Using panel data to partially identify HIV prevalence when HIV status is not missing at random

Bruno Arpino
Universitat Pompeu Fabra, Barcelona, Spain
Elisabetta De Cao
University of Groningen, Groningen, The Netherlands†
Franco Peracchi
University of Rome 'Tor Vergata', Rome, Italy

Abstract. Good estimates of HIV prevalence are important for policy makers in order to plan control programs and interventions. Although population-based surveys are now considered the "gold standard" to monitor the HIV epidemic, they are usually plagued by problems of non-ignorable nonresponse. We show how the partial identification approach can be adapted to exploit the availability of panel data and the absorbing nature of HIV and get narrower bounds on the HIV prevalence without imposing assumptions on the missing-data mechanism. Applied to longitudinal data from rural Malawi, our approach results in a considerable reduction of the width of the worst-case bounds.

1. Introduction

The prevalence of HIV in a population is defined as the proportion of people who are infected or, equivalently, the probability that a randomly drawn individual has the disease. Having reliable estimates of the HIV prevalence is essential for policy makers in order to plan control programs and interventions. Since the mid-1980s, the mainstay for monitoring the HIV epidemic has been facility-based sentinel surveillance data. Based on these data, HIV prevalence in developing countries has been found to be higher among women, sexually active people, and in urban areas. In many cases, estimates have been derived from pregnant women attending antenatal clinics (ANC) (Brookmeyer, 2010). ANC data have several sources of bias. First, they are only representative of pregnant women who are sexually active, and exclude men. Second, they may provide biased estimates

†*Address for correspondence:* Elisabetta De Cao, University of Groningen, University Center for Pharmacy, Department of Pharmacoepidemiology and Pharmacoeconomics, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands.
E-mail: elisabetta.decao@gmail.com

even for the sub-population of pregnant women because of the selective location of the clinics, mostly concentrated in urban areas. As a result, ANC-based estimates of HIV prevalence may be substantively biased upward (Gouws et al., 2008; Montana et al., 2008; Reniers and Eaton, 2009).

In recent years, many large-scale national surveys began to include biomarker modules to collect information on HIV serostatus. These biometric surveys are an important new source of data because they accurately measure HIV status and, unlike ANC-based surveys, are not restricted to a selected sub-population. Estimates of HIV prevalence derived from biometric surveys are, in general, considerably lower than those based on ANC data (Gouws et al., 2008; Montana et al., 2008). Based on these new results, UNAIDS corrected downward HIV prevalence estimates in several countries (Brookmeyer, 2010).

Although population-based surveys are now considered the "gold standard" to monitor the HIV epidemic (Boerma et al., 2003; Gouws et al., 2008; Mishra et al., 2008; Martin-Herz et al., 2006; Garcia-Calleja et al., 2006; Sakarovitch et al., 2007), these data may be affected by a different but not necessarily less severe source of bias, due to missing data on the respondents' HIV status. There are two main causes of missing data: refusal to take the HIV test and temporary absence or migration of the respondent. Approaches that discard cases with missing HIV status (complete-case analysis) implicitly rely on the assumption that data are missing completely at random (MCAR) (Rubin, 1976). Because MCAR implies that the distribution of observable characteristics should be the same for cases with and without missing data, this assumption is easily testable and is often rejected by the data. Failure of the MCAR assumption is likely to produce biased estimates of HIV prevalence.

To relax the MCAR assumption, imputation and weighting techniques are frequently used. These methods, based on the weaker assumption that data are missing at random (MAR) (Little and Rubin, 1987; Rubin, 1989), produce unbiased estimates only if the missing data mechanism does not depend on unobservables. In fact, many important sources of differences between individuals (such as knowledge or perceptions about one's HIV status), are unobservable, so HIV prevalence estimates based on the MAR assumption may be severely biased. For example, there is evidence that people refusing to be tested have higher risk of being infected (Reniers and Eaton, 2009). It has also been found that those who are not interviewed because of migration have higher risk of being HIV infected (Marston et al., 2008; Crampin et al., 2003; Obare, 2010). Anglewicz (2007) analyzes this phenomenon using data from a follow-up specifically designed to interview respondents who did not participate in one wave of a panel survey for Malawi because of absence. He finds that migrants are likely to report a higher number of sexual partners and to be HIV positive. An explanation is that HIV infected people are more likely to migrate as a consequence of union dissolution due to death of the partner or divorce.

Unlike MCAR, the MAR assumption is essentially untestable and several approaches have been proposed to avoid it (see Vella, 1998, for a survey). These

approaches have recently been used to estimate HIV prevalence (Lachaud, 2007; Reniers and Eaton, 2009; Barnighausen et al., 2011). For example, using a Heckman-type selection model (Heckman, 1979) Barnighausen et al. (2011) show that prevalence estimates can be very severely underestimated when missingness depends on unobserved variables. One problem with these alternative approaches, however, is that they tend to impose strong restrictions on the distribution of the unobservables.

The aim of our paper is to study what can be learned about HIV prevalence when data are subject to non-ignorable missing data mechanisms avoiding strong untestable assumptions. We follow Horowitz and Manski (1998) and Manski (1995, 2003) and switch the focus away from point identification, which typically relies on a combination of strong requirements about the data and strong assumptions about the model, to partial identification. We first use the empirical evidence alone to identify a region of credible values for HIV prevalence. We then exploit the availability of panel data and the absorbing nature of HIV infection to narrow the width of this region. Although additional assumptions, such as instrumental variable (IV) and monotone instrumental variable (MIV) restrictions, may be used to further narrow the width of the identification region, our main contribution is to show the power of combining substantive information about the HIV process with the longitudinal nature of the data.

Our data come from the Malawi Diffusion and Ideational Change Project (MDICP), a longitudinal survey conducted every two years in rural Malawi since 1998. Starting from 2004, a biometric survey has been added to the main survey allowing the estimation of HIV prevalence. Malawi is one of the countries most affected by the HIV epidemic. Out of a population of 15 million people, 80% of them living in rural areas, almost one million people are living with HIV and AIDS is the leading cause of death among adults. The national HIV prevalence rate, based on the 2004 Malawi Demographic and Health Survey (MDHS), is equal to 11.8% for people aged 15–49. Like for most countries in sub-Saharan Africa, where HIV is mainly transmitted via heterosexual contact, HIV prevalence is estimated to be higher for women than for men (13.3% against 10.2%), and to be higher in urban than in rural areas (17.1% versus 10.8%). Although the MDICP may only be considered representative of the population of rural Malawi, it has the advantage over the MDHS of being a longitudinal survey. Further, unlike the MDHS for which a biomarker module is currently available only for 2004, biometric data from the MDICP are available for 2004, 2006 and 2008.

The remainder of this paper is organised as follows. Section 2 describes the data and the problem of missing information on HIV status. Section 3 reviews the partial identification approach and shows how to exploit the longitudinal nature of the data and the absorbing nature of HIV infection to narrow the worst-case bounds. It also discusses how to use plausible IV and MIV restrictions to further narrow the bounds. Section 4 presents the estimated HIV prevalence bounds for the whole population and, separately, by region, gender and cohort.

Finally, Section 5 offers some conclusions.

2. Data

We use data from the Malawi Diffusion and Ideational Change Project (MDICP), a longitudinal survey conducted in rural Malawi (the data can be freely downloaded from the following website: `http://www.malawi.pop.upenn.edu`). The MDICP is the result of a collaboration of the University of Pennsylvania with the College of Medicine and Chancellor College at the University of Malawi. This data set is particularly interesting for our purposes because it is longitudinal and includes HIV tests for the years 2004, 2006 and 2008.

2.1. MDICP survey

The MDICP survey has been carried out in three of the 28 Malawian districts, one for each of the three administrative regions of the country: Balaka in the South, Mchinji in the Center and Rumphi in the North. The three regions are significantly different in terms of ethnic composition, language, religious practice, population density, literacy, and prevailing social system (e.g. patrilocal or matrilocal residence).

The first wave of the survey was carried out in 1998, interviewing 1,541 ever-married women of childbearing age and 1,198 men, most of them husbands of the married women in the sample. The second wave, carried out in 2001, followed-up the respondents and interviewed the new spouses of respondents who got married between the first and the second wave (Watkins et al., 2003). The third wave, carried out in 2004, augmented the original sample with a random sample of about 1,500 people aged 15–28 (both married and never-married), to correct for ageing of the baseline sample and the fact that the original sample was restricted to ever-married women and their husband. With this addition, the survey may be regarded as broadly representative of the population of rural Malawi (see `http://www.malawi.pop.upenn.edu` for further details about the sampling strategy). The fourth (2006) and fifth (2008) waves added the spouses of newly married people.

The survey instrument asks about sexual relations, risk assessments, marriage and partnership histories, household rosters and transfers, as well as income and other measures of wealth. It also includes information on village-level variables, regional market prices, and weather conditions. The survey instrument was translated from English in the three most common local languages (Yao, Chichewa, and Tumbuka). Interviews were carried out face-to-face by interviewers who spoke the same language as the interviewees and were hired and trained locally.

Starting from 2004, a biometric survey, called the voluntary consulting and test (VCT) survey, has been added to the main survey. The VCT survey consists of a short questionnaire, submitted a few days after the main survey and focused

on sexual behaviour and AIDS related questions, and free tests for HIV and other sexually transmitted infections administered by nurses from outside the area. Respondents to the VCT survey are also offered pre-test counselling about HIV prevention strategies. In 2004, oral swabs were used for the HIV test and results were given to respondents 2–4 months after testing. In 2006 and 2008, the MDICP team tested only for HIV using an improved testing procedure consisting of rapid response blood test. Measurement error in the two types of tests (oral swabs and blood test) is very limited and, being due only to the accuracy limit of the measuring instruments, can be considered as random.

Although the survey was not designed to be representative of the population in rural Malawi, the characteristics of the 2004 sample closely match those of the 2004 MDHS for rural Malawi (Thornton, 2008). We focus on people interviewed in 2004, excluding new entrants in 2006 and 2008, and dropping from the sample people who were never successfully contacted. Because prevalence is defined on the population of alive people, our working sample consists of 4,062 persons who were alive in 2004. When computing HIV prevalence for 2006 and 2008, we exclude people who died after 2004.

2.2. Missing data

In each of the three waves considered, HIV status is missing for a substantial fraction of the sample. Missing HIV status may arise from either unit or item nonresponse. We define as unit nonresponse the case in which both the main and the VCT survey are missing because of failure to establish a contact or refusal to cooperate. Item nonresponse occurs when HIV status is not available for responding units.

There are different patterns of unit nonresponse across our three waves. About 55% of the sample are unit respondents in all three waves, about 12% are unit respondents in 2004 and unit nonrespondents in 2006 and 2008, about 11% are unit respondents in 2004 and 2006 and unit nonrespondents in 2008, about 8% are unit respondents in 2004 and 2008 and unit nonrespondents in 2006, while the remaining 14% include the other patterns of unit nonresponse.

Table 1 shows the various sources of missing data. Overall, the fraction with missing HIV status is 29% in 2004, about 37% in 2006 and rises to 42% in 2008 due to the increase in item nonresponse from 14% to 18% and to a larger increase in unit nonresponse from 15% to 24%. The main reason for unit nonresponse, and for its increase across waves, is migration. Hospitalisation and refusal to participate are relatively unimportant. Other reasons for unit nonresponse are lumped into the residual category 'other', consisting mainly of people who did not fill the questionnaire because too old or too sick, or for unknown reasons. People who are unit nonrespondents because of migration, unknown reasons or 'other' reasons will be assumed to be alive when computing the bounds.

The main reason for item nonresponse is refusal to get tested although, in 2004, the refusal rate in the MDICP (6.3%) is lower than for the MDHS in rural

Table 1. Distribution of types of unit respondents and nonrespondents by wave.

	2004		*2006*		*2008*	
	Freq	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
UNIT RESPONDENTS						
HIV negative	2700	66.47	2408	59.28	2116	52.09
HIV positive	177	4.36	123	3.03	117	2.88
Item nonresponse						
Test refused	256	6.30	200	4.92	172	4.23
Indeterminate	14	0.34	6	0.15	1	0.02
Results lost	24	0.59	0.00	0.00	0.00	0.00
Other†	319	7.85	313	7.71	569	14.01
UNIT NONRESPONDENTS						
Refused	27	0.66	11	0.27	58	1.43
Moved	184	4.53	479	11.79	470	11.57
Temporarily absent	36	0.89	41	1.01	76	1.87
Hospitalized	6	0.15	5	0.12	1	0.02
Other‡	319	7.85	432	10.64	359	8.84
Dead	/	/	44	1.08	123	3.03
Total§	4062	100	4062	100	4062	100

†People that fulfil the first part of the questionnaire, but not the second, for example because they were temporarily absent during the biomarker collection.
‡People who did not fulfil the questionnaire for unknown reasons or because too old or too sick.
§The new entrants 2006/2008 are excluded.

areas (21.7%). Thornton (2008) argues that this may be due to the method of testing (oral swabs) and the fact that the MDICP does not require respondents to learn their results at the time of testing. However, low refusal rates in the MDCIP are also found in 2006 and 2008. In very few cases the results of the HIV test are indeterminate or have been lost. Other reasons for item nonresponse are lumped into the category 'other', consisting of people who completed the main survey but not the VCT survey, for example because they were temporarily absent. The importance of this residual category almost doubled between 2004 and 2008.

The classification of the different sources of missing data is important. Ignoring missing data due to refusal to be tested or migration may bias the HIV prevalence estimate downward (Reniers and Eaton, 2009; Obare, 2010). On the other hand, missing data due to loss of test results are not a major source of concern and may be considered as purely random.

3. Partial identification of HIV prevalence

To formalise our problem, consider a population that, at a given time t, consists of N_t living individuals who can be either susceptible to HIV or infected. A

susceptible individual is a member of the population who, at a given point in time, is at risk of becoming infected by the disease.

HIV status of individual i at time t is represented by the binary indicator y_{it}, which is equal to one if individual i is HIV positive and to zero otherwise. HIV prevalence at time t is just the proportion, $\pi_t = N_t^{-1} \sum_{i=1}^{N_t} y_{it}$, of HIV infected people, which in turn is equal to $\Pr(Y_t = 1)$, where Y_t is a binary random variable equal to one if a randomly selected individual is HIV positive at time t and to zero otherwise.

Our aim is to construct informative bounds for π_t when HIV status is missing for a fraction of individuals in the population. As argued in the previous section, in our data measurement error is negligible and may be considered as purely random. Thus, unlike Kreider and Pepper (2007) and Nicoletti et al. (2011), we ignore this problem and focus on the uncertainty about π_t caused by missing data.

3.1. Bounds with cross-sectional data

We first consider the problem of bounding HIV prevalence when data are only available at a given point in time, as in a single cross-section or when the longitudinal dimension of a panel survey is ignored.

By the law of total probability, we can write HIV prevalence at time t as

$$\pi_t = \Pr(Y_t = 1|D_t = 1)\Pr(D_t = 1) + \Pr(Y_t = 1|D_t = 0)\Pr(D_t = 0), \quad (1)$$

where D_t is a binary indicator equal to one if HIV status is known and to zero otherwise. As pointed out by Manski (1989), the missing data problem arises because the data tell us nothing about $\Pr(Y_t = 1|D_t = 0)$. However, because $0 \leq \Pr(Y_t = 1|D_t = 0) \leq 1$, substituting the lower and upper bounds for $\Pr(Y_t = 1|D_t = 0)$ into (1) gives the following lower and upper bounds on π_t

$$\begin{aligned} LB_t &= \Pr(Y_t = 1|D_t = 1)\Pr(D_t = 1) = \Pr(Y_t = 1, D_t = 1), \\ UB_t &= \Pr(Y_t = 1|D_t = 1)\Pr(D_t = 1) + \Pr(D_t = 0), \\ &= \Pr(Y_t = 1, D_t = 1) + \Pr(D_t = 0). \end{aligned}$$

These bounds are often referred to as worst-case bounds. If only a cross-section is available, these bounds are sharp because they use all the available information.

The identification region for π_t consists of all the points in the interval between LB_t and UB_t. The width $W_t = UB_t - LB_t$ of this region is equal to the nonresponse probability $\Pr(D_t = 0)$, which therefore represents a direct measure of the uncertainty about HIV prevalence caused by nonresponse (Horowitz and Manski, 1998). Without nonresponse, there is no uncertainty about π_t. When nonresponse is frequent, the uncertainty is large. In this case, an important issue is whether there exist credible restrictions on the HIV process that may be used to narrow the worst-case bounds.

3.2. Bounds with panel data

HIV infection is an absorbing state: a person infected at any given time has zero probability of becoming susceptible at later times, while a person susceptible at any given time has probability one of being susceptible at earlier times. These simple considerations help narrow the worst-case bounds when panel data are available and HIV status of people who are nonrespondent in one wave may be observed in other waves. We will refer to the resulting bounds as 'dynamic', because they use restrictions on the dynamics of the HIV epidemic. To keep things simple, we only present results for the case of short panels with two or three waves. The Appendix A presents the results for the general case of a panel with $P \geq 1$ waves before wave t, or $F \geq 1$ waves after wave t, or both.

Suppose first that we use only two waves of a panel, at times t and $t+1$. To narrow the worst-case bounds on π_t, consider again equation (1) and notice that

$$\begin{aligned}\Pr(Y_t = 1|D_t = 0) = \Pr(Y_t = 1|D_{t+1} = 0, D_t = 0)\Pr(D_{t+1} = 0|D_t = 0) + \\ + \Pr(Y_t = 1|D_{t+1} = 1, D_t = 0)\Pr(D_{t+1} = 1|D_t = 0),\end{aligned}$$

where

$$\begin{aligned}&\Pr(Y_t = 1|D_{t+1} = 1, D_t = 0) = \\ &\quad = \Pr(Y_t = 1|Y_{t+1} = 1, D_{t+1} = 1, D_t = 0)\Pr(Y_{t+1} = 1|D_{t+1} = 1, D_t = 0),\end{aligned}$$

since $\Pr(Y_t = 1|Y_{t+1} = 0, D_{t+1} = 1, D_t = 0) = 0$ due to the absorbing nature of HIV status. Thus, we can rewrite (1) as

$$\begin{aligned}\Pr(Y_t = 1) = \Pr(Y_t = 1, D_t = 1) + \\ + \Pr(Y_t = 1|D_{t+1} = 0, D_t = 0)\Pr(D_{t+1} = 0, D_t = 0) + \\ + \Pr(Y_t = 1|Y_{t+1} = 1, D_{t+1} = 1, D_t = 0)\times \\ \times \Pr(Y_{t+1} = 1|D_{t+1} = 1, D_t = 0)\Pr(D_{t+1} = 1, D_t = 0).\end{aligned} \tag{2}$$

From (2) we obtain lower and upper bounds on π_t by assuming that the unknown probabilities $\Pr(Y_t = 1|D_{t+1} = 0, D_t = 0)$ and $\Pr(Y_t = 1|Y_{t+1} = 1, D_{t+1} = 1, D_t = 0)$ are respectively equal to their lower bound of zero and their upper bound of one. Setting both probabilities equal to zero gives the lower bound

$$LB_t^{(+1)} = LB_t,$$

while setting both of them equal to one gives the upper bound

$$\begin{aligned}UB_t^{(+1)} &= \Pr(Y_t = 1, D_t = 1) + \Pr(D_{t+1} = 0, D_t = 0) + \\ &\quad + \Pr(Y_{t+1} = 1|D_{t+1} = 1, D_t = 0)\Pr(D_{t+1} = 1, D_t = 0) \\ &= \Pr(Y_t = 1, D_t = 1) + \Pr(D_t = 0)\times \\ &\quad \times \{\Pr(Y_{t+1} = 1, D_{t+1} = 1|D_t = 0) + \Pr(D_{t+1} = 1|D_t = 0)\} \\ &= UB_t - \Pr(D_t = 0)\times \\ &\quad \times \{1 - \Pr(Y_{t+1} = 1, D_{t+1} = 1|D_t = 0) - \Pr(D_{t+1} = 1|D_t = 0)\},\end{aligned}$$

where the term in square brackets in the last relationship is equal to the conditional probability that $Y_{t+1} = 0$ and $D_{t+1} = 1$ given $D_t = 0$, and is therefore bounded between zero and one. With a 2-wave panel, unlike the worst-case bounds, these new bounds are sharp, as they use all the available information. The width of the resulting identification region for π_t is

$$W_t^{(+1)} = UB_t^{(+1)} - LB_t^{(+1)} = W_t - \Pr(Y_{t+1} = 0, D_{t+1} = 1, D_t = 0).$$

Because $\Pr(Y_{t+1} = 0, D_{t+1} = 1, D_t = 0)$ is bounded between zero and one, and cannot exceed $\Pr(D_t = 0)$, we have that $0 \leq W_t^{(+1)} \leq W_t$.

Notice that simply knowing the HIV status at $t + 1$ of people with missing HIV status at t is not enough to narrow the worst-case bounds. In fact, among the respondents at $t+1$, only the information about negative HIV status can be used to infer HIV status at t, so only the upper bound can be reduced relative to the worst-case. Respondents at $t + 1$ who are found to be HIV positive cannot be assumed to have been already HIV positive at t, so the lower bound is the same as in the worst-case.

If the two waves of the panel are at times $t-1$ and t, then we can rewrite the unknown probability in (1) by exploiting past rather than future information. This gives

$$\begin{aligned}\Pr(Y_t = 1|D_t = 0) = \Pr(Y_t = 1|D_t = 0, D_{t-1} = 0) \Pr(D_{t-1} = 0|D_t = 0) + \\ + \Pr(Y_t = 1|D_t = 0, D_{t-1} = 1) \Pr(D_{t-1} = 1|D_t = 0),\end{aligned}$$

where

$$\begin{aligned}&\Pr(Y_t = 1|D_t = 0, D_{t-1} = 1) = \\ &\quad = \Pr(Y_t = 1|D_t = 0, D_{t-1} = 1, Y_{t-1} = 0) \Pr(Y_{t-1} = 0|D_t = 0, D_{t-1} = 1) + \\ &\qquad + \Pr(Y_{t-1} = 1|D_t = 0, D_{t-1} = 1),\end{aligned}$$

since $\Pr(Y_t = 1|D_t = 0, D_{t-1} = 1, Y_{t-1} = 1) = 1$ due to the absorbing nature of HIV status. Proceeding as before, we obtain the following bounds

$$\begin{aligned}LB_t^{(-1)} &= LB_t + \Pr(Y_{t-1} = 1, D_{t-1} = 1, D_t = 0), \\ UB_t^{(-1)} &= UB_t.\end{aligned}$$

Notice that, unlike the case when future information is used, here the upper bound is the same as in the worst-case, while the lower bound is greater. This is because past negative HIV status is uninformative, as we cannot assume that a person who was HIV negative in the past remains HIV negative in the future, while past positive HIV status is informative, as a person who was HIV positive in the past remains HIV positive in the future. The width of the resulting identification region for π_t is

$$W_t^{(-1)} = UB_t^{(-1)} - LB_t^{(-1)} = W_t - \Pr(Y_{t-1} = 1, D_{t-1} = 1, D_t = 0).$$

Again, $0 \leq W_t^{(-1)} \leq W_t$.

Using three waves of a panel, we can further narrow the identification region for π_t. Suppose that, in addition to wave t, we use one wave before t and one after t. Then it follows immediately from our previous results that

$$\begin{aligned} LB_t^{(-1,+1)} &= LB_t^{(-1)}, \\ UP_t^{(-1,+1)} &= UB_t^{(+1)}, \\ W_t^{(-1,+1)} &= W_t - \Pr(Y_{t+1} = 0, D_{t+1} = 1, D_t = 0) - \\ &\quad - \Pr(Y_{t-1} = 1, D_{t-1} = 1, D_t = 0). \end{aligned}$$

Using wave t and two waves after t we instead have

$$\begin{aligned} LB_t^{(+2)} &= LB_t^{(+1)}, \\ UB_t^{(+2)} &= UB_t^{(+1)} - \Pr(Y_{t+2} = 0, D_{t+2} = 1, D_{t+1} = D_t = 0), \\ W_t^{(+2)} &= W_t^{(+1)} - \Pr(Y_{t+2} = 0, D_{t+2} = 1, D_{t+1} = D_t = 0), \end{aligned}$$

while using wave t and two waves before t we have

$$\begin{aligned} LB_t^{(-2)} &= LB_t^{(-1)} + \Pr(Y_{t-2} = 1, D_{t-2} = 1, D_{t-1} = D_t = 0), \\ UB_t^{(-2)} &= UB_t^{(-1)}, \\ W_t^{(-2)} &= W_t^{(-1)} - \Pr(Y_{t-2} = 1, D_{t-2} = 1, D_{t-1} = D_t = 0). \end{aligned}$$

In the last two cases, the uncertainty about π_t due to missing data decreases because of either an increase in the lower bound or a decrease in the upper bound, in the first case because of a combination of the two effects. Increasing the number of available waves further decreases the uncertainty due to missing data as it is shown in the Appendix A.

3.3. IV and MIV restrictions

To further narrow the identification region for π_t, the restrictions discussed in Section 3.2 may be combined with those implied by additional assumptions on the HIV process.

One possibility are instrumental variable (IV) restrictions (Manski, 1994, 2003). A random variable is an IV if it helps predict nonresponse but does not help predict HIV status, possibly after conditioning on a set of observable covariates. Although it is generally difficult to find valid instrumental variables, a convincing case can be made for data collection characteristics (characteristics of the interviewer, interview mode, length and design of the questionnaire, etc.), because they help predict nonresponse (Lepkowski and Couper, 2002; Nicoletti and Peracchi, 2006), but lack predictive power for HIV status.

Since IV restrictions are often controversial, another possibility is to impose weaker monotone instrumental variable (MIV) restrictions (Manski and Pepper,

2000). A random variable is a MIV if it shifts HIV prevalence monotonically, possibly after conditioning on a set of observable covariates.

4. Results

We illustrate our results by presenting complete-case estimates, worst-case bounds and dynamic bounds for HIV prevalence in rural Malawi constructed from the MDICP data for 2004, 2006 and 2008. Since it is of interest for both research and policy-making to know how the HIV epidemic is spread among different demographic groups, we present estimates for the whole population and for subgroups defined by region, gender, and birth cohort. We distinguish between four cohorts: i) Cohort A: born 1984–1989 (aged 15–20 in 2004), ii) Cohort B: born 1975–1983 (aged 21–29 in 2004), iii) Cohort C: born 1965–1974 (aged 30–39 in 2004), and iv) Cohort D: born before 1965 (aged 40+ in 2004).

4.1. Complete-case estimates

The complete-case estimates of HIV prevalence in rural Malawi are 6.2% for 2004, 4.9% for 2006, and 5.1% for 2008. These estimates are substantially lower than the 2004 MDHS estimate of 10.8% for rural Malawi, possibly because the MDICP sample does not include peri-urban areas (Obare et al., 2009), and show no clear trend.

Details about the complete-case estimates are reported in Table 4 of the supporting materials. In particular, for the youngest cohort (born 1984–1989), estimated HIV prevalence is very low in all three waves (between 0 and 4%). Among males it is always highest for the cohort born before 1965 (about 4-6%) while, among females, it is highest for the 1975–83 cohort in 2004 (about 9%) and the 1965–74 cohort in 2006 and 2008 (about 10%). However, since the fraction of the sample with missing HIV status is very high in each year, uncertainty about the complete-case estimates is also high.

4.2. Worst-case and dynamic bounds

The bounds introduced in Section 3 are easily estimated non-parametrically by their sample counterparts. Since they are estimated, their sampling variability must be taken into account. We do this by constructing 95%-level bootstrap confidence intervals based on the percentile method with 999 bootstrap replications. The interval between the upper limit of the 95%-level confidence interval for the upper bound and the lower limit of the 95%-level confidence interval for the lower bound is a 95%-level confidence interval for the identification region.

Figure 1 displays graphically the worst-case and the dynamic bounds on HIV prevalence in rural Malawi, along with the complete-case estimates. Using the worst-case bounds, the identification region is the interval between 3.8% and 34.2% in 2004, the interval between 2.6% and 40.2% in 2006, and the interval

between 2.4% and 46.6% in 2008 (see also Table 2, All regions). Notice that the width of these intervals increases over time following the pattern of missing data. From Figure 1, we can also notice that the complete-case estimates are always very close to the lower bound of the identification region.

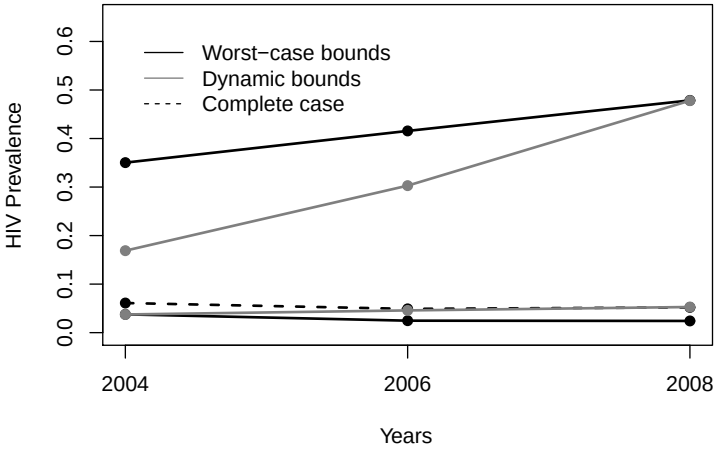

Figure 1. HIV prevalence for the whole sample by survey year.

Using the dynamic bounds, the identification region is the interval between 3.8% and 15.9% in 2004, between 4.5% and 28.9% in 2006, and between 4.9% and 46.6% in 2008. Thus, for the first two waves, we have a sizeable reductions of the uncertainty about HIV prevalence compared to the worst-case bounds (amounting to a reduction of their width by about 18.2 percentage points in 2004 and 13.2 percentage points in 2006). For the last wave, the reduction of the bound width is instead limited (only 2.4 percentage points). This pattern reflects the number of waves available before and after the point in time where HIV prevalence is estimated. In 2004 only future information about HIV status can be used. As a consequence, the dynamic upper bound is lower than the worst-case upper bound but the lower bound is unchanged. In 2006, both previous and future information about HIV status help reduce the uncertainty, resulting in a decrease of the upper bound and an increase of the lower bound. In 2008, since no subsequent wave of the panel is available, only previous information about HIV status helps reduce the uncertainty, resulting in an increase of the lower bound with the upper bound unchanged.

Table 2 shows bounds for the different geographical regions of Malawi: North, Center and South. According to the MDHS, Southern Malawi is the region with

Table 2. Bootstrapped bounds for the whole sample and by regions.

Year	Region	Bounds type	L†	U‡	W§
2004	North	Worst-case	0.023	0.285	0.262
		Dynamic	0.023	0.133	0.110
	Center	Worst-case	0.031	0.453	0.422
		Dynamic	0.031	0.199	0.168
	South	Worst-case	0.043	0.339	0.296
		Dynamic	0.044	0.195	0.15
	All	Worst-case	0.038	0.342	0.304
		Dynamic	0.038	0.159	0.122
2006	North	Worst-case	0.018	0.361	0.343
		Dynamic	0.028	0.275	0.247
	Center	Worst-case	0.018	0.440	0.421
		Dynamic	0.033	0.293	0.259
	South	Worst-case	0.028	0.471	0.443
		Dynamic	0.060	0.372	0.312
	All	Worst-case	0.026	0.402	0.376
		Dynamic	0.045	0.289	0.244
2008	North	Worst-case	0.023	0.472	0.449
		Dynamic	0.037	0.473	0.437
	Center	Worst-case	0.015	0.436	0.420
		Dynamic	0.037	0.438	0.401
	South	Worst-case	0.026	0.554	0.528
		Dynamic	0.068	0.555	0.486
	All	Worst-case	0.024	0.466	0.442
		Dynamic	0.049	0.466	0.417

†Lower bound. ‡Upper bound. §Width.

the highest HIV prevalence, followed by the Center and the North. Although the dynamic bounds are much narrower than the worst-case bounds, they are still too wide to support this conclusion.

Table 3 shows that the dynamic bounds are much narrower than the worst-case bounds also if we consider subgroups characterized by gender and birth cohort. Again, this is especially true for 2004 and 2006. Although the width of the dynamic bounds is generally lower for males, meaning that there is more uncertainty about HIV prevalence among females, the identification regions remain too wide to allow us to establish a rank by gender.

4.3. Imposing additional IV and MIV restrictions

The IVs considered are: gender differences between the interviewer and the interviewee, interviewer's experience, interviewer's age categorised in two classes, and the month of the first interview attempt. The latter is the only IV available in 2008. As MIV, we consider the number of sexual partners each respondent had till that year. This is a valid MIV if the probability of being HIV infected does

Table 3. Bootstrapped bounds by gender and birth cohort.

			Gender					
			Male			Female		
Year	Cohort	Bounds type	L†	U‡	W§	L†	U‡	W§
2004	A	Worst-case	0.000	0.265	0.265	0.002	0.359	0.357
		Dynamic	0.000	0.094	0.094	0.002	0.169	0.167
	B	Worst-case	0.008	0.321	0.313	0.041	0.42	0.379
		Dynamic	0.008	0.139	0.131	0.045	0.218	0.173
	C	Worst-case	0.020	0.452	0.432	0.042	0.364	0.323
		Dynamic	0.020	0.216	0.196	0.04	0.198	0.158
	D	Worst-case	0.048	0.395	0.347	0.042	0.331	0.289
		Dynamic	0.049	0.211	0.162	0.042	0.154	0.113
2006	A	Worst-case	0.000	0.455	0.455	0.002	0.529	0.526
		Dynamic	0.000	0.322	0.322	0.006	0.385	0.379
	B	Worst-case	0.000	0.485	0.485	0.027	0.465	0.438
		Dynamic	0.008	0.370	0.362	0.057	0.333	0.275
	C	Worst-case	0.010	0.403	0.392	0.055	0.381	0.326
		Dynamic	0.023	0.309	0.286	0.074	0.280	0.207
	D	Worst-case	0.024	0.393	0.369	0.016	0.362	0.346
		Dynamic	0.047	0.305	0.258	0.039	0.236	0.196
2008	A	Worst-case	0.000	0.583	0.583	0.008	0.616	0.607
		Dynamic	0.000	0.583	0.583	0.015	0.611	0.597
	B	Worst-case	0.006	0.577	0.572	0.027	0.457	0.429
		Dynamic	0.020	0.577	0.558	0.067	0.462	0.395
	C	Worst-case	0.009	0.485	0.476	0.041	0.440	0.399
		Dynamic	0.024	0.485	0.462	0.075	0.437	0.362
	D	Worst-case	0.016	0.480	0.464	0.015	0.386	0.371
		Dynamic	0.047	0.476	0.429	0.039	0.388	0.349

†Lower bound. ‡Upper bound. §Width.

not fall as the number of sexual partners increases. Further, because information on IVs and MIVs is not available for unit nonrespondents, the following analysis is restricted to the subsample of unit respondents. Detailed results by year, gender and cohort are contained in Tables 5–7 in the supporting materials.

The identification region for HIV prevalence in 2004, produced by the dynamic bounds, is the interval between 4.1% and 13.6% in the benchmark case, the interval between 4.3% and 12% when using the interview month as an IV, and the interval between 4.2% and 13% when using our MIV. The identification region for HIV prevalence in 2006, is the interval between 3.5% and 16.6% in the benchmark case, the interval between 3.7% and 15.1% when using the interview month as an IV, and the interval between 3.6% and 16.6% when using our MIV. The identification region for HIV prevalence in 2008 is the interval between 4.3% and 30.6% in the benchmark case, the interval between 4.7% and 26.5% when using the interview month as an IV, and the interval between 4.4% and 30.3% when using our MIV. Thus, using the interview month as an IV reduces the width of the identification region relative to the benchmark case by 1.7 percentage points in 2004 and in 2006, and by 4.4 percentage points in 2008. On the other hand, the number of sexual partners does not appear to be an effective MIV, as it is of little help in narrowing the identification region.

Figure 2 shows the dynamic bounds on HIV prevalence by survey year, separately by gender and birth cohort, along with the complete-case estimates. The 'best IV' available, namely the one that most reduces the width of the identification region, varies with gender and cohort. In 2004 the best IVs are either the interview month or the interviewer's experience, while in 2006 the best IV is always the interview month. Unlike the case of the whole sample, the MIV restriction now seems to be more effective in reducing the width of the identification interval, although its effectiveness varies with gender and cohort.

5. Discussion

Having reliable estimates of HIV prevalence is critical for policy makers. Today, the gold-standard is estimates based on biomarkers collected in population based surveys. These surveys, however, are plagued by non-ignorable missing data problems, which in turn translate into substantial uncertainty about HIV prevalence in the population.

Our paper uses a bounding approach to assess what can be learnt from this type of data. Its main contribution is to show how worst-case bounds, which are often distressingly wide, can be narrowed when panel data are available by exploiting the absorbing nature of HIV infection.

Panel data are typically used to estimate HIV incidence rates. However, they can also be used to estimate HIV prevalence at different points in time for the same population.

We show that the identifying power of panel data comes from the fact that we are able to observe in other waves the HIV status of current nonrespondents.

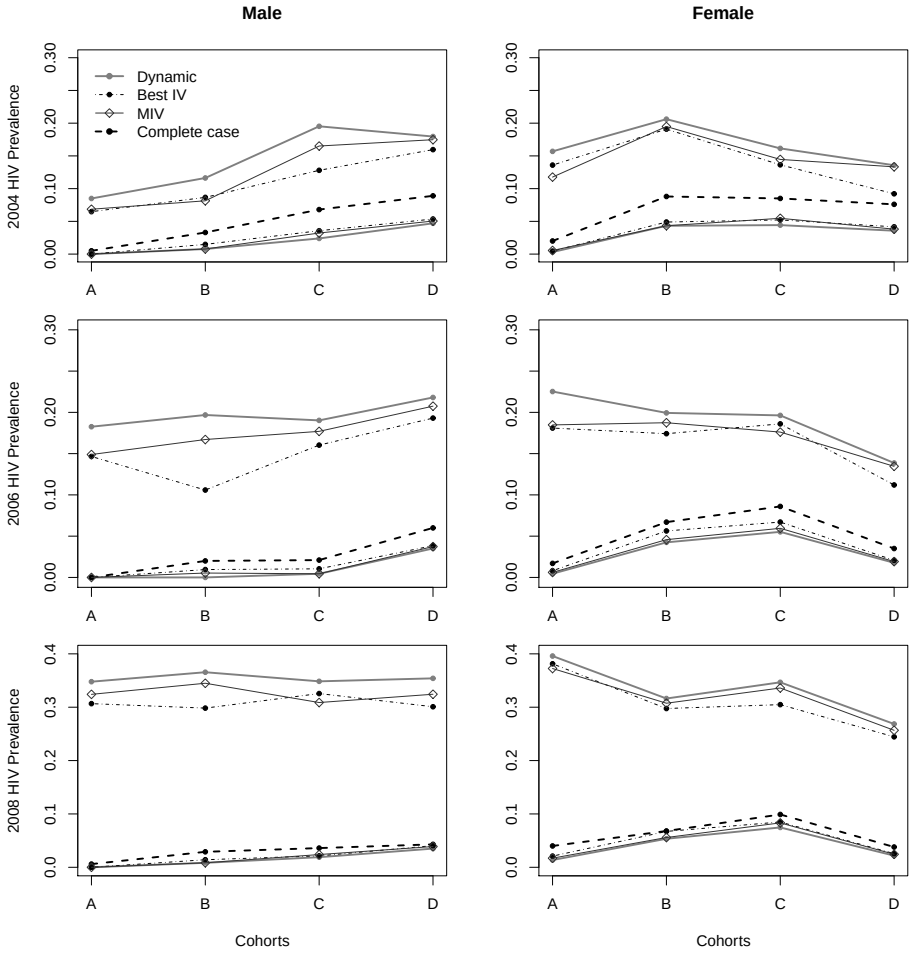

Figure 2. HIV prevalences for unit respondents by year, gender and cohort. The graph shows complete-case estimates, and bootstrapped dynamic bounds in the benchmark case, when the best IV or MIV restrictions are applied. Cohort A is the cohort born in 1984-1989, cohort B is the cohort born in 1975-1983, cohort C is the cohort born in 1965-1974, and cohort D is the cohort born before 1965.

By itself, this is not enough to narrow the worst-case bounds. In fact, among the respondents in future waves, only the information about negative HIV status can be used to infer HIV status in the current wave, so only the upper bound can be reduced relative to the worst-case. Similarly, information on past HIV status is helpful only if some of the nonrespondents in the current wave are found to be HIV positive in the past. In these cases, the availability of panel data helps

because it decreases the upper bound when future information is exploited and increases the lower bound when past information is exploited.

Applying our dynamic bounds to longitudinal data from Malawi, we obtain a reduction of the width of the worst-case bounds by about 18.2 percentage points in 2004, 13.2 percentage points in 2006, and 2.4 percentage points in 2008. Introducing plausible IV and MIV restrictions helps to further narrow the bounds. Ignoring the missing data problem and only using the complete cases, would give a point estimate of HIV prevalence that is very close to our lower bound. This estimate may be too optimistic because, according to our bounds, HIV prevalence could be much higher.

Our approach is easy to implement, it does not require assumptions about the nature of the missing data mechanism, and it allows to obtain relatively small and precisely estimated intervals for HIV prevalence. It could also be used for other applications where panel data are available and credible restrictions may be placed on the transition probabilities for the outcome of interest.

Our results confirm the importance of keeping low the nonresponse rates, and to consider unit and item nonresponse separately. They also illustrate the importance of including in the data information on interviewers' characteristics, fieldwork procedures etc, as these variables can be used as Instrumental Variables.

6. Supporting Material

Table 4. Number of observations and complete-case estimates by survey year, gender and cohort.

Cohort	Gender	†	*2004*	*2006*	*2008*
All	All	n	4008‡	3926	3733
		$Prev^{cc}$	0.062	0.049	0.051
A	Male	n	404	400	398
		$Prev^{cc}$	0.003	0.000	0.011
B	Male	n	374	359	355
		$Prev^{cc}$	0.029	0.015	0.040
C	Male	n	398	385	338
		$Prev^{cc}$	0.060	0.035	0.040
D	Male	n	691	662	636
		$Prev^{cc}$	0.094	0.056	0.045
A	Female	n	474	473	471
		$Prev^{cc}$	0.015	0.020	0.042
B	Female	n	560	559	552
		$Prev^{cc}$	0.092	0.069	0.070
C	Female	n	530	528	439
		$Prev^{cc}$	0.082	0.105	0.098
D	Female	n	577	560	544
		$Prev^{cc}$	0.079	0.040	0.038

†Prev^{cc} is the complete-case estimate of the prevalence.
‡The total number of individuals is 4,008 instead of 4,062 because we drop 54 individuals for which age is missing.

Table 5. 2004 Bootstrapped bounds for unit respondents.

Cohort	*Gender*		*Benchmark*		*IV Diff Gender*		*IV Experience*		*IV Age*		*IV Month*		*MIV*	
†		‡	Worst	Dyn	Worst	Dyn	Worst	Dyn	Worst	Dyn	Worst	Dyn	Worst	Dyn
All	All	L	0.041	0.041	0.043	0.043	0.043	0.044	0.045	0.044	0.042	0.043	0.042	0.042
(*n=2758*)		U	0.274	0.136	0.268	0.133	0.264	0.125	0.269	0.129	0.223	0.120	0.273	0.130
Prev^{cc}=0.062		W	0.233	0.095	0.225	0.090	0.221	0.081	0.224	0.085	0.181	0.078	0.230	0.087
A	Male	L	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(*n=259*)		U	0.209	0.085	0.201	0.077	0.185	0.065	0.196	0.078	0.151	0.073	0.202	0.069
Prev^{cc}=0.005		W	0.209	0.085	0.201	0.077	0.185	0.065	0.196	0.078	0.151	0.073	0.202	0.069
B	Male	L	0.012	0.008	0.014	0.014	0.015	0.015	0.014	0.014	0.012	0.013	0.011	0.008
(*n=258*)		U	0.264	0.116	0.250	0.107	0.231	0.087	0.227	0.100	0.225	0.095	0.244	0.081
Prev^{cc}=0.033		W	0.252	0.109	0.236	0.093	0.217	0.072	0.214	0.086	0.213	0.082	0.232	0.074
C	Male	L	0.024	0.024	0.028	0.028	0.029	0.030	0.033	0.036	0.029	0.030	0.031	0.032
(*n=251*)		U	0.406	0.195	0.330	0.182	0.364	0.177	0.342	0.128	0.384	0.181	0.383	0.165
Prev^{cc}=0.068		W	0.382	0.171	0.302	0.153	0.334	0.147	0.308	0.092	0.355	0.151	0.352	0.133
D	Male	L	0.049	0.047	0.054	0.054	0.053	0.054	0.052	0.052	0.056	0.057	0.050	0.050
(*n=468*)		U	0.321	0.180	0.302	0.160	0.309	0.177	0.308	0.169	0.300	0.167	0.310	0.175
Prev^{cc}=0.089		W	0.271	0.133	0.248	0.106	0.256	0.122	0.255	0.117	0.244	0.110	0.260	0.124
A	Female	L	0.003	0.003	0.007	0.004	0.006	0.006	0.005	0.006	0.005	0.005	0.006	0.006
(*n=325*)		U	0.317	0.157	0.301	0.142	0.292	0.140	0.296	0.145	0.216	0.136	0.298	0.118
Prev^{cc}=0.020		W	0.314	0.154	0.294	0.138	0.286	0.134	0.290	0.140	0.212	0.131	0.292	0.112
B	Female	L	0.043	0.043	0.048	0.049	0.047	0.049	0.048	0.048	0.047	0.049	0.043	0.043
(*n=393*)		U	0.364	0.206	0.348	0.198	0.348	0.191	0.348	0.192	0.339	0.197	0.351	0.195
Prev^{cc}=0.088		W	0.321	0.163	0.300	0.149	0.301	0.142	0.300	0.144	0.291	0.148	0.308	0.152
C	Female	L	0.047	0.044	0.052	0.053	0.060	0.058	0.052	0.052	0.052	0.052	0.055	0.055
(*n=384*)		U	0.292	0.161	0.275	0.152	0.276	0.151	0.266	0.155	0.230	0.136	0.270	0.145
Prev^{cc}=0.085		W	0.245	0.117	0.224	0.098	0.216	0.093	0.214	0.103	0.178	0.084	0.215	0.090
D	Female	L	0.036	0.036	0.043	0.043	0.041	0.042	0.041	0.041	0.042	0.042	0.039	0.038
(*n=420*)		U	0.262	0.136	0.252	0.128	0.249	0.118	0.249	0.121	0.225	0.092	0.255	0.133
Prev^{cc}=0.076		W	0.226	0.100	0.209	0.085	0.208	0.076	0.207	0.080	0.183	0.050	0.216	0.095

†Prev^{cc} is the complete-case estimate of the prevalence.

‡L is the Lower bound, U the Upper bound and W the width.

Table 6. 2006 Bootstrapped bounds for unit respondents.

Cohort	*Gender*		*Benchmark*		*IV Diff Gender*		*IV Experience*		*IV Age*		*IV Month*		*MIV*	
†		‡	Worst	Dyn	Worst	Dyn	Worst	Dyn	Worst	Dyn	Worst	Dyn	Worst	Dyn
All	All	L	0.029	0.035	0.032	0.040	0.032	0.037	0.034	0.040	0.031	0.037	0.030	0.036
(*n=2236*)		U	0.225	0.166	0.215	0.160	0.217	0.158	0.214	0.157	0.193	0.151	0.220	0.166
Prev^{cc}=0.044		W	0.196	0.131	0.183	0.120	0.185	0.121	0.180	0.117	0.161	0.114	0.190	0.130
A	Male	L	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(*n=230*)		U	0.278	0.183	0.260	0.170	0.261	0.168	0.256	0.170	0.214	0.147	0.226	0.149
Prev^{cc}=0.000		W	0.278	0.183	0.260	0.170	0.261	0.168	0.256	0.170	0.214	0.147	0.226	0.149
B	Male	L	0.000	0.000	0.000	0.007	0.000	0.000	0.000	0.010	0.000	0.010	0.000	0.005
(*n=188*)		U	0.261	0.197	0.246	0.186	0.215	0.161	0.235	0.169	0.161	0.106	0.212	0.167
Prev^{cc}=0.020		W	0.261	0.197	0.246	0.179	0.215	0.161	0.235	0.159	0.161	0.096	0.212	0.162
C	Male	L	0.004	0.004	0.006	0.006	0.007	0.010	0.009	0.009	0.007	0.010	0.004	0.004
(*n=226*)		U	0.235	0.190	0.220	0.180	0.219	0.174	0.213	0.168	0.180	0.160	0.201	0.177
Prev^{cc}=0.021		W	0.230	0.186	0.214	0.174	0.212	0.164	0.204	0.159	0.173	0.150	0.197	0.173
D	Male	L	0.030	0.035	0.037	0.042	0.034	0.038	0.036	0.040	0.037	0.039	0.033	0.037
(*n=399*)		U	0.263	0.218	0.241	0.202	0.241	0.193	0.250	0.207	0.245	0.205	0.255	0.208
Prev^{cc}=0.060		W	0.233	0.183	0.204	0.161	0.208	0.155	0.214	0.167	0.209	0.166	0.223	0.170
A	Female	L	0.000	0.005	0.000	0.008	0.000	0.007	0.000	0.008	0.000	0.008	0.000	0.006
(*n=222*)		U	0.284	0.225	0.266	0.215	0.250	0.192	0.265	0.208	0.229	0.181	0.247	0.185
Prev^{cc}=0.017		W	0.284	0.221	0.266	0.207	0.250	0.185	0.265	0.200	0.229	0.173	0.247	0.179
B	Female	L	0.033	0.042	0.037	0.055	0.040	0.051	0.041	0.056	0.039	0.051	0.033	0.046
(*n=306*)		U	0.268	0.199	0.255	0.184	0.252	0.187	0.250	0.174	0.252	0.185	0.251	0.187
Prev^{cc}=0.067		W	0.235	0.157	0.218	0.129	0.212	0.136	0.209	0.118	0.213	0.134	0.218	0.142
C	Female	L	0.049	0.055	0.056	0.062	0.055	0.063	0.062	0.067	0.057	0.060	0.054	0.059
(*n=326*)		U	0.233	0.196	0.203	0.181	0.219	0.185	0.222	0.186	0.218	0.179	0.213	0.176
Prev^{cc}=0.086		W	0.184	0.141	0.147	0.119	0.164	0.123	0.161	0.119	0.160	0.120	0.159	0.117
D	Female	L	0.012	0.018	0.019	0.021	0.019	0.021	0.017	0.022	0.018	0.021	0.015	0.019
(*n=339*)		U	0.221	0.139	0.202	0.126	0.209	0.127	0.196	0.124	0.205	0.112	0.214	0.135
Prev^{cc}=0.035		W	0.209	0.121	0.184	0.105	0.190	0.105	0.179	0.102	0.187	0.091	0.199	0.115

†Prev^{cc} is the complete-case estimate of the prevalence.

‡L is the Lower bound, U the Upper bound and W the width.

Table 7. 2008 Bootstrapped bounds for unit respondents.

Cohort	*Gender*		*Benchmark*		*IV Month*		*MIV*	
†		‡	Worst	Dyn	Worst	Dyn	Worst	Dyn
All	All	L	0.030	0.043	0.033	0.047	0.032	0.044
(n=*2825*)		U	0.305	0.306	0.268	0.265	0.303	0.303
$Prev^{cc}$=*0.050*		W	0.276	0.262	0.235	0.218	0.272	0.259
A	Male	L	0.000	0.000	0.000	0.000	0.000	0.000
(n=*253*)		U	0.348	0.348	0.303	0.307	0.324	0.324
$Prev^{cc}$=*0.006*		W	0.348	0.348	0.303	0.307	0.324	0.324
B	Male	L	0.004	0.008	0.008	0.014	0.004	0.008
(n=*238*)		U	0.357	0.366	0.298	0.298	0.346	0.345
$Prev^{cc}$=*0.029*		W	0.353	0.357	0.290	0.284	0.341	0.337
C	Male	L	0.011	0.019	0.015	0.021	0.015	0.024
(n=*264*)		U	0.341	0.348	0.324	0.326	0.307	0.309
$Prev^{cc}$=*0.036*		W	0.330	0.330	0.309	0.304	0.292	0.285
D	Male	L	0.018	0.035	0.020	0.041	0.018	0.039
(n=*514*)		U	0.354	0.354	0.300	0.301	0.325	0.324
$Prev^{cc}$=*0.043*		W	0.337	0.319	0.280	0.260	0.307	0.285
A	Female	L	0.010	0.014	0.012	0.021	0.014	0.017
(n=*293*)		U	0.396	0.396	0.376	0.381	0.369	0.373
$Prev^{cc}$=*0.040*		W	0.386	0.382	0.365	0.360	0.354	0.355
B	Female	L	0.033	0.053	0.043	0.067	0.037	0.056
(n=*430*)		U	0.314	0.316	0.295	0.298	0.306	0.307
$Prev^{cc}$=*0.068*		W	0.281	0.263	0.253	0.230	0.269	0.252
C	Fgemale	L	0.053	0.075	0.058	0.085	0.058	0.083
(n=*375*)		U	0.344	0.347	0.305	0.305	0.336	0.336
$Prev^{cc}$=*0.099*		W	0.291	0.272	0.246	0.220	0.279	0.253
D	Female	L	0.015	0.022	0.021	0.026	0.017	0.024
(n=*458*)		U	0.269	0.269	0.245	0.244	0.258	0.257
$Prev^{cc}$=*0.038*		W	0.253	0.247	0.224	0.219	0.240	0.232

†$Prev^{cc}$ is the complete-case estimate of the prevalence

‡L is the Lower bound, U the Upper bound and W the width.

A. Appendix: Dynamic bounds for arbitrary number of waves

Consider bounding HIV prevalence at time t in the general case when several waves of a panel survey are available, either before or after wave t.

A.1. F waves after t

With information on F waves after wave t, the lower bound on π_t does not change while the upper bound is characterised by the following recursion

$$\begin{aligned}
t: \quad & UB_t, \\
t, t+1: \quad & UB_t^{(+1)} = UB_t - \Pr(Y_{t+1}=0, D_{t+1}=1, D_t=0), \\
t, t+1, t+2: \quad & UB_t^{(+2)} = UB_t^{(+1)} - \Pr(Y_{t+2}=0, D_{t+2}=1, D_{t+1}=0, D_t=0), \\
& \dots \\
t, \dots, t+F: \quad & UB_t^{(+F)} = UB_t^{(+(F-1))} - \\
& \quad - \Pr(Y_{t+F}=0, D_{t+F}=1, D_{t+F-1}=0, \dots, D_t=0).
\end{aligned}$$

Thus we obtain

$$LB_t^{(+F)} = LB_t,$$

$$UB_t^{(+F)} = UB_t - \sum_{f=1}^{F} \Pr(Y_{t+f}=0, D_{t+f}=1, D_{t+f-1}=0, \dots, D_{t+1}=0, D_t=0),$$

and

$$W_{t(+F)} = W_t - \sum_{f=1}^{F} \Pr(Y_{t+f}=0, D_{t+f}=1, D_{t+f-1}=0, \dots, D_{t+1}=0, D_t=0).$$

It is easy to see that increasing the number of future waves decreases the width of the identification region.

A.2. P waves before t

With information on P waves before wave t, the upper bound does not change while the lower bound is characterized by the following recursions

$$\begin{aligned}
t: \quad & LB_t, \\
t-1, t: \quad & LB_t^{(-1)} = LB_t + \Pr(Y_{t-1}=1, D_{t-1}=1, D_t=0), \\
t-2, t-1, t: \quad & LB_t^{(-2)} = LB_t^{(-1)} + \Pr(Y_{t-2}=1, D_{t-2}=1, D_{t-1}=0, D_t=0), \\
& \dots \\
t-P, \dots, t: \quad & LB_t^{(-P)} = LB_t^{(-(P-1))} + \\
& \quad + \Pr(Y_{t-P}=1, D_{t-P}=1, D_{t-P+1}=0, \dots, D_t=0).
\end{aligned}$$

Thus we obtain

$$LB_t^{(-P)} = LB_t + \sum_{p=1}^{P} \Pr(Y_{t-p} = 1, D_{t-p} = 1, D_{t-p+1} = 0, \ldots, D_{t-1} = 0, D_t = 0),$$

$$UB_t^{(-P)} = UB_t,$$

and

$$W_t^{(-P)} = W_t - \sum_{p=1}^{P} \Pr(Y_{t-p} = 1, D_{t-p} = 1, D_{t-p+1} = 0, \ldots, D_{t-1} = 0, D_t = 0).$$

It is easy to see that increasing the number of past waves decreases the width of the identification region.

A.3. P waves before and F waves after t

Combining the previous results gives

$$\begin{aligned} LB_t^{(-P,+F)} &= LB_t^{(-P)} \\ &= LB_t + \sum_{p=1}^{P} \Pr(Y_{t-p} = 1, D_{t-p} = 1, D_{t-p+1} = 0, \ldots, D_{t-1} = 0, D_t = 0), \\ UB_t^{(-P,+F)} &= UB_t^{(+F)} \\ &= UB_t - \sum_{f=1}^{F} \Pr(Y_{t+f} = 0, D_{t+f} = 1, D_{t+f-1} = 0, \ldots, D_{t+1} = 0, D_t = 0), \end{aligned}$$

and

$$\begin{aligned} W_t^{(-P,+F)} = W_t &- \sum_{p=1}^{P} \Pr(Y_{t-p} = 1, D_{t-p} = 1, D_{t-p+1} = 0, \ldots, D_{t-1} = 0, D_t = 0) - \\ &- \sum_{f=1}^{F} \Pr(Y_{t+f} = 0, D_{t+f} = 1, D_{t+f-1} = 0, \ldots, D_{t+1} = 0, D_t = 0). \end{aligned}$$

Acknowledgments

We thank Hans-Peter Kohler and Philip Anglewicz for providing the data from the Malawi Diffusion and Ideational Change Project (MDICP) and for support. The MDICP has been funded by the following grants: ICHD R01HD053781, NICHD R01 HD044228, NICHD R01HD/MH41713. We also thank seminar participants to the Italian Congress of Econometrics and Empirical Economics

2011 Conference (ICEEE, Pisa, Italy, January 19-21), the Understanding Society/BHPS Conference 2011 (University of Essex, Colchester, UK, June 30 - July 1) and the International Statistical Institute Conference 2011 (ISI, Dublin, Ireland, August 21-26, 2011).

References

Anglewicz, P., 2007. *Migration, HIV Infection, and Risk Perception in Malawi*. Ph.D. thesis, Philadelphia: Graduate Group in Demography, Population Studies Center, University of Pennsylvania.

Barnighausen, T., J. Bor, S. Wandira-Kazibwe, and D. Canning, 2011. Correcting hiv prevalence estimates for survey non-participation: An application of heckman-type selection models to the zambian demographic and health survey. *Epidemiology*, 22(1):27–35.

Boerma, J., P. Ghys, and N. Walker, 2003. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet*, 363(9399):1929–1931.

Brookmeyer, R., 2010. Measuring the HIV/AIDS Epidemic: Approaches and Challenges. *Epidemiologic Reviews*, 32:26–37.

Crampin, A. C., J. R. Glynn, B. M. M. Ngwira, F. D. Mwaungulu, J. M. Ponnighauss, D. K. Warndorff, and P. Fine, 2003. Trends and measurement of HIV prevalence in Northern Malawi. *AIDS*, 17:1817–1825.

Garcia-Calleja, J., E. Gouws, and P. Ghys, 2006. National population based hiv prevalence surveys in sub-saharan africa: Results and implications for hiv and aids estimates. *Sexually Transmitted Infections*, 82(Suppl III):iii64–iii70.

Gouws, E., V. Mishra, and T. Fowler, 2008. Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalized epidemics: Implications for calibrating surveillance data. *Sexually Transmitted Infections*, 84(Suppl 1):i17–i23.

Heckman, J. J., 1979. Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Horowitz, J. L. and C. F. Manski, 1998. Censoring of Outcomes and Regressors Due to Survey Nonresponse: Identification and Estimation Using Weights and Imputation. *Journal of Econometrics*, 84:37–58.

Kreider, B. and J. V. Pepper, 2007. Disability and employment: Reevaluating the evidence in light of reporting errors. *Journal of American Statistical Association*, 102(478):432–441.

Lachaud, J. P., 2007. HIV prevalence and poverty in Africa: Micro- and macroeconometric evidences applied to Burkina Faso. *Journal of Health Economics*, 26:483–504.

Lepkowski, J. M. and M. P. Couper, 2002. *Survey Nonresponse*, chapter Nonresponse in Longitudinal Household Surveys, pages 259–272. eds R.M. Groves, D. Dillman, J. Eltinge, and R. Little, New York: Wiley.

Little, J. A. and D. B. Rubin, 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Manski, C. F., 1989. Anatomy of the Selection Problem. *Journal of Human Resources*, 24:343–360.

Manski, C. F., 1994. The Selection Problem. In C. Sims and C. C. U. Press, editors, *Advances in Econometrics, Sixth World Congress*, pages 143–170.

Manski, C. F., 1995. *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA.

Manski, C. F., 2003. *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. F. and J. Pepper, 2000. Monotone Instrumental Variables with an Application to the Returns to Schooling. *Econometrica*, 68:997–1010.

Marston, M., K. Harriss, and E. Slaymaker, 2008. Nonresponse bias in estimates of HIV prevalence due to the mobility of absentees in national population-based surveys: a study of nine national surveys. *Sexually Transmitted Infections*, 84(1):i71–i77.

Martin-Herz, S., A. Shetty, M. Bassett, C. Ley, M. Mhazo, S. Moyo, A. Herz, and D. Katzenstein, 2006. Perceived risks and benefits of hiv testing, and predictors of acceptance of hiv counseling and testing among pregnant women in Zimbabwe. *International Journal of Sexually Transmitted Diseases and AIDS*, 17:835–841.

Mishra, V., B. Barrere, R. Hong, and S. Khan, 2008. Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexually Transmitted Infections*, 84(Suppl I):i63–i70.

Montana, L., V. Mishra, and R. Hong, 2008. Measuring the HIV/AIDS Epidemic: Approaches and Challenges. *Sexually Transmitted Infections*, 84(1):i78–i84.

Nicoletti, C. and F. Peracchi, 2006. The effects of income imputation on micro analyses: Evidence from the ECHP. *Journal of the Royal Statistical Society, Series A*, 169:625–646.

Nicoletti, C., F. Peracchi, and F. Foliano, 2011. Estimating income poverty in the presence of missing data and measurement error. *Journal of Business and Economic Statistics*, 29(1):61–72.

Obare, F., 2010. Nonresponse in Repeat Population-Based Voluntary Counseling and Testing for HIV in Rural Malawi. *Demography*, 47(3):651–665.

Obare, F., P. Fleming, R. Anglewicz, R. Thornton, F. Martinson, A. Kapatuka, M. Poulin, S. Watkins, and H. Kohler, 2009. Acceptance of repeat population-based voluntary counselling and testing for HIV in rural Malawi. *Sexually transmitted infections*, 85(139).

Reniers, G. and J. Eaton, 2009. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS*, 23(5):621–629.

Rubin, D. B., 1976. Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B., 1989. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sakarovitch, C., A. Alioum, D. Ekouevi, P. Msellati, V. Leroy, and F. Dabis, 2007. Estimating incidence of HIV infection in childbearing age African women using serial prevalence data from antenatal clinics. *Stat Med*, 26:320–335.

Thornton, R., 2008. The demand for, and impact of, learning HIV status. *American Economic Review*, 98:1829–1863.

Vella, F., 1998. Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources*, 33:127–169.

Watkins, S. C., E. M. Zulu, H. P. Kohler, and J. R. Behrman, 2003. Introduction to: Social interactions and HIV/AIDS in rural Africa. *Demographic Research*, Special Collection 1(1):1–30.