A method for estimating incidence from annual antenatal prevalence and occasional household prevalence surveys

Introduction/Background

In the early years of the HIV/AIDS epidemic, when mortality was not a serious confounder, tracking prevalence was sufficient to give one an idea of the growth of the epidemic. However, as the epidemic develops and one has to deal with not only the confounding of high mortality due to infection, but more recently with a confounding of that effect by the provision of antiretroviral therapy. Unfortunately measuring incidence is more difficult than measuring prevalence. Tracking a cohort through frequent re-testing is the direct approach and often considered to be the 'gold standard', but is so expensive that it is confined to a few, usually nationally unrepresentative studies which are, in addition, prone to dropout and non-participation which is likely to bias results. Unfortunately the alternatives to a longitudinal study all have their own drawbacks. Laboratory tests are, at present, of doubtful accuracy and require very large samples to reduce the confidence intervals to a small enough range to allow one to assert a change in incidence between two time points. Simulation models on the other hand rely on assumptions, some of which may be questionable, and also contribute to extensive uncertainty around the estimates. Finally, one can resort to estimating incidence from the occasional household prevalence surveys, but at best these provide estimates are an average over the inter-survey period and hence do not provide up-to-date estimates, and also require extremely large surveys before the confidence intervals are small enough to allow the estimates to indicate a change in incidence between two time points.

South Africa runs a large national antenatal survey annually (sample size of 16 000 up to 2005 and double that since then), and it is the purpose of this study to examine the potential for using antenatal prevalence, adjusted to approximate the national prevalence, to track incidence on an annual basis.

Method

The rational of the method is as follows. If one had age-group specific estimates of the prevalence in a population at two time points and an estimate of the proportion of those surviving between these two points in time of those in the age group already infected at the time of the first survey, one can estimate the cohort rate of incidence for the age group, ${}_n\lambda_{x,t}$, by making use of synthetic cohorts as follows, for a one-year time interval, (Hallett, Zaba, Todd *et al.* 2008):

$${}_n\lambda_{x,t} = \frac{{}_nH_{x+1,t+1} - {}_n\pi_{x,t} \cdot {}_nH_{x,t}}{\frac{1}{2}\left[\left({}_nN_{x,t} - {}_nH_{x,t}\right) + \left({}_nN_{x+1,t+1} - {}_nH_{x+1,t+1}\right)\right]}$$

where ${}_nN_{x,t}$ represents the number of people in the population at time *t* aged x to $x+n$, ${}_nH_{x,t}$ represents the number of those people at time *t* aged x to $x+n$ in the population who are estimated to be positive on the basis of testing a representative sample of the population, and ${}_n\pi_{x,t}$ represents the proportion of those infected and aged x to $x+n$ at time *t* surviving to time *t*+1.

Dividing through by ${}_nN_{x,t}$ this can be rewritten as

$${}_n\lambda_{x,t} = \frac{\frac{{}_nN_{x+1,t+1}}{{}_nN_{x,t}}\,{}_nh_{x+1,t+1} - {}_n\pi_{x,t} \cdot {}_nh_{x,t}}{\frac{1}{2}\left[\left(1 - {}_nh_{x,t}\right) + \frac{{}_nN_{x+1,t+1}}{{}_nN_{x,t}}\left(1 - {}_nh_{x+1,t+1}\right)\right]}$$

where ${}_nh_{x,t} = {}_nH_{x,t} / {}_nN_{x,t}$ for all x and *t*.

In order to apply this to data on the prevalence of women aged x to $x+n$ attending public antenatal care clinics, ${}_{n}h^{a}_{x,t}$, one needs to estimate ${}_{n}N_{x+1,t+1}/{}_{n}N_{x,t}$, ${}_{n}\pi_{x,t}$, and ${}_{n}h_{x,t}$. If we assume that the first two terms can be estimated from any reasonably representative population projection model of the population in question (which allows for the demographic impact of HIV), particularly since we are dealing with short (annual) periods of survival, then the only remaining problem is to estimate ${}_{n}h_{x,t}$ from ${}_{n}h^{a}_{x,t}$, or more specifically, ${}_{n}r_{x,t}$ such that ${}_{n}h_{x,t} = {}_{n}r_{x,t} \cdot {}_{n}h^{a}_{x,t}$.

The annual antenatal survey tests women attending public antenatal care clinics for the first visit of their current pregnancy in a sample drawn in proportion to probability at a provincial level (Department of Health (no date)). Age-specific antenatal prevalence rates were derived from unit record data for the years 1998 to 2008[1].

Prevalence rates of all women in the population were estimated from the HSRC numbers tested and numbers positive from the 2005 and 2008 surveys for standard age groups, and for age groups one year older provided by Thomas Rehle (personal communication)[2].

The ratio ${}_{n}N_{x+1,t+1}/{}_{n}N_{x,t}$, which can be expected to be quite close to 1, and the one-year survival probabilities for various age groups of those who were infected at time *t*, ${}_{n}\pi_{x,t}$, were estimated from the ASSA2008 model. (The sensitivity of the results to these assumptions was tested by using the UNAIDS projection for South Africa underlying the estimates of population prevalence and numbers infected included in the Department of Health's annual antenatal survey prevalence report (Department of Health 2009).)

In order to estimate the ratios, ${}_{n}r_{x,t}$, we first estimate the ratio of the age-specific prevalence from the 2005 HSRC survey to the average of the antenatal survey prevalence rates for 2004 and 2005, and the ratio of the age-specific prevalence from the 2008 HSRC survey to that from the antenatal survey for 2008[3]. Similar ratios were calculated annually using estimates of the population prevalence from both the ASSA2008 model and the UNAIDS model, for comparative purposes. As we have estimates of ${}_{n}r_{x,t}$ at only two time points it was decided fit curves to ${}_{n}r_{x,t}$ and ratio of ratios derived from the average of the ASSA and UNAIDS models scaled to pass through the average of the two empirical estimates of ${}_{n}r_{x,t}$ in the middle of 2006, as described in more detail in the appendix.

The estimate of incidence after applying the fitted ratios to the prevalence of women attending public antenatal clinics produced a time series with a marked drop in incidence to almost zero for the year between the 2005 and 2006 survey, before rising again to be in line with the overall trend over time. Since this is exactly the point at which the antenatal survey was expanded substantially (which saw a doubling in the number of women tested and a more than tripling in the number of clinics surveyed) it is possible that the change in prevalence of women attending public antenatal clinics between 2005 and 2006 is due the fact that estimates of prevalence based

[1] In 2000 age was recorded, in error, quinquennially for most women tested and in order to include them in this study the prevalence by individual ages was approximated using Beers ordinary coefficients for sub-division 1976 (Sheyock and Seigel 1976).

[2] Data from the 2002 survey were not used as various inconsistencies between them and the more recent surveys call into question the quality of that survey.

[3] The specific antenatal surveys were chosen to match most closely the average date of the two HSRC surveys. According to the report on the 2005 survey (Shisana et al 2005: 16) some of the 2005 survey took place from October to December 2004, with the bulk from mid-Jan to June 2005, say centred on March 31 2005, and according to the report on the 2008 survey (Shisana et al 2008: 15) most of the testing took place from the end of May 2008 to beginning March 2009, say on average, mid-October 2008.

on surveys prior to 2006 were biased upward. Investigation of the trend by age suggested that this effect was probably limited to ages under 30, and an adjusted set of prevalence rates for the years prior to 2006 was estimated by reducing the prevalence by 7% and 8% in the 15-19 and 16-20 age groups, 5 and 6% in the 20-24 and 21-25 age groups, 2% and 1% in the 25-29 and 26-30 age groups, respectively. This adjustment was chosen so as to produce an estimate of incidence between 2005 and 2006 that was approximately the average of those in the year before and the year after.

Results

Figure 1 presents the ratio of the prevalence of all women in the population in the age groups 15-49, 15-24 and 30-49 to the prevalence of women attending public antenatal clinics in those age groups respectively. The first column shows the ratios based on the antenatal prevalence adjusted for bias in the years before 2006, which results in a slight increase in points prior to 2006, the second column gives the unadjusted ratios. (The triangle represents the ratio of the 2002 HSRC survey to the antenatal prevalence, which was not used in determining the fitted line.

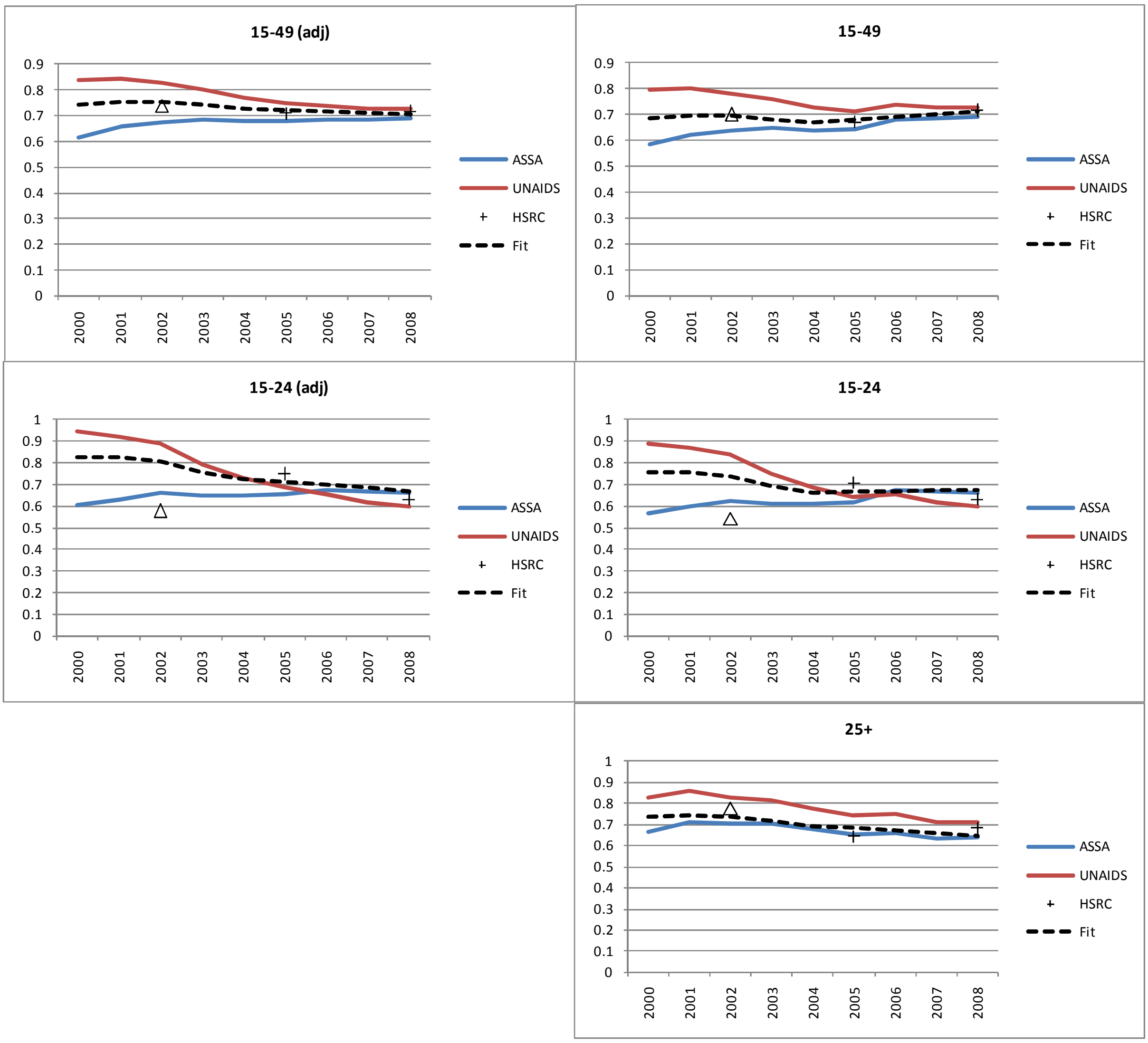


Figure 1 The ratio of population to antenatal prevalence, $_n r_{x,t}$, ASSA2008, UNAIDS, HSRC and fit (triangle represents HSRC ratio for 2002 □ not used to determine the fit)

From this we see that the fit to the two adjusted empirical ratios in 2005 and 2008 is very good in the 15-49 age group but not as good in the 15-24 age group, suggesting that estimates of incidence derived using the ratios in the 15-49 age group are probably reasonably reliable, but they are less reliable in the 15-24 age group. The lack of reliability of the estimates for the 25+ age group is partly a function of the smaller numbers of women attending antenatal clinics at these ages and lies somewhere between the other two.

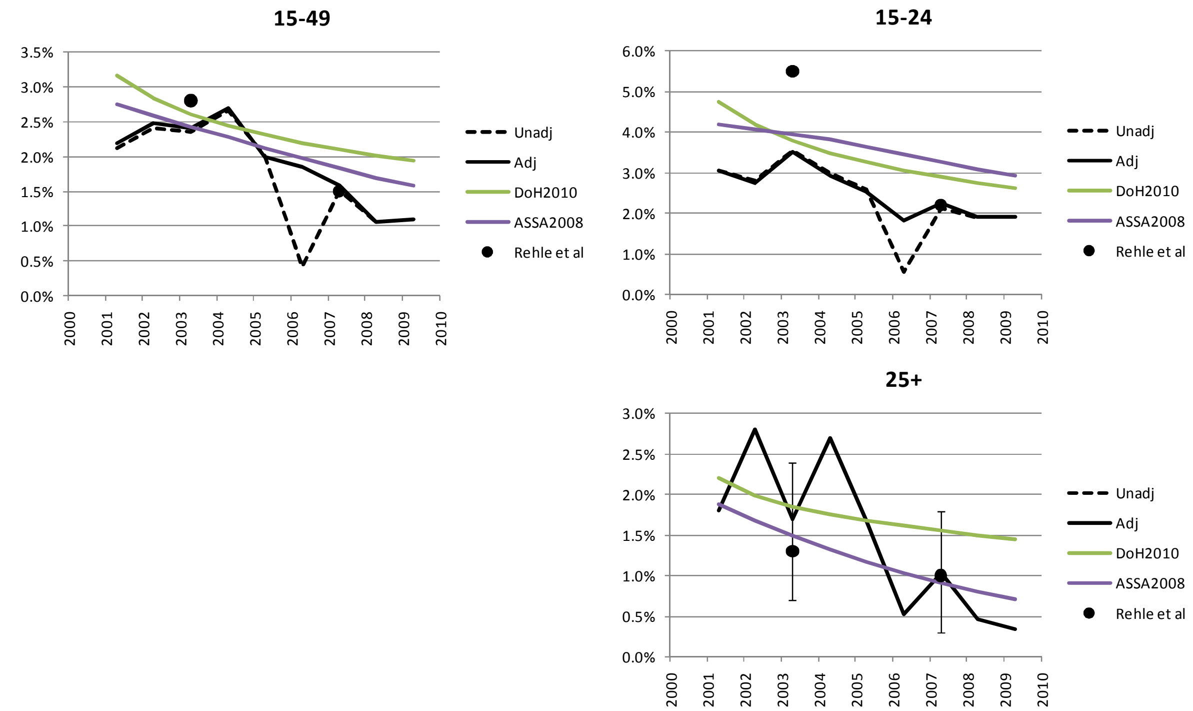


Figure 2 Incidence per annum: ASSA2008, UNAIDS, population prevalence estimated from the antenatal prevalence (adjusted and unadjusted)

Comparison of the results with those from the latest Spectrum model fitted for the National Department of Health, the ASSA2008 model and empirical estimates produced by Rehle, Hallett, Shisana *et al.* (2010) show a comforting consistency, from which a number of conclusions can be drawn. The first is that incidence appears to have fallen in at least the past five years in all age groups. The second is that the models appear to exaggerate the incidence in the youth, and the third is that it is unlikely that incidence in the youth has fallen by as much as suggested by Rehle *et al.* (60%).

Discussion

Given the consistency in the ratios of the population prevalence for women aged 15-49 to the prevalence of women attending public antenatal clinics using both the empirical survey results of the HSRC for the years 2005 and 2008, and even 2002, and the average of the estimated prevalence from the ASSA2008 and UNAIDS models shown in Figure 1, it appears possible to produce reasonably accurate annual estimates of the incidence of women in the population aged 15-49 by applying either the ratio estimated from the most recent HSRC survey or the extension of the straight line joining the 2005 and 2008 ratios. These estimates could be reconsidered when a new estimate of the prevalence of women in the population aged 15-49 is available from another survey.

Unfortunately one has less confidence in the method to produce accurate estimates of the incidence in the 15-24 age range (in particular in the 15-19 age range) because there is some difference between the two empirical ratios and the curve fitted doesn't pass very closely to the empirical points. It will thus be necessary to consider a further survey before it will be possible to decide to what extent the method might be useful. For the time being the best one could do is to extrapolate the trend fitted to the ratios (after adjusting for the bias in the surveys prior to 2006).

The fit to the trend in the ratios fits quite closely to the empirical points for ages 25-34 and the 25-49 age group, the numbers tested in the survey tends to be smaller for these age groups and thus the estimate is prone to much random fluctuation, even if we were to group several years.

In terms of deciding on estimate of incidence for 2007, the base year for the current National Strategic Plan, the best estimates would be the following: 1.5% for the 15-49 age group, 3% for the 15-24 age group, and 1% for the 25-49 age group.

Department of Health. 2009. *National HIV and Syphilis Antenatal Sero-Prevalence Survey in South Africa 2008* Pretoria, South Africa: Directorate: Health Systems Research, Research Coordination and Epidemiology, Department of Health. Available: http://www.doh.gov.za/docs/nassps-f.html

Department of Health. (no date). *Final Protocol: Annual antenatal HIV and Syphilis seroprevalence survey.* Pretoria: South African National Department of Health. Available:

Hallett, T. B., Zaba, B., Todd, J., Lopman, B. A., Mwita, W., Biraro, S. *et al.* 2008. "Estimating incidence from prevalence in generalised HIV epidemics: methods and validation", *PLoS Med* **5**(4):e80:doi:10.1371/journal.pmed.0050080.

Rehle, T. M., Hallett, T. B., Shisana, O., Pillay-van Wyk, V., Zuma, K., Carrara, H. *et al.* 2010. "A Decline in New HIV Infections in South Africa: Estimating HIV Incidence from Three National HIV Surveys in 2002, 2005 and 2008", *PLoS ONE* **5**(6):e11094. doi:11010.11371/journal.pone.0011094

Shisana, O., Rehle, T., Simbayi, L. C., Parker, W., Zuma, K. B., A, Connolly, C. *et al.* 2005. *South African National HIV Prevalence, HIV Incidence, Behavioural and Communications Survey, 2005.* Cape Town: HSRC. Available:

Shisana, O., Rehle, T., Simbayi, L. C., Zuma, K., Jooste, S., Pillay-van Wyk, V. *et al.* 2009. *South African National HIV Prevalence, HIV Incidence, Behavioural and Communications Survey, 2008.* Cape Town: HSRC. Available:

Shryock, H. S. and Siegel, J. S. 1976. *The Methods and Materials of Demography (Condensed Edition).* San Diego: Academic Press.

Appendix

Estimation of ${}_n r_{x,t}$ and ${}_n r_{x+1,t+1}$.

The equation for estimating the cohort incidence,

$${}_n\lambda_{x,t} = \frac{\frac{{}_nN_{x+1,t+1}}{{}_nN_{x,t}}\,{}_nr_{x+1,t+1}\cdot{}_nh^a_{x+1,t+1} - {}_n\pi_{x,t}\cdot{}_nr_{x,t}\cdot{}_nh^a_{x,t}}{\frac{1}{2}\left[\left(1-{}_n\,r_{x,t}\cdot{}_nh^a_{x,t}\right)+\frac{{}_nN_{x+1,t+1}}{{}_nN_{x,t}}\left(1-{}_n\,r_{x+1,t+1}\cdot{}_nh^a_{x+1,t+1}\right)\right]}$$

, can be rewritten as

$${}_n\lambda_{x,t} = \frac{{}_nr_{x,t}\left\{\frac{{}_nN_{x+1,t+1}}{{}_nN_{x,t}}\,\frac{{}_nr_{x+1,t+1}}{{}_nr_{x,t}}\,{}_nh^a_{x+1,t+1} - {}_n\pi_{x,t}\cdot{}_nh^a_{x,t}\right\}}{\frac{1}{2}\left[\left(1-{}_nr_{x,t}\cdot{}_nh^a_{x,t}\right)+\frac{{}_nN_{x+1,t+1}}{{}_nN_{x,t}}\left(1-\frac{{}_nr_{x+1,t+1}}{{}_nr_{x,t}}\,{}_nr_{x,t}\cdot{}_nh^a_{x+1,t+1}\right)\right]}$$

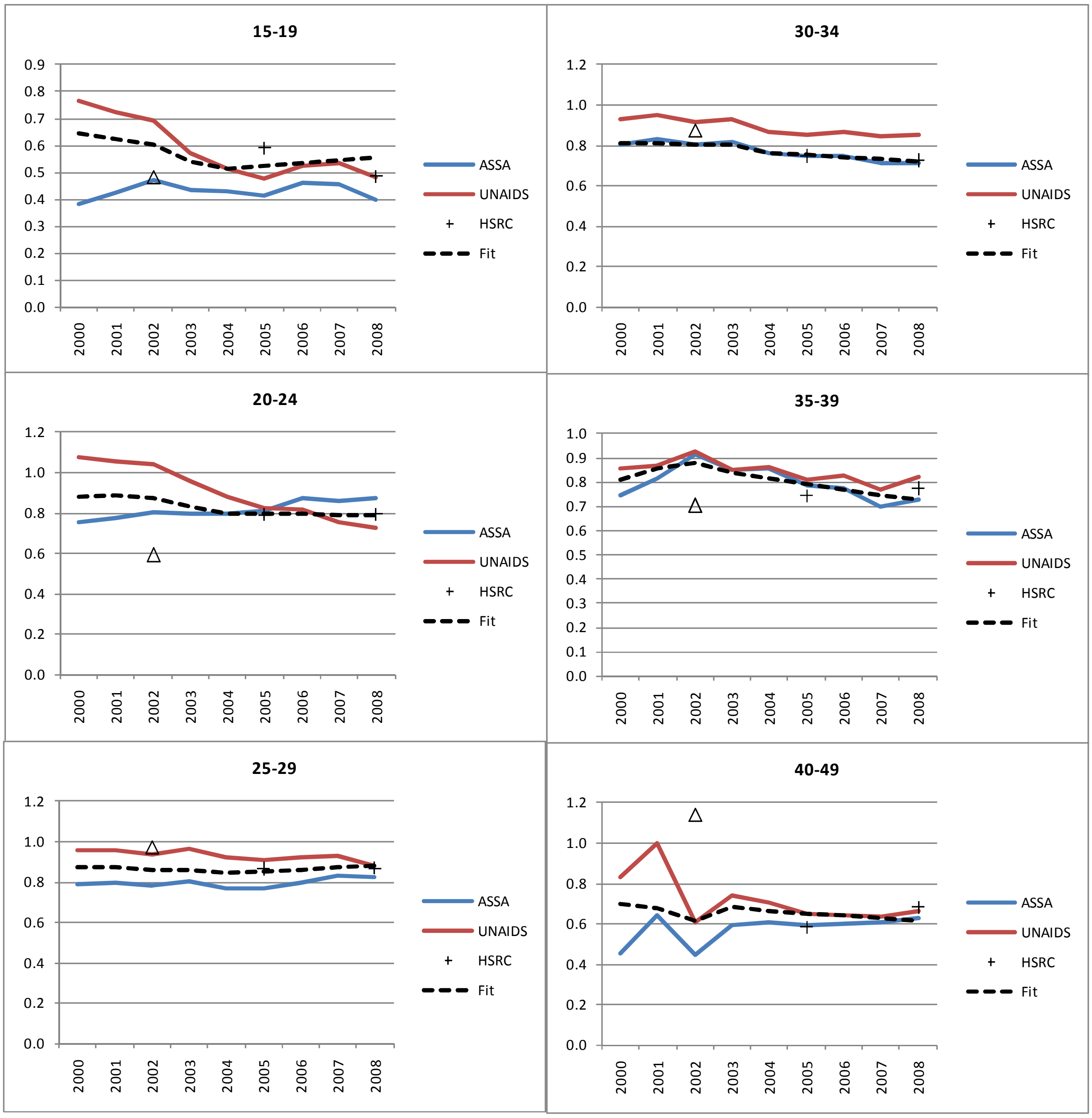
, from which it is apparent that the estimate is more sensitive to errors in the ratio of ratios $\frac{{}_nr_{x+1,t+1}}{{}_nr_{x,t}} = {}_nR_{x,t}$, than it is to errors in either ${}_nr_{x,t}$ or ${}_nr_{x+1,t+1}$, per se. Thus instead of fitting curves to the ${}_nr_{x,t}$ and ${}_nr_{x+1,t+1}$ observations separately, we decided to fit curves to ${}_nr_{x,t}$ and ${}_nR_{x,t}$.

The only empirical data available for estimating ${}_nR_{x,t}$, are ${}_n\hat{r}_{x,2005}$, ${}_n\hat{r}_{x,2008}$, ${}_n\hat{r}_{x+1,2005}$ and ${}_n\hat{r}_{x+1,2008}$ the ratios of the HSRC survey prevalence to the antenatal prevalence at the mid-point of the HSRC surveys in 2005 and 2008. As it can be expected that these ratios will be subject to random fluctuation they were used together to estimate an average ratio of ratios for the period 2005 to 2008, assumed to apply the ratio of prevalence at the start of 2006.

In order to apply the method to earlier years, mainly for illustrative purposes, the ratio of ratios, ${}_nR_{x,t}$ were estimated from a linear curve fitted to the ratio of ratios derived from the average of the ratios of the population to antenatal prevalence, ${}_n\bar{r}_{x,t}$ and ${}_n\bar{r}_{x+1,t}$, using the ASSA2008 model and the UNAIDS models to estimate the population prevalence, scaled to pass through the empirical estimate.

The ratios ${}_nr_{x,t}$ for the years 2001 to 2004 were estimated as the average of ${}_n\bar{r}_{x,t}$, and ${}_n\bar{r}_{x+1,t+1}/{}_nR_{x,t}$. For the years 2005 to 2008 the ${}_nr_{x,t}$ were estimated from the straight line which passes between the ratio for 2004 and the average of the empirical ratios ${}_n\hat{r}_{x,2005}$ and ${}_n\hat{r}_{x,2008}$, at mid-year 2006. The ratios ${}_nr_{x+1,t+1}$ are estimated as ${}_nr_{x,t}\cdot{}_nR_{x,t}$.

Comparison of the ratio of population prevalence to antenatal prevalence for various age groups (unadjusted)



15-19

30-34

20-24

35-39

25-29

40-49

Figure A.1 Comparison of the ratio of population prevalence to antenatal prevalence for various age groups (unadjusted)