Appendix

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

Epidemiological rates in the simulation platform

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

Fertility

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

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

Incidence

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

1. a lognormal term that is only a function of age, with incidence being zero until age 14 (no mother to child transmission) and peaking at age 20 (see figure A1), and
2. a lognormal term that is only a function of time, becoming non zero from 1986, and peaking in 2000 (see figure A2)

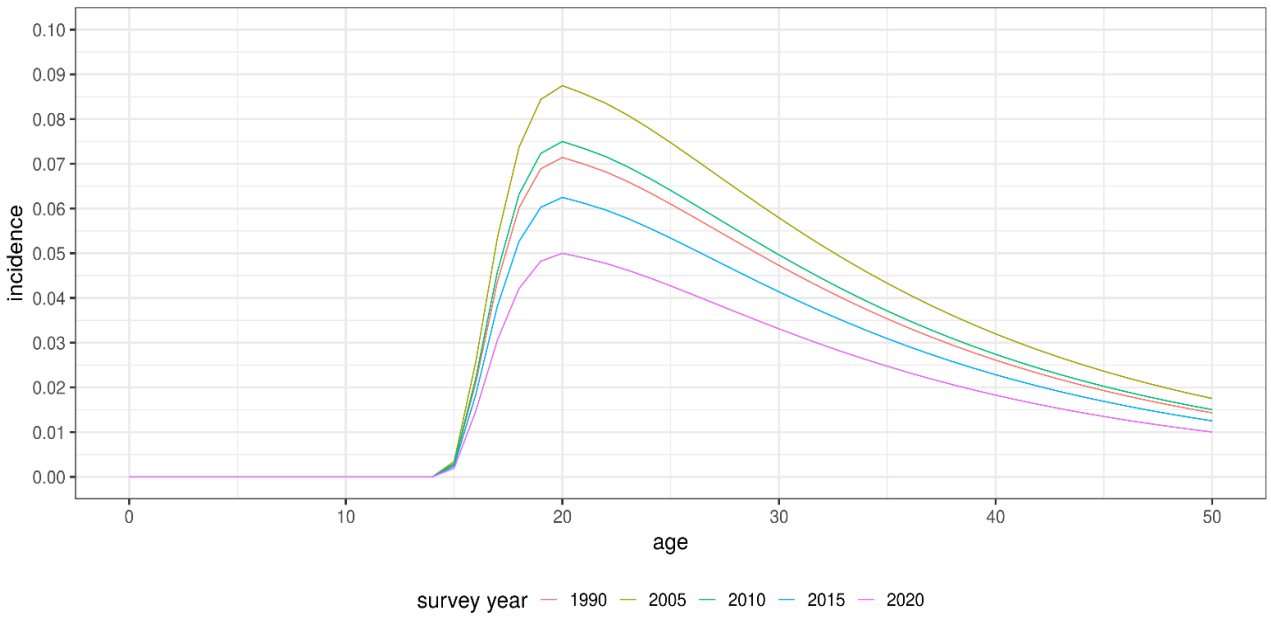

Figure A1: *Incidence as a function of age $I(a)$ at selected times considered in the investigations*

Background mortality

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

Infection-associated mortality

Upon infection, individuals experience an age-at-infection and time-since-infection dependent excess mortality which is a calendar time independent power (2.28) of time-since-infection.

The prevalence which emerges from the interplay of incidence and mortality is summarized in figure (A2)

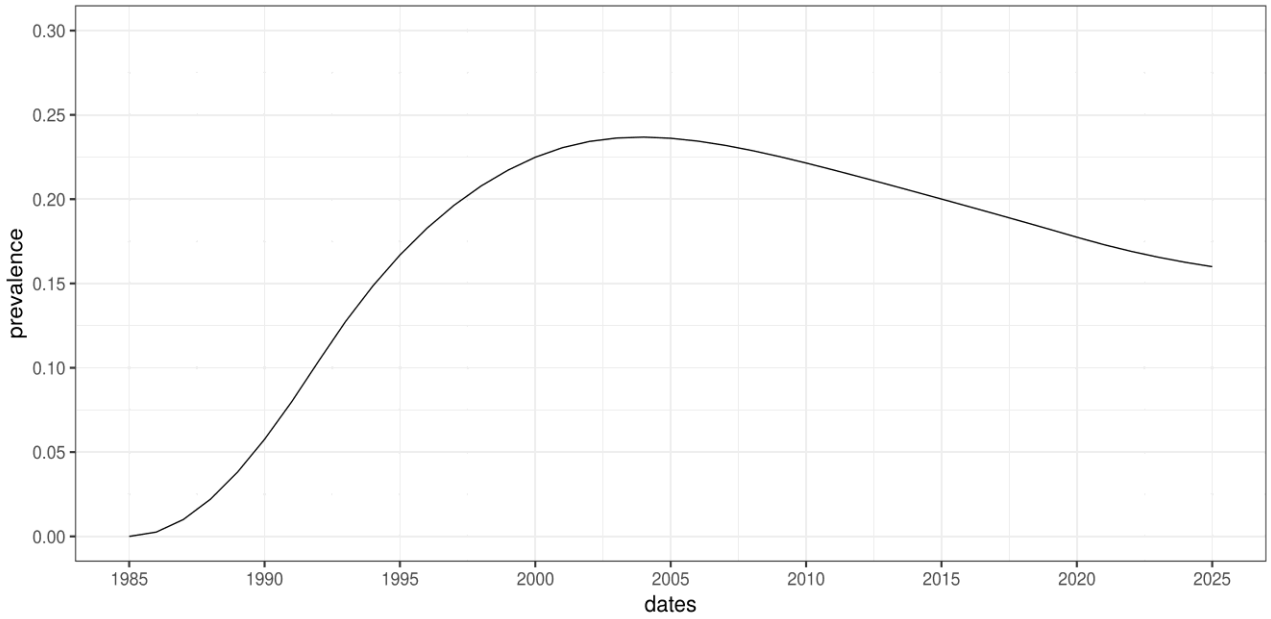

Figure A2: *Age weighted prevalence output as a function of time $P(t)$ for (ages 15-45).*

Recent Infection

After infection, individuals tested for 'recent infection' have a probability of giving the result 'recent' according to a Weibull survival curve with scale factor 0.5 (years) and shape parameter 5. This leads to a mean duration of recent infection (MDRI) of 167.7 days and a negligible false recent rate (FRR). This simplifies all our analysis by freeing us from the real-world problem of estimating the FRR.

The interplay of all of the abovementioned parameters leads to a 'prevalence' of recent infection as shown in figure A3. Note that this prevalence is only defined among HIV positives, not over the entire population.

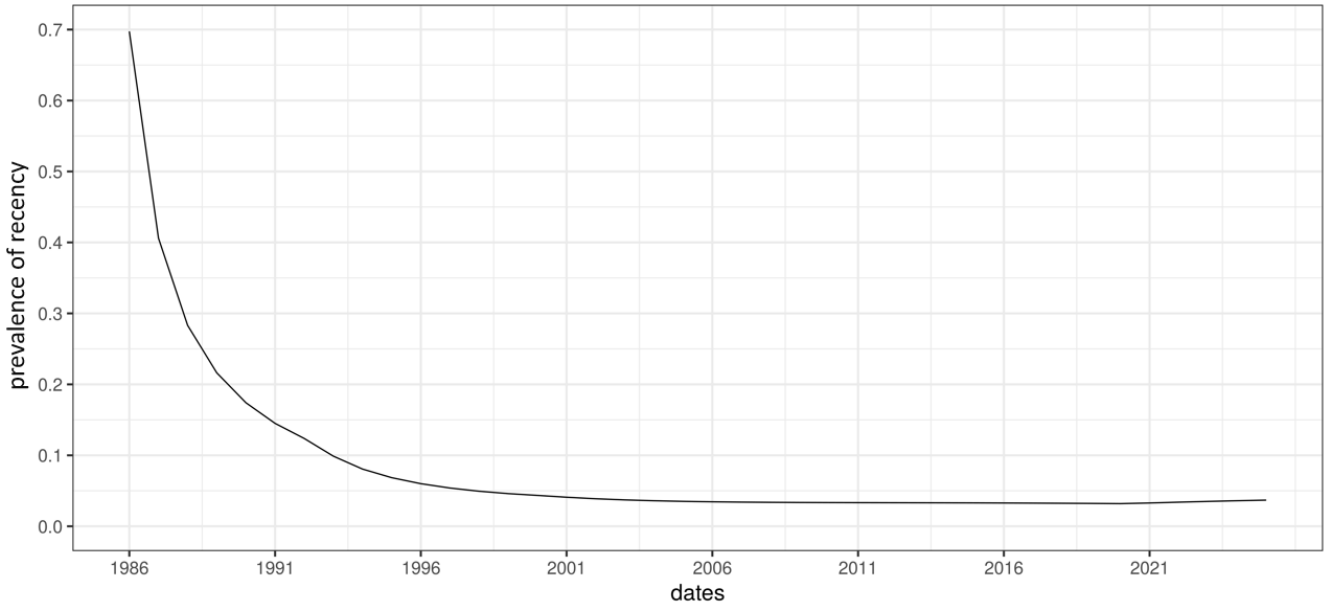

Figure A3: *Age weighted prevalence of recent infection output as a function of time* $R(t)$.

Further error analysis

Figure 5 in the body of the manuscript shows distributions of relative root mean square errors for prevalence, recency and incidence for various choices of survey data fitting parameters (polynomial order and data inclusion age range around age of interest). The following three plots present the underlying distributions of the relative standard errors (red), and relative bias (teal), of, respectively, prevalence (figure A4), prevalence of recent infection (figure A5), and incidence (figure A6). The plots are based a range of simulated epidemic stages (the standard times at which the cross-sectional surveys were simulated) and ages in the range 15 to 45. At each of these times and ages, we varied the polynomial order of the regression formula, as well as the data inclusion distance.

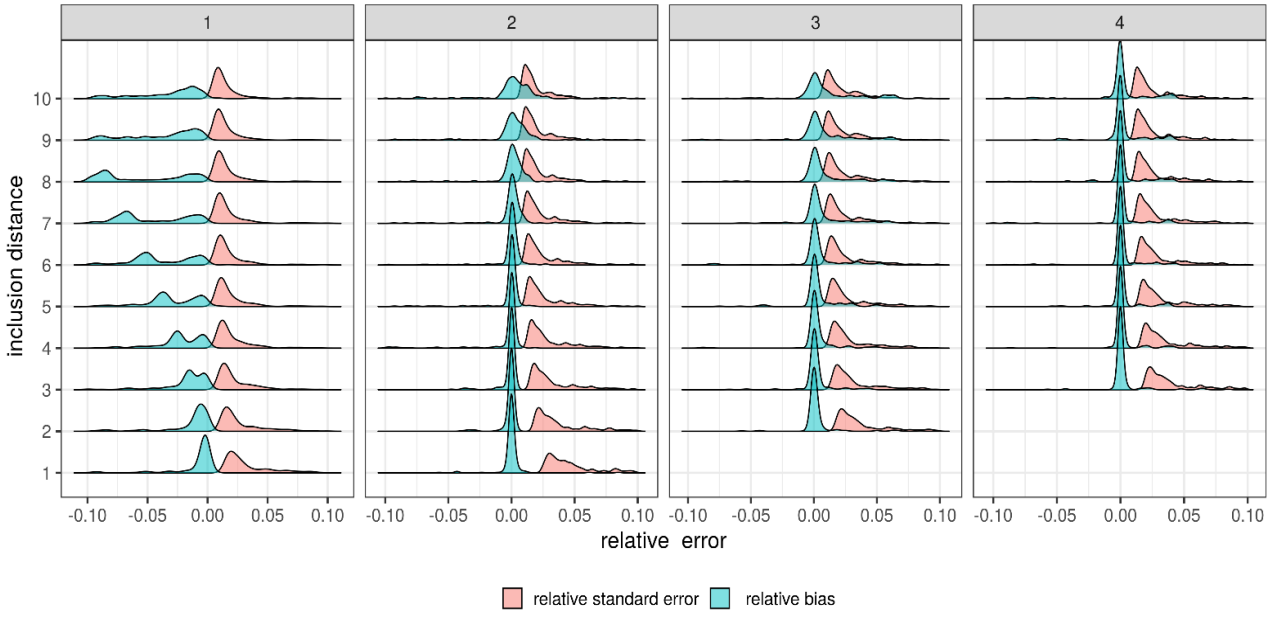

Figure A4: *Relative standard error and relative bias of P(prevalence) estimates using the logit link function*. Each facet represent the 4 polynomial order (1 -4) compared to each other and the x-axis represents the inclusion distance, sample = 4000/5 year age range.

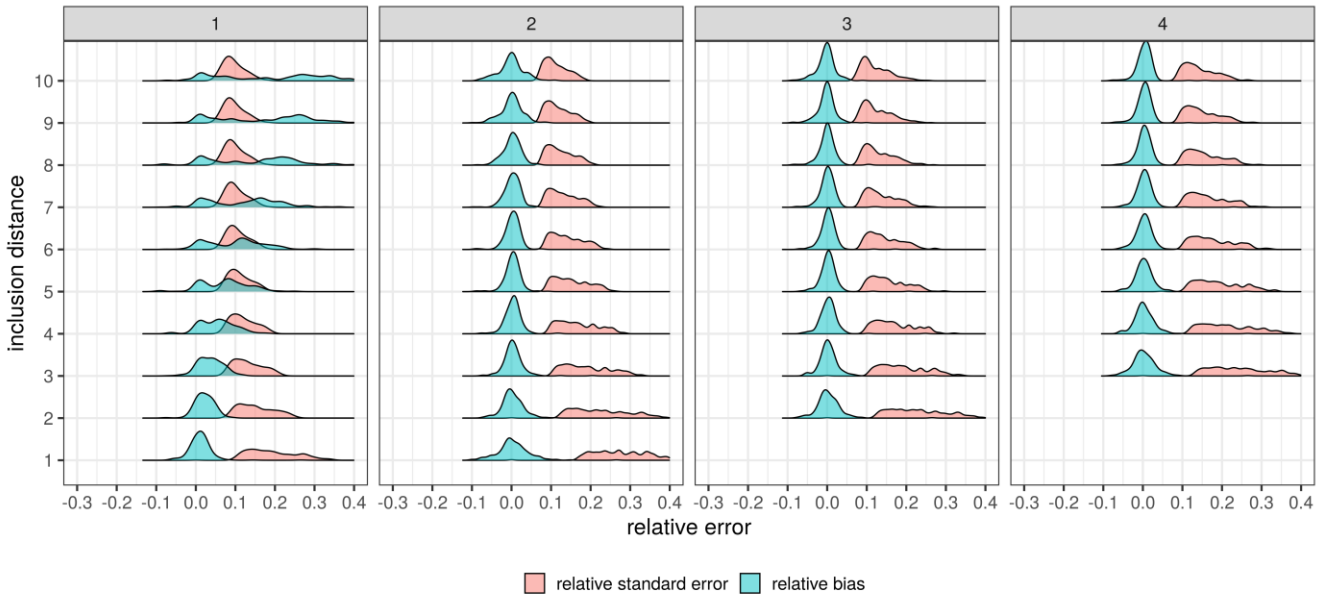

Figure A5: *Relative standard error and relative bias of R (recency) estimates using the clog log link function.* Each facet represent the 4 polynomial order being compared and the inclusion distance is on the x-axis. The sample size was set to 4000 per 5 year age ranges

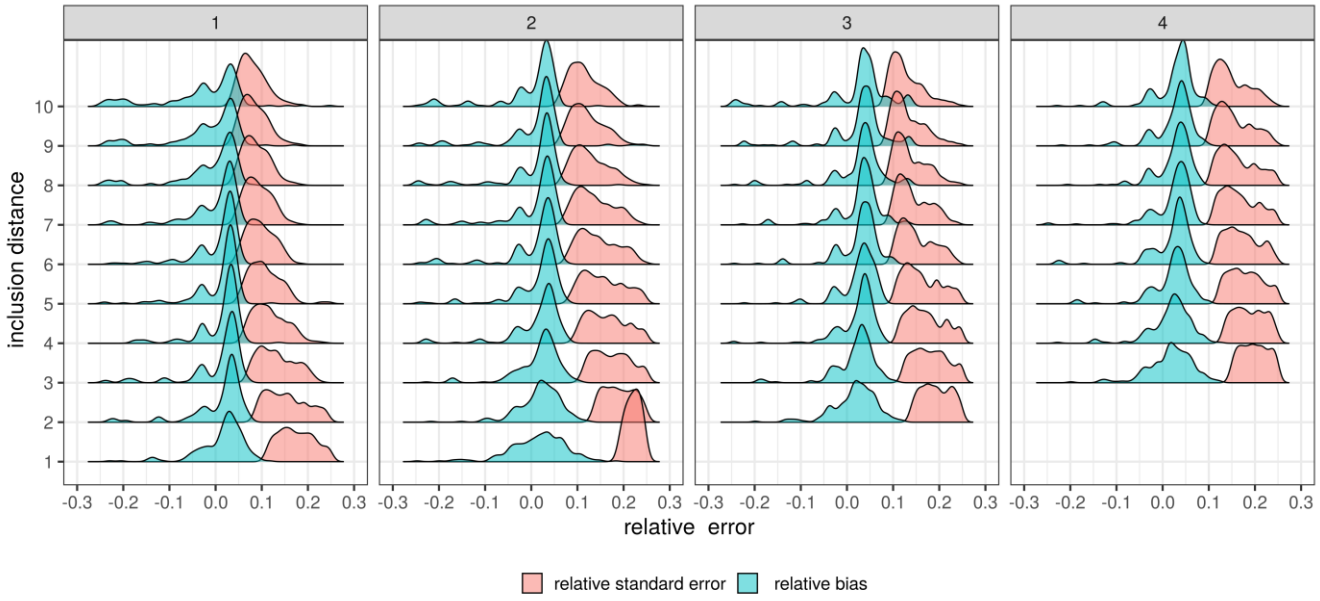

Figure A6: *Relative standard error and relative bias for **incidence estimates** disaggregated by polynomial order and the inclusion distance on the x-axis*.

Incidence difference

For the purpose of evaluating the ability to detect incidence differences, we simulated 2 pairs of cross-sectional surveys; the first pair of cross-sectional surveys depicted an epidemic where incidence was rapidly increasing (1993 and 1998 in our canonical scenario); the second pair of cross-sectional surveys portrays an epidemic that is steadily decreasing (2010 and 2015 in our canonical scenario). For each cross-sectional survey, we independently estimated the age specific incidence and for the survey pairs (set 5 years apart) we estimated the incidence differences. For each analysis we varied the sample size to highlight the effect of sample size on yielding informative incidence difference estimates. Figures XX and XX in the body indicate differences in various proposed incidence age-range-averages, and figures A7 and A8 show the underlying detailed integer-age specific estimates

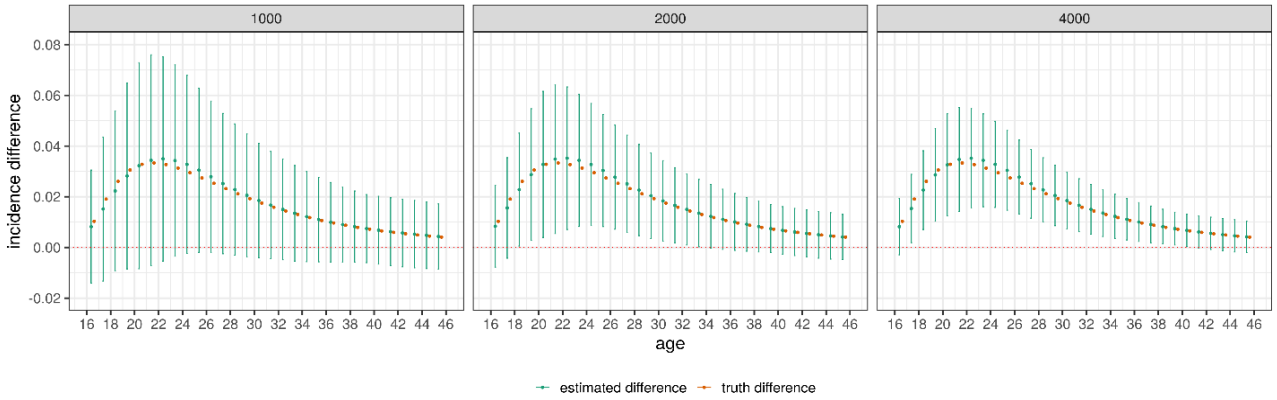

Figure A7: *Incidence difference estimates calculated from two cross sectional surveys simulated in 1993 and 1998 with size of 1000, 2000, and 4000 per 5-year age bin.*

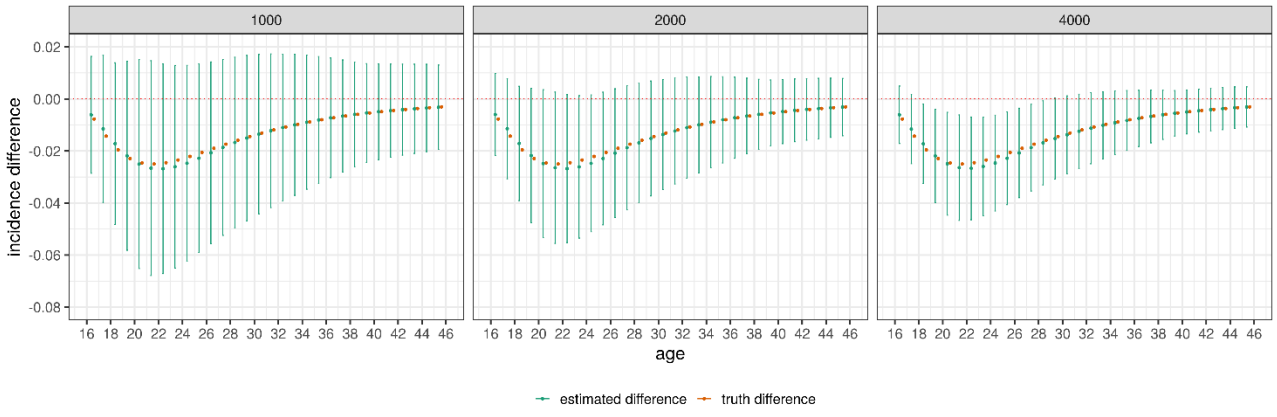

Figure A8 *Incidence difference estimates calculated from two cross sectional surveys simulated in 2010 and 2015 each with sample size 1000, 2000, 4000 per 5-year age bin.*