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

Laurette Mhlanga (laurette@aims.ac.tz)
SACEMASouth African DST-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University

Eduard Grebe
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 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-996585/v2

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

(running head: The value of recency data)

Laurette Mhlanga, Eduard Grebe, Alex Welte

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 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.

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 (1–4).

In two companion articles (5,6), we have systematically explored the optimal extraction of this age/time structure from population survey data. To recap:

- 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 separately (7).
- We proposed a generic approach to age/time regression in order to use the approach of Kassanjee et al (8), 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. (9), 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 – an approach 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 that are however correlated in complex ways as they both rely on the same underlying serostatus data which always comprises the bulk of the data set. For the present analysis, 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 ascertainment, which invariably imply substantial increases in both the cost and the complexity of surveys that are already major undertakings even without this requirement.

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

1. Simulating demonstrative epidemics, defined by incidence and mortality, leading to an emergent (age-, time- and time-since-infection- structured) population state.
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 extract age and time specific estimates of prevalence of HIV, and prevalence of 'recent infection' amongst HIV positives.
4. Estimating incidence, and incidence differences/trends, from these smoothed functions, using the Kassanjee and Mahiane frameworks - both separately, and in conjunction.

5. Evaluating the relative merits of the various combinations of approaches – which we are able to do by comparing the estimates with the known incidence parameter values which were used in the simulations.
6. Proposing guidance on the use of, and value added by, 'recent infection' ascertainment (for the purpose of HIV incidence estimation).

Methods

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 (9) and Kassanjee et al. (8), 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

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

$$I_K = \frac{P(R-\beta)}{(1-P)(\Omega - \beta \cdot T)} \qquad 2$$

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' (a.k.a. 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:

$$\mathrm{I_{Opt}} = W \cdot I_M + I_K \cdot (1 - \mathrm{W}) \qquad 3$$

$$se(I_{Opt}) = \sqrt{\mathrm{W}^2 \cdot \sigma_{I_M}^2 + (1-\mathrm{W})^2 \cdot \sigma_{I_K}^2 + 2 \cdot W \cdot (1-W) \cdot \mathrm{CoV(I_M, I_K)}} \qquad 4$$

We find the optimal weight by differentiating equation 4 with respect to W and setting that to zero:

$$\mathrm{W} = \frac{\sigma_{\mathrm{I_K}}^2 - \rho \cdot \sigma_{I_K} \cdot \sigma_{I_M}}{\sigma_{I_M}^2 + \sigma_{I_K}^2 - 2 \cdot \mathrm{CoV(I_M, I_K)}} \qquad 5$$

Where, $\sigma_{I_M}^2$ is the variance of I_M, $\sigma_{I_K}^2$ is the variance I_K and $\mathrm{CoV(I_M, I_K)}$ is the covariance of I_M, and I_K. According to delta method analysis (10,11) the $COV\ (I_M, I_K)$ is given by;

$$COV\ (I_M, I_K) = \frac{\partial I_M}{\partial P} \cdot \frac{\partial I_K}{\partial P} \cdot \sigma_P^2$$

The derivatives of the estimators are given by

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

and

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

$$COV\ (I_M, I_K) = \left[\left(\frac{P(R-\beta)}{(1-P)^2(\Omega-\beta\cdot T)}\right) + \cdot \left(\frac{(R-\beta)}{(1-P)(\Omega-\beta\cdot T)}\right)\right] \cdot \left[\frac{dP}{(1-\ P)^2} + M\right]\sigma_P^2 \qquad 6$$

The covariance can be estimated either by equation 6, 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.

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).

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 (12–14). Hence, to avoid this distraction for the present purposes, surveys are simulated with the same recency test.

Incidence trends.

Incidence trends are a crucial indicator of whether interventions or emergent changes in habits and services are reducing the transmission of HIV. To investigate the prospects for estimation of an incidence trends 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 $\frac{\partial}{\partial t}P$ in both I_M estimates.

Results/Discussion

Single cross-sectional survey

Figure 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 size (or sampling density) is likely to be smaller, and the bias shown here may be substantially swamped by poor precision.

Midpoint incidence estimates comparison (I_K,I_M and I_{Opt}).

Figure 2 and Figure 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.

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 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 obstacle to implementing an ultimately more robust bootstrap based calculation.

Sensitivity of the standard error I_{opt} to W (midpoint)

Figure 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.

Incidence estimates at Survey times

Figure 5 and Figure 6 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 2 and Figure 3, probably because incidence is being estimated where the data points actually is, 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.

Incidence trends

Two surveys

Our attempts to estimate incidence trends/difference from two cross sectional surveys, using all 3 approaches I_M, I_K, and I_{Opt} are shown in Figure 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.

Three surveys

Figure 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 post-hoc age averaging (see Figure 9) which we previously introduced in our companion piece (15) based on two cross sectional survey with recency ascertainment. Figure 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.

Conclusion

In our preceding companion pieces, we explored the fine points to consider when estimating P, dP/dt, and R for use in each of the incidence estimators Mahiane et al., (9) and Kassanjee et al., (8). 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 (6), 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.
- *Multiple estimates of recency test properties:* 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.

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.

References

1. William K. Maina, Andrea A. Kim, Rutherford GW, Harper M, K'Oyugi BO, Sharif; S, et al. Kenya AIDS Indicator Surveys 2007 and 2012: Implications for Public Health Policies for HIV Prevention and Treatment William. J Acquir Immune Defic Syndr. 2014;66(Suppl 1):1–14.

2. DHS. DHS Methodology [Internet]. 2017 [cited 2021 Oct 7]. Available from: https://dhsprogram.com/What-We-Do/Survey-Types/DHS-Methodology.cfm

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. Mhlanga L, Grebe E, 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

6. Mhlanga L, Grebe E, Welte A. Smoothing age/time structure of HIV prevalence, for optimal use in synthetic cohort based incidence estimation. 2021 [cited 2021 Oct 18]; Available from: https://www.researchsquare.com/article/rs-959136/latest.pdf

7. Mhlanga L, Grebe E, Welte A. Notes on the age/time structured population simulations. *forthcoming*

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

9. 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

10. Ku HH, others. Notes on the use of propagation of error formulas. J Res Natl Bur Stand (1934). 1966;70(4):75–9.

11. Oehlert GW. A note on the delta method. Am Stat. 1992;46(1):27–9.

12. 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.

13. Grebe E, Welte A, Johnson LF, Cutsem G Van, Puren A, Ellman T, et al. Population-level HIV incidence estimates using a combination of synthetic cohort and recency biomarker approaches in KwaZulu-Natal, South Africa. Blackard J, editor. PLoS One. 2018 Sep 13;13(9):1–16.

14. Mahy M, Brown T, Stover J, Walker N, Stanecki K, Kirungi W, et al. Producing HIV estimates: From global advocacy to country planning and impact measurement. Glob Health Action [Internet]. 2017;10(1). Available from: https://doi.org/10.1080/16549716.2017.1291169

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

Figures

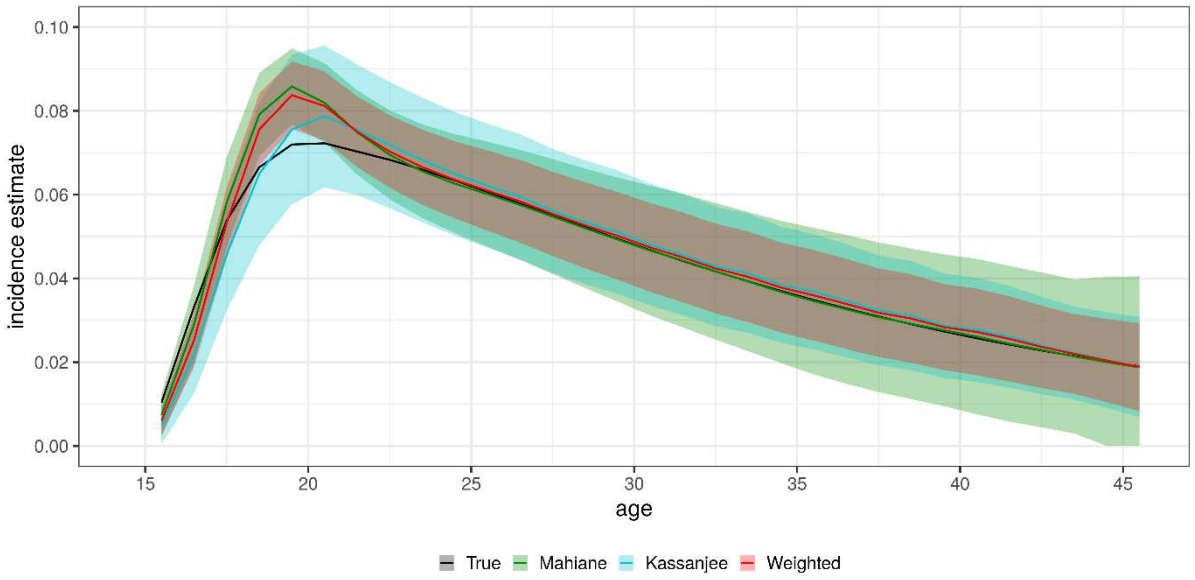

Figure 1: *Incidence estimates from single cross-sectional surveys simulated* 2017.

The plot compares 3 incidence estimates (Kassanjee et al, Mahiane et al., and optimally weighted estimators) to the true (simulated) incidence. Each survey has a sample size of 4000 per 5-year age range with link functions logit for P and c log-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 simulates no treatment.

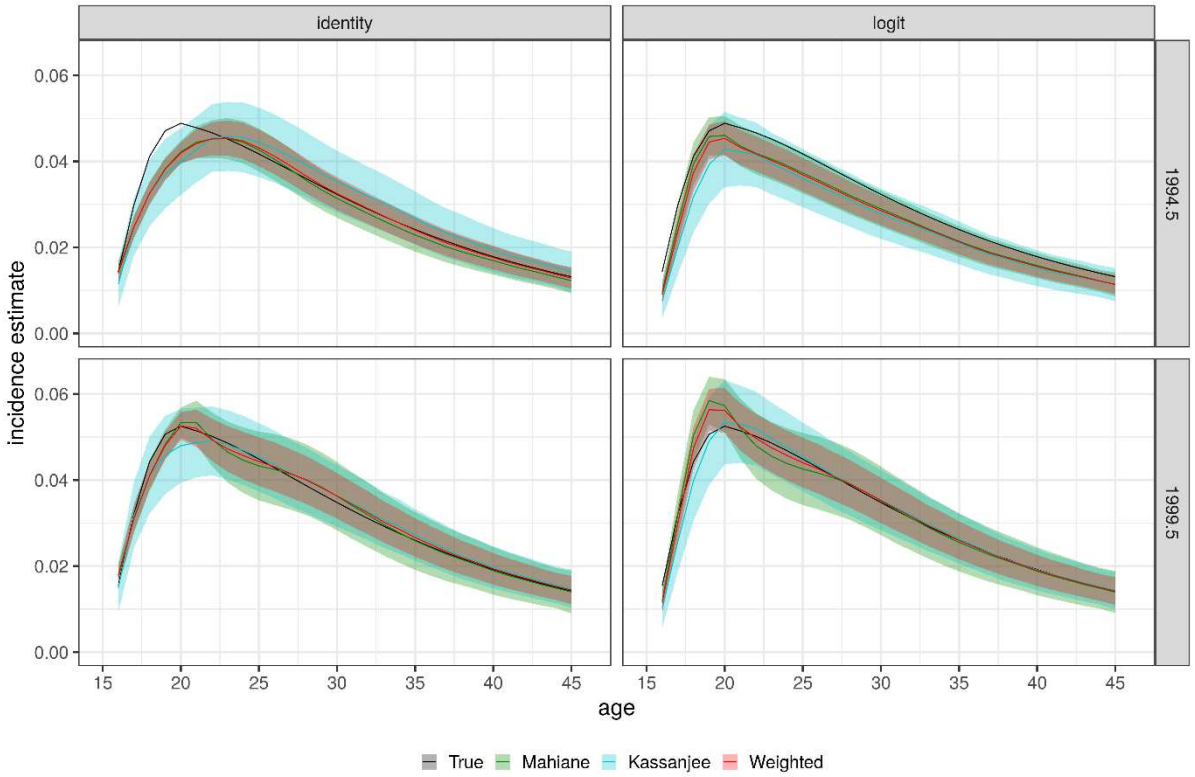

Figure 2: *Midpoint incidence estimates from pairs of simulated cross-sectional surveys (1992, 1997), and (1997, 2002)).*

The estimates are based on the Kassanjee estimator, the Mahiane estimator, and the optimally weighted incidence estimators, as shown. Generous sample sizes of 4000 per 5-year age range were used either with identity or logit link functions for P as shown in column labels. A c log-log link function was used throughout for R. All regressions used a cubic order polynomial (in age, truncated at linear in time because there are only two time points across all observations in any given regression) and an observation inclusion distance of 6 years in the age direction.

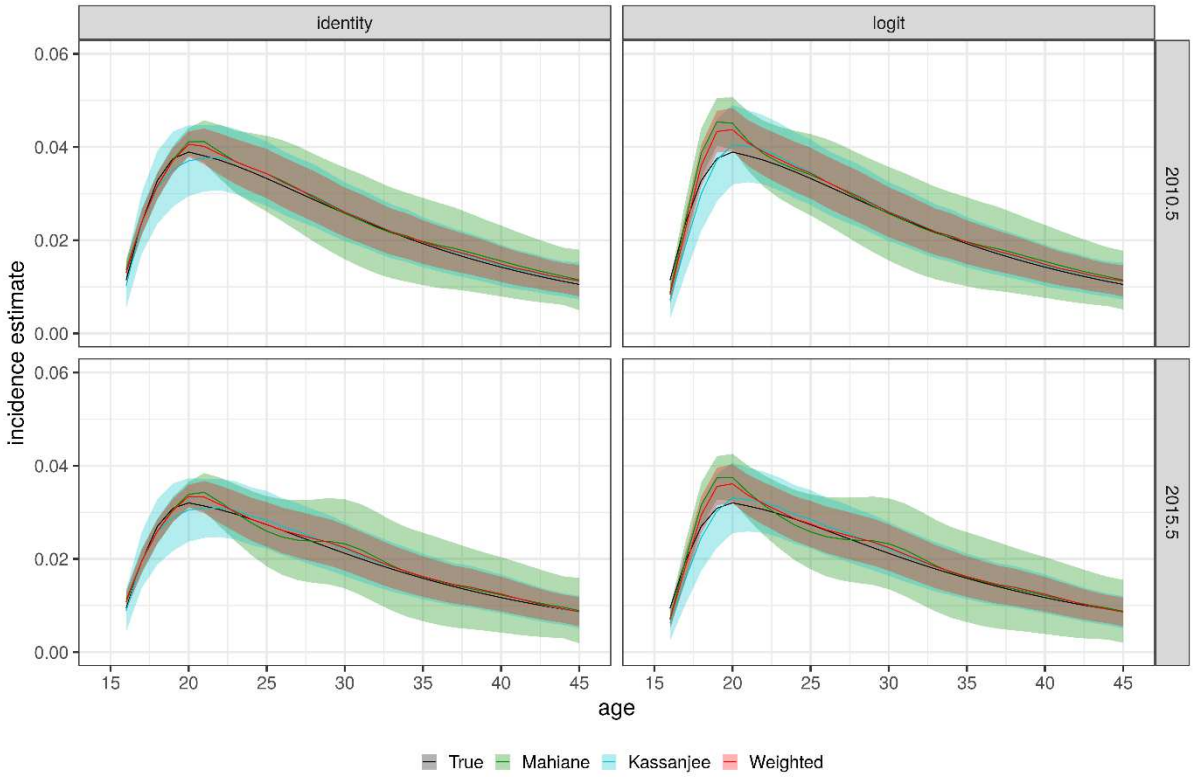

Figure 3: *Midpoint incidence estimates from 5 pairs of simulated cross-sectional surveys (2008, 2013), and (2013, 2017)) estimated using the link functions, identity (left) and logit (right).*

The estimates are based on the Kassanjee estimator, the Mahiane estimator, and the optimally weighted incidence estimators, as shown. Generous sample sizes of 4000 per 5-year age range were used either with identity or logit link functions for P as shown in column labels. A c log-log link function was used throughout for R. All regressions used a cubic order polynomial (in age, truncated at linear in time because there are only two time points across all observations in any given regression) and an observation inclusion distance of 6 years in the age direction.

Table 1: *I_{Opt} estimates derived from two estimates of W_1. W_1– Optimal weight calculated from delta method (analytical function) $COV(I_M, I_K)$ versus W_2 – derived from 10000 bootstrap estimates of I_M and I_K* 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** (% p.a.)	**Bootstrap** (% p.a.)	**Delta Method / Bootstrap Concordance** (% agreement)	**Delta Method** (% p.a.)	**Bootstrap** (% p.a.)	**Delta Method / Bootstrap Concordance** (% agreement)
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	1.10	**99.1**	0.195	0.199	**97.9**

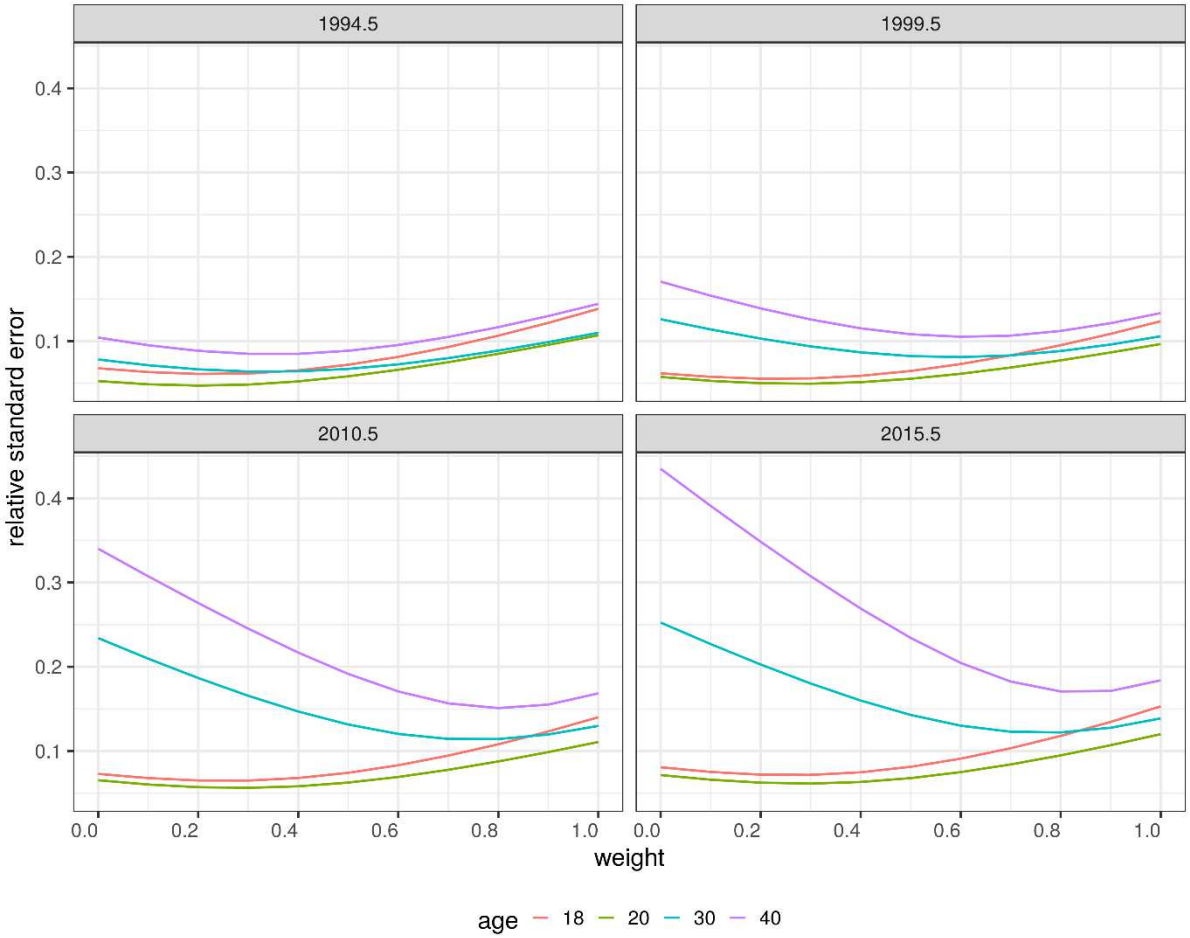

Figure 4: *Relative standard error of the optimally weighted incidence estimator as a function of weights ranging from 0 to 1 weighted to the "Recency" estimator*. The plot shows the relative standard errors for selected ages 18, 20, 30, and 40 at epidemic stages 1994.5, 1999.5, 2010.5, 2012.5, and 2015.5.

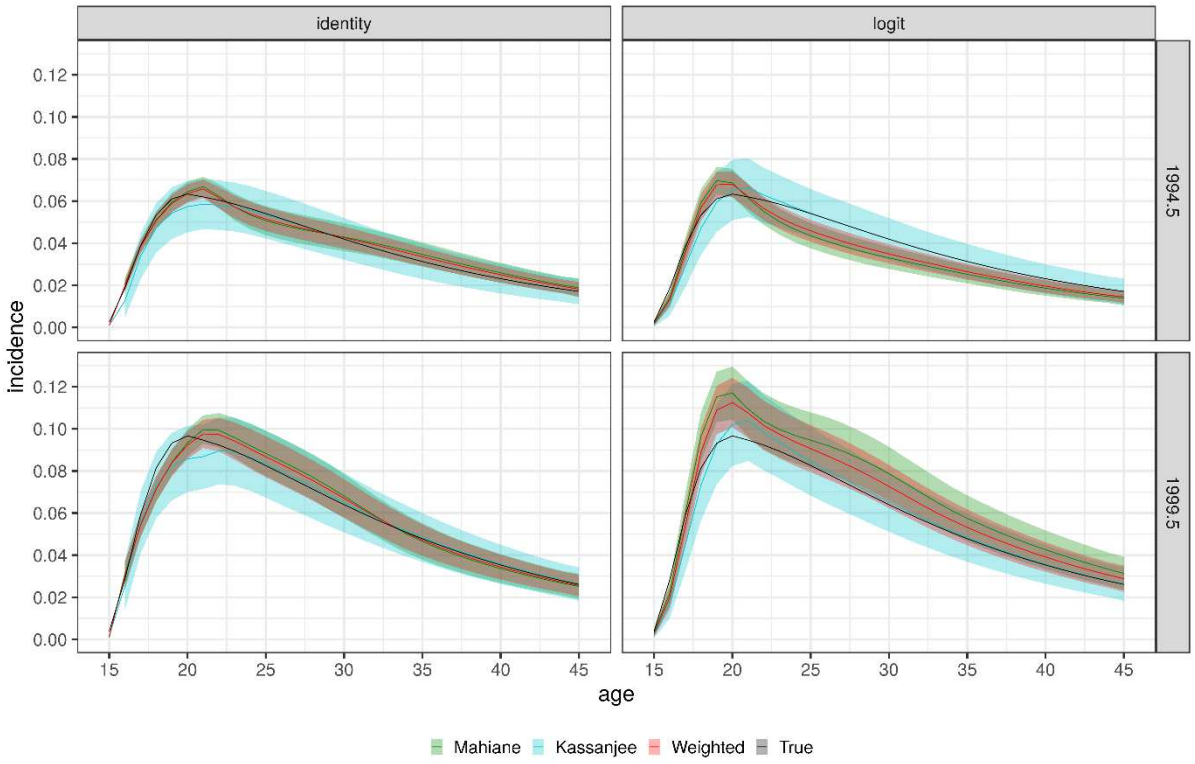

Figure 5: *Incidence estimates at the simulated survey dates – rapidly rising epidemic (1994.5 and 1999.5).*

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 increasing in time (1994.5, 1999.5). The surveys each 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 column depicts the link function (*logit vs identity*) used for the fitting P. R is fitted using a clog-log link function.

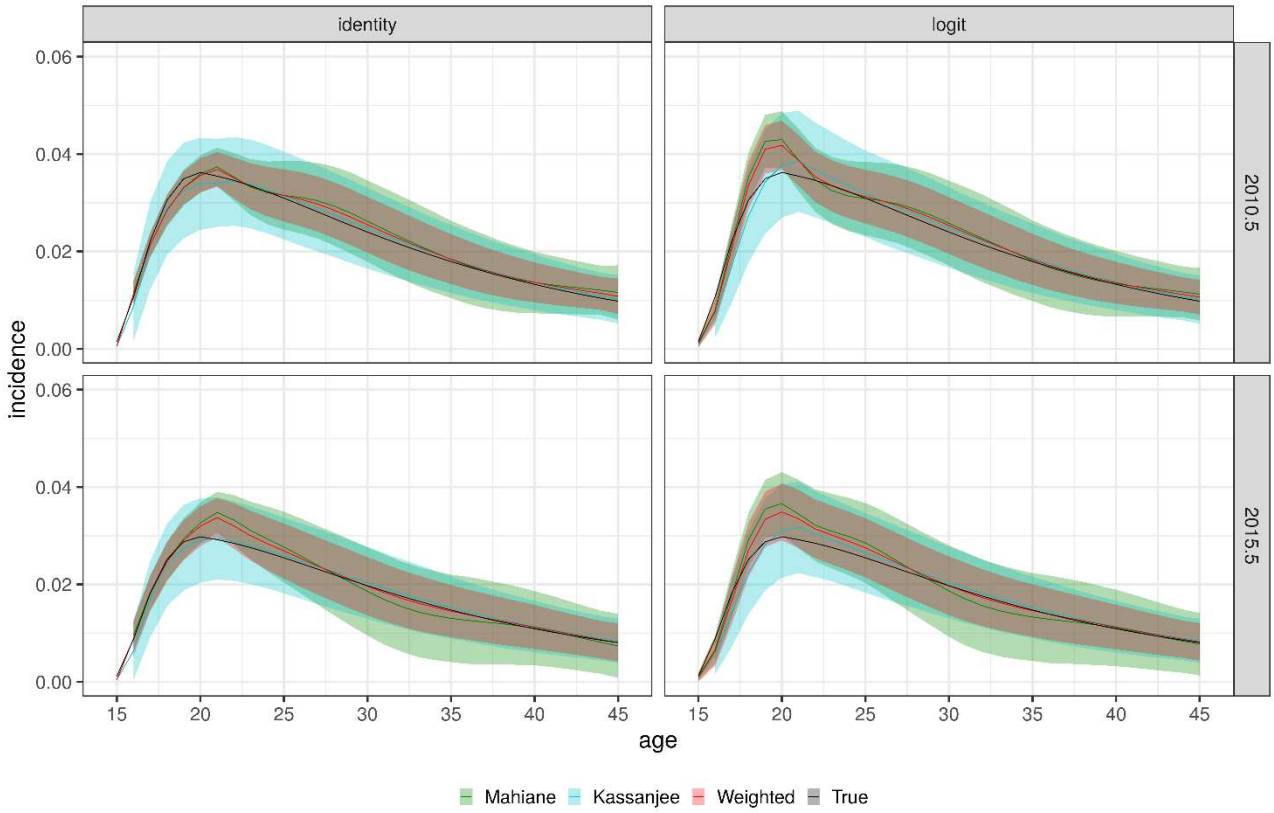

Figure 6: *Incidence estimates at the simulated survey dates steadily declining epidemic (2010.5, 2015.5).*

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 decreasing in time (*2010.5, 2015.5*). The surveys each 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 column depicts the link function (*logit vs identity*) used for the fitting P. R is fitted using a clog-log link function.

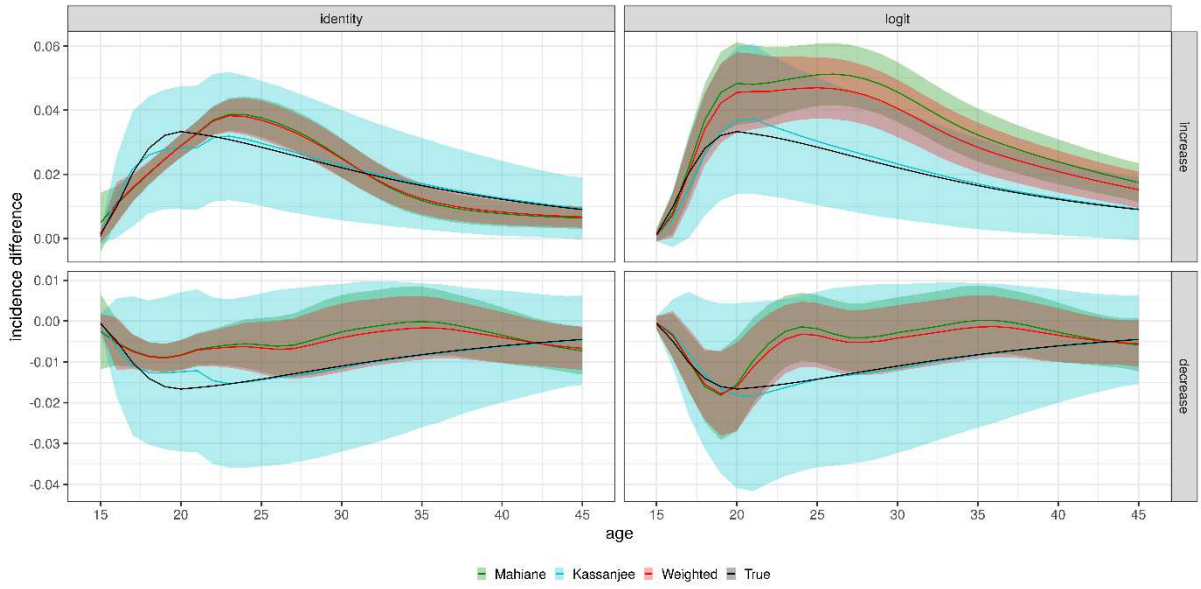

Figure 7: *Incidence difference estimate for pairs of surveys simulated (1993, 1998) - a rapid increase and (2010, 2015) –decrease.*

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 a logit /identity link functions (columns) are used to estimate P and clog log link function for R.

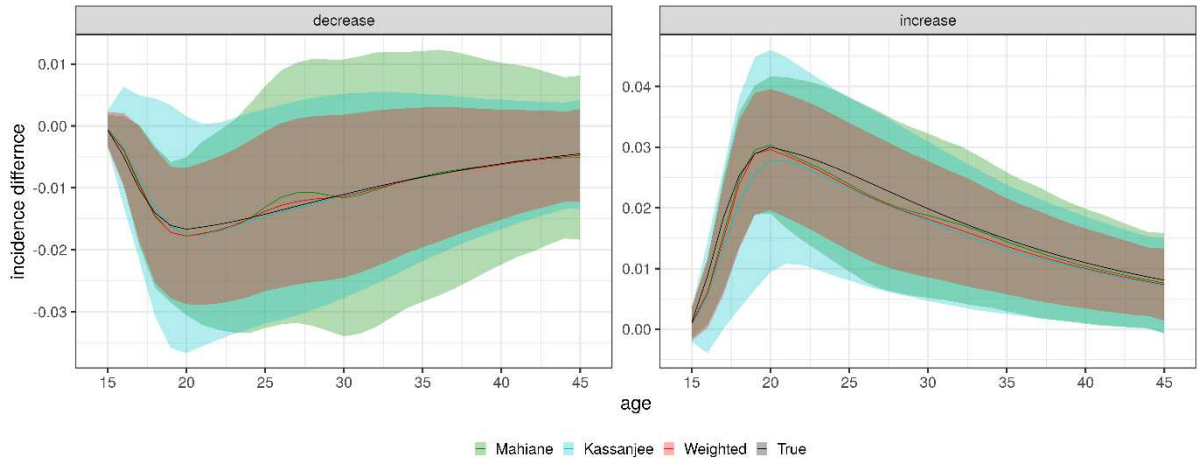

Figure 8: *Incidence difference estimates from midpoint incidence estimates of three cross sectional surveys*.

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

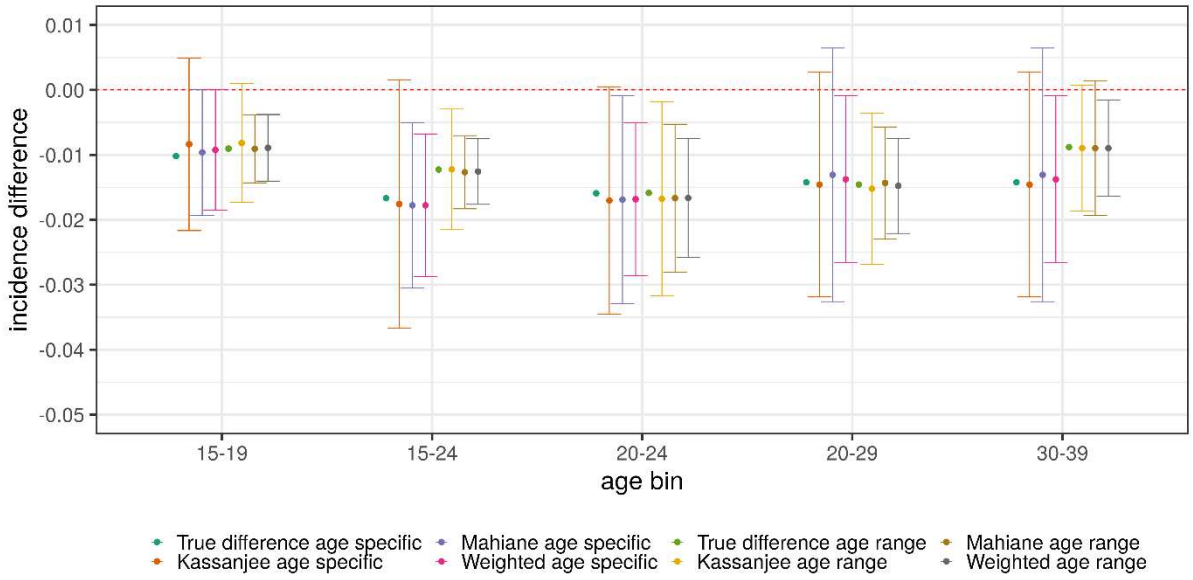

Figure 9: *Post hoc average Incidence difference estimates from midpoint incidence estimates of three cross sectional surveys (2007.5 and 2012.5)*.

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.