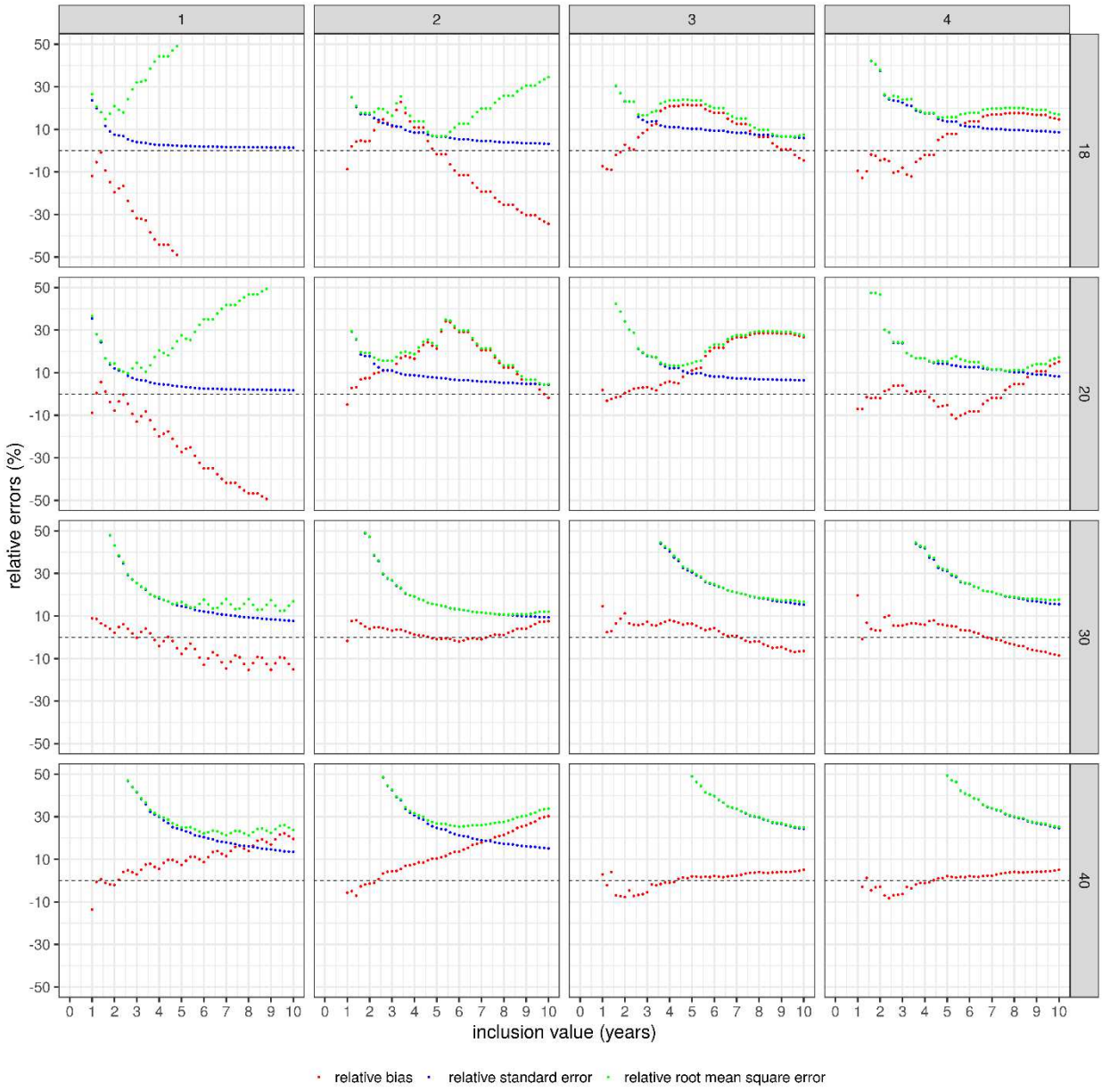

Figure 1. *Relative errors, as a function of the inclusion distance, for various ages and choices of the polynomial order of the prevalence fitting function*. The plot shows the relative bias (green), relative standard error (red), and relative root mean square error (blue) for estimates of incidence 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.

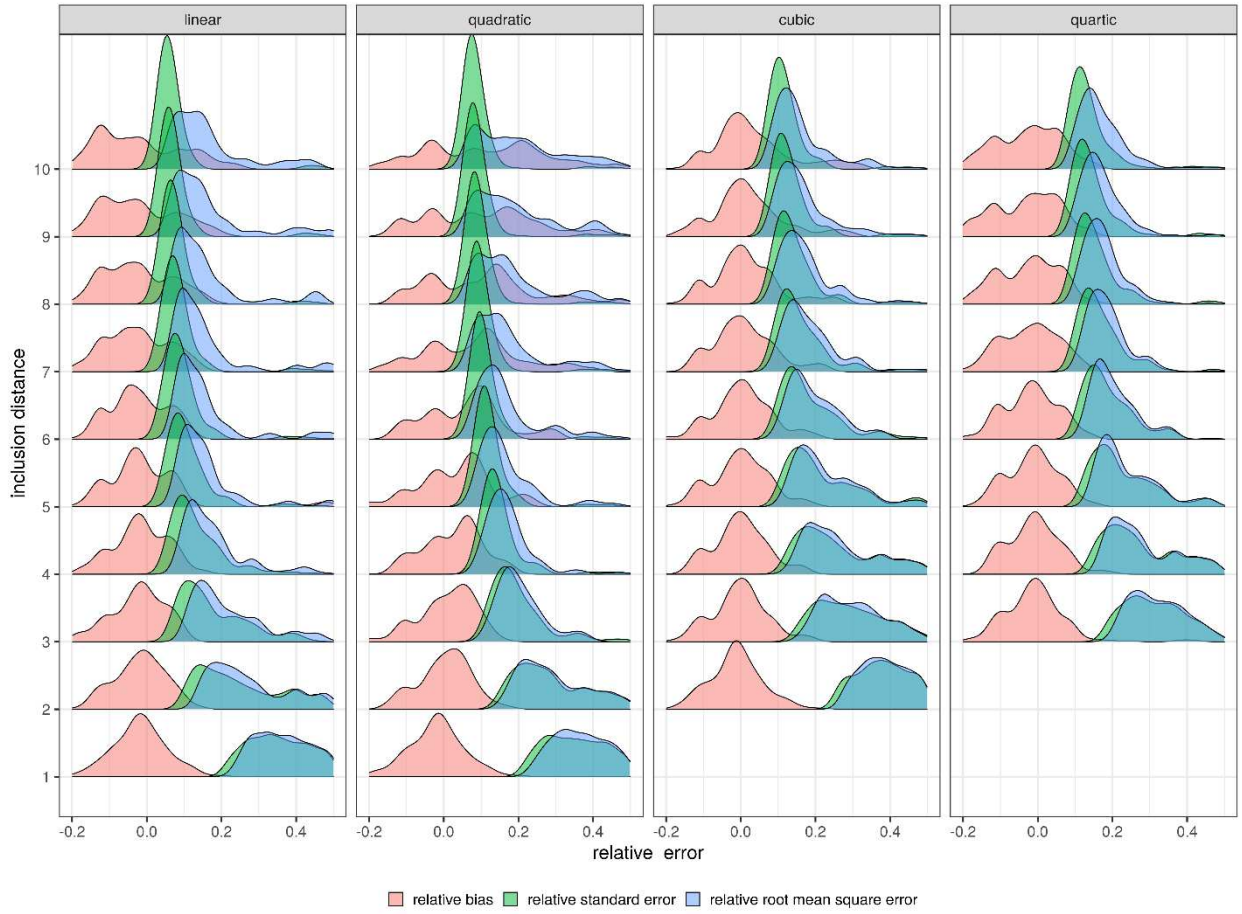

Figure 2: *Distributions of relative errors in incidence estimates, arising over the range of integer ages and inter-survey midpoint times in the standard canonical epidemiological scenario.* The plots show the relative bias, relative standard error and relative root mean square error. The inter survey 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.

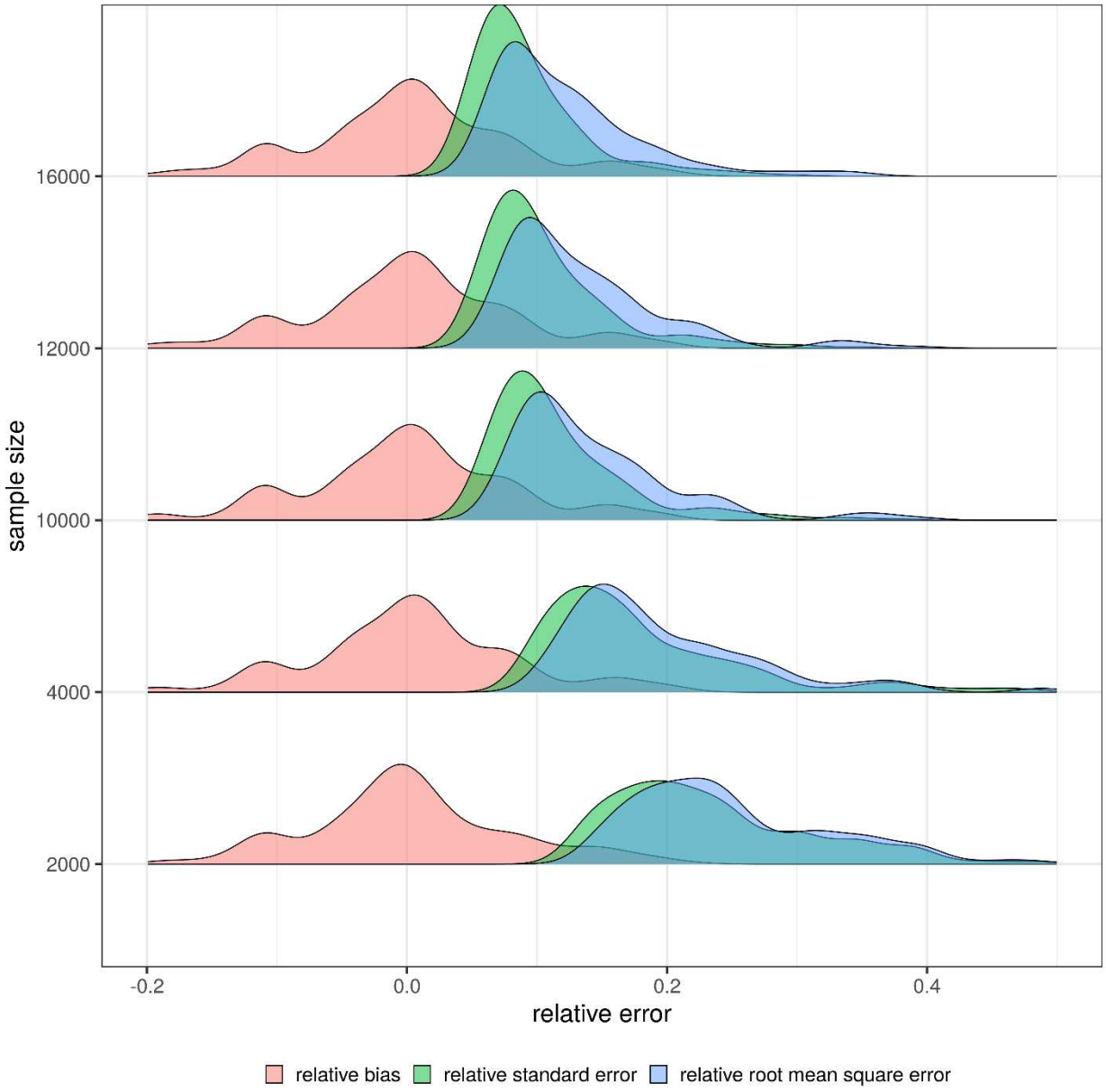

Figure 3. *Density plots of the overall errors for varying sample sizes.* The effect of sample size on the incidence estimates is summarised by the distribution of the relative errors. The sample sizes varied are 2000, 4000, 10000, 12000, and 16000 per 5 year survey bin and the inter survey interval is 5. The distributions are based on 248 data points (8 timepoints for all ages 15:45).

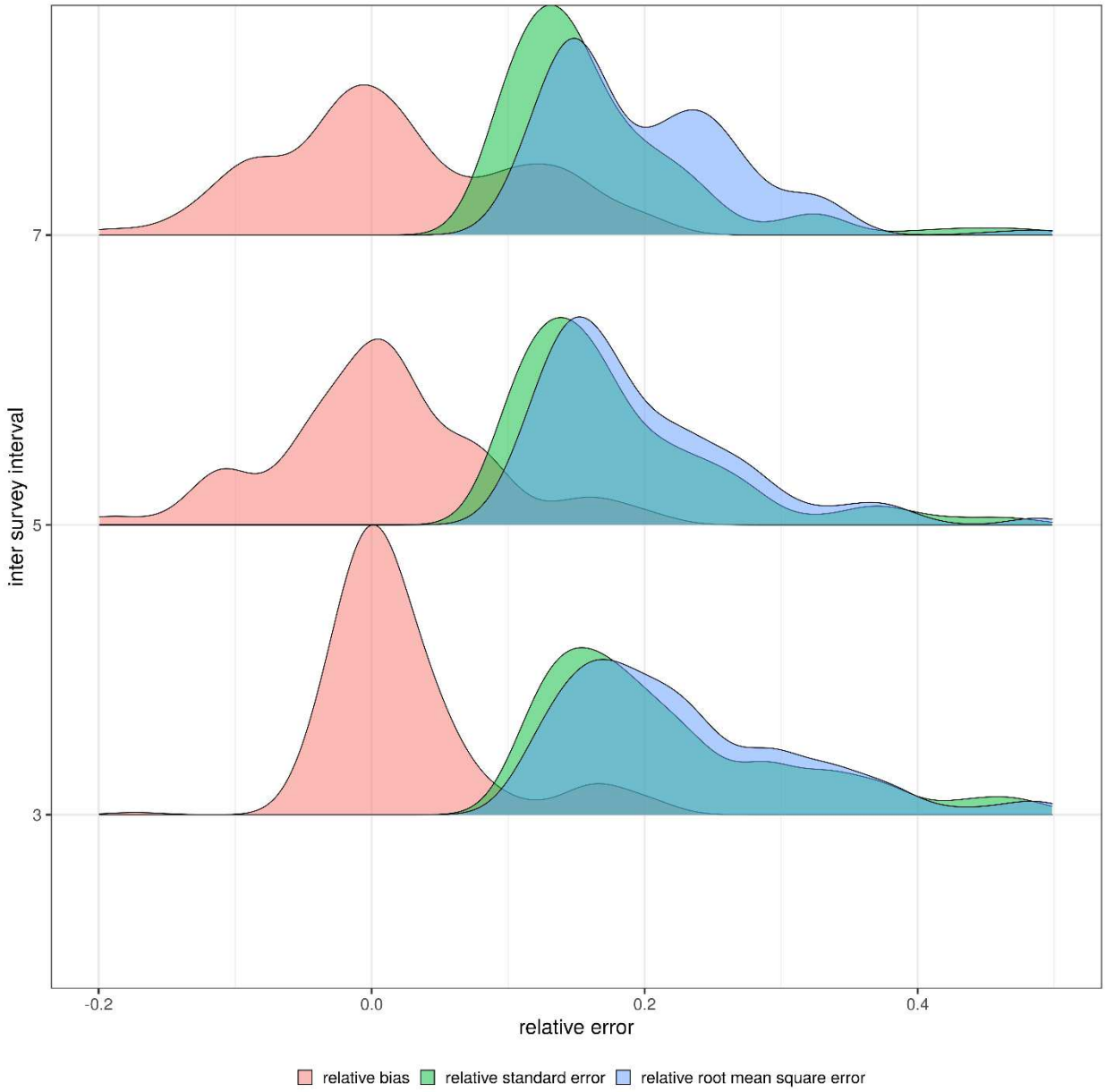

Figure 4. *Density plot of the relative errors (bias, standard error, and root mean square error) for 3 different inter-survey intervals (3, 5, and 7 years*). The Figure shows the distribution of the relative errors for a range of simulated pairs of survey each with a sample of 4000 per 5 year age bin. Pairs of surveys are simulated 3, 5, and 7 years apart, around points 1994.5, 1997.5, 2000.5, 2002.5, 2007.5, 2010.5, 2012.5, 2017.5. The plot is based on a total of 248 data points (ages 15 to 45, for each of the 8 time points).

Table 1: *Bootstrap (10000 samples) versus delta method standard errors for prevalence, gradient of the prevalence and incidence.* The table shows the prevalence, gradient of the prevalence and incidence's standard errors, for selected ages 18, 20, 30, and 40 at time 2017.5.

	Prevalence Standard Error (%)			**Prevalence gradient standard error (% p.a.)**			**Incidence standard error (% p.a.)**		
Age	**Bootstrap**	**Delta Method**	**Ratio**	**Bootstrap**	**Delta Method**	**Ratio**	**Bootstrap**	**Delta Method**	**Ratio**
18	0.289	0.286	1.01	0.209	0.233	0.900	0.218	0.227	0.96
20	0.400	0.414	0.96	0.231	0.231	1.00	0.248	0.239	1.04
30	0.518	0.514	1.01	0.327	0.326	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

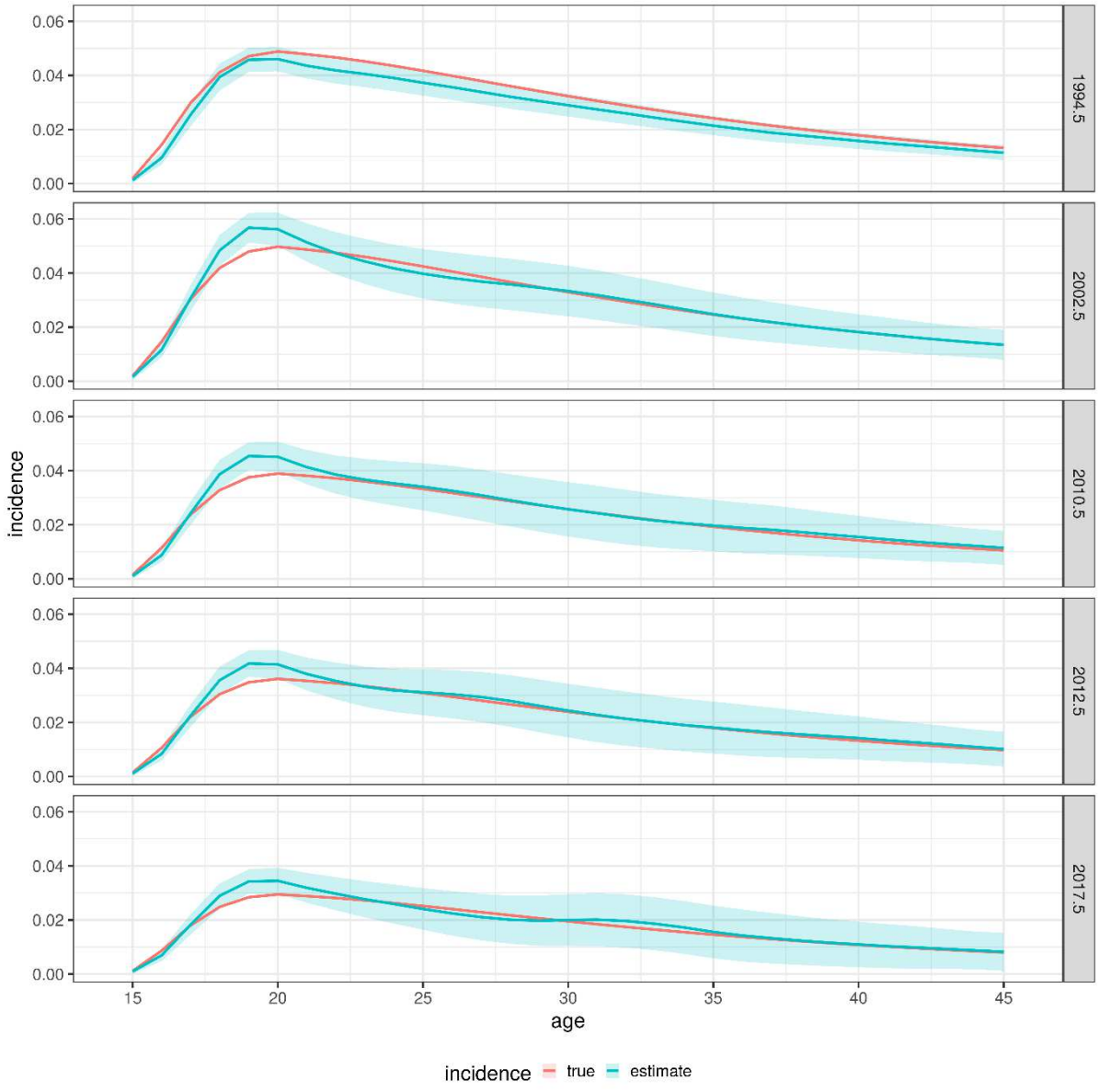

Figure 4. *Midpoint incidence estimates for selected epidemic stages in (1992, 1997), (2000, 2005), (2008, 2013), (2010, 2015), and (2015, 2020) with inter survey interval 5*.

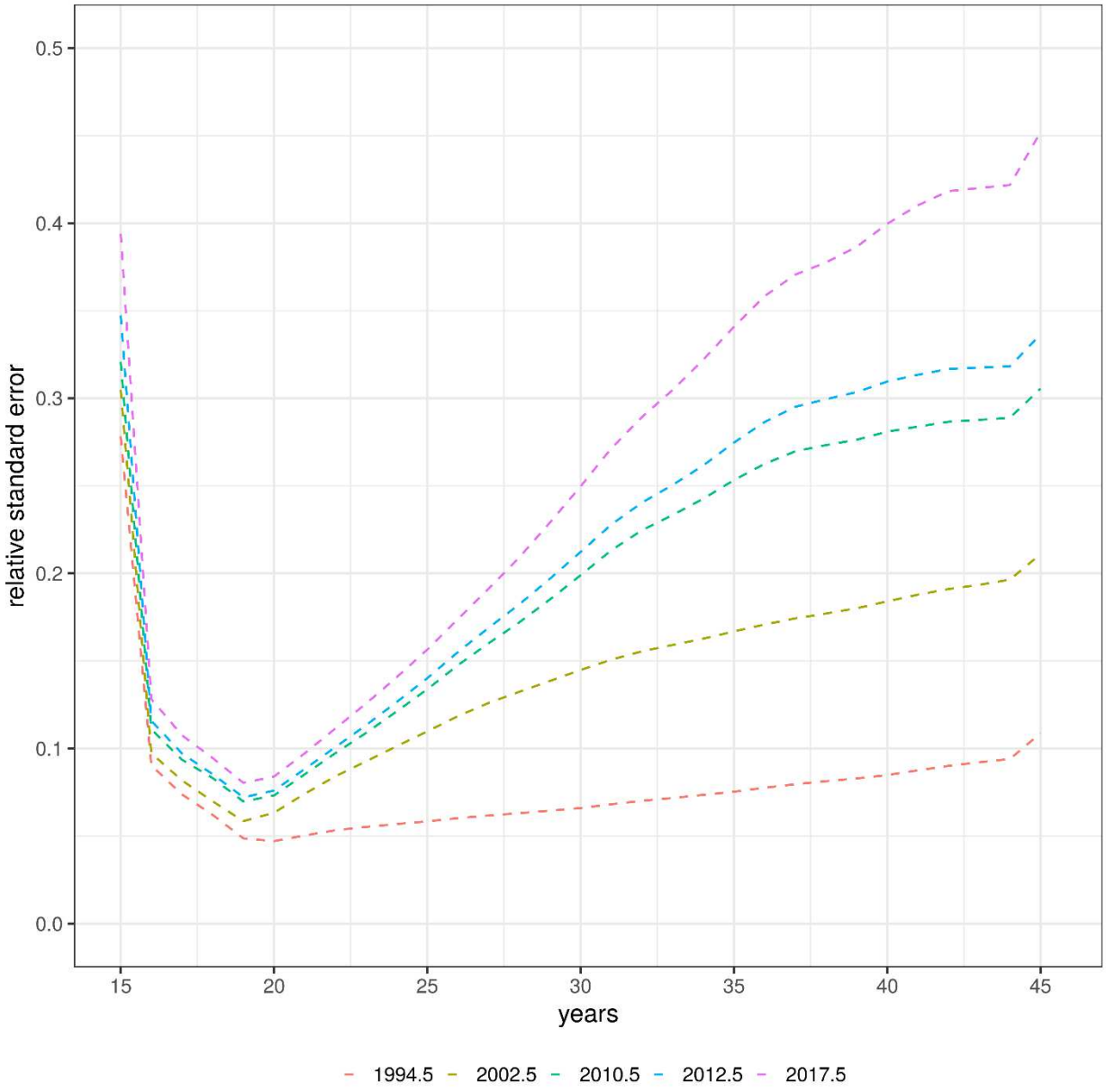

Figure 5. *Relative standard error of the midpoint incidence estimates for selected epidemic stages.* The Figure shows relative standard errors for a range of simulated pairs of survey each with a sample of 4000 per 5 year age bin. The surveys are simulated 5 years apart and the corresponding midpoints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

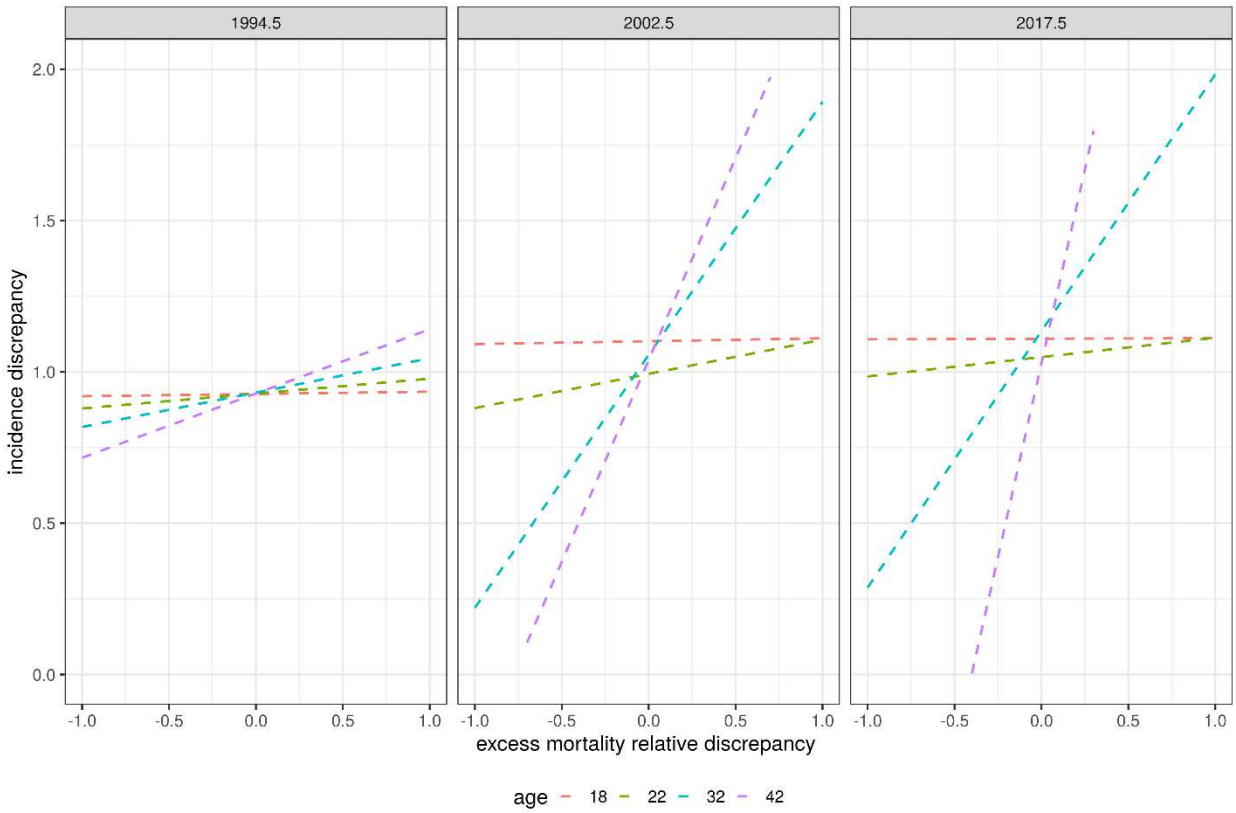

Figure 6. *Sensitivity analysis of the excess mortality's discrepancy ratio.* The picture depicts the bias of the incidence as a function of the bias in the excess mortality for selected ages at different epidemic stages.

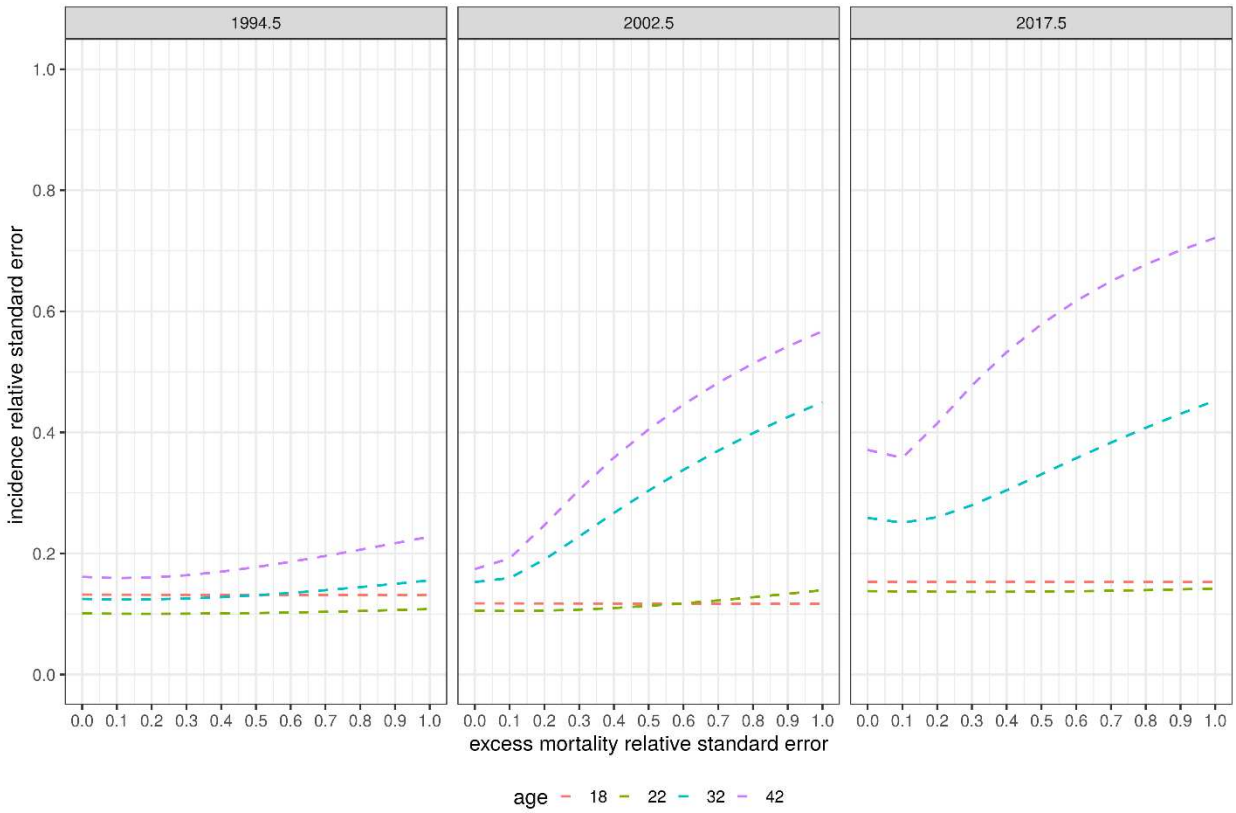

Figure 7. *Sensitivity analysis of the excess mortality's standard error.* The Figure depicts the relative standard error of the incidence estimate at selected ages 18, 22, 32, and 42 as functions of the relative standard error of the excess mortality, each age represents a specific epidemic stage for example age 18 represents an early epidemic state, the time the incidence is rapidly rising and there is not much excess (disease induced) mortality. The analysis is based on a pair of cross-sectional surveys simulated in (2015, 2020).

Supplementary Files

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

- AppendixoptimisingMahiane.pdf