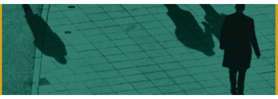
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2014)
177, *Part* 3, *pp.* 587–606

Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random

Bruno Arpino,

Universitat Pompeu Fabra, Barcelona, Spain

Elisabetta De Cao

University of Groningen, The Netherlands

and Franco Peracchi

University of Rome 'Tor Vergata' and Einaudi Institute for Economics and Finance, Rome, Italy

[Received October 2011. Final revision May 2013]

Summary. Population-based surveys are often considered the 'gold standard' to estimate the prevalence of human immunodeficiency virus (HIV) but typically suffer from serious missing data problems. This causes considerable uncertainty about HIV prevalence. Following the partial identification approach, we produce worst-case bounds for HIV prevalence. We then exploit the availability of panel data and the absorbing nature of HIV infection to narrow the width of these bounds. Applied to panel data from rural Malawi, our approach considerably reduces the width of the worst-case bounds. It also allows us to check the credibility of the additional assumptions that are imposed by methods that point-identify HIV prevalence.

Keywords: Human immunodeficiency virus prevalence; Malawi Diffusion and Ideational Change Project data; Non-ignorable non-response; Panel data; Partial identification

1. Introduction

The prevalence of human immunodeficiency virus (HIV) in a population is the fraction of people who are infected or, equivalently, the probability that a randomly drawn individual has the disease. Credible estimates of the prevalence of HIV are essential for policy makers to plan control programmes and interventions.

When available, administrative data could be used to identify those who are infected accurately (see, for example, Antoniou *et al.* (2011)). However, in countries with generalized epidemics, national HIV estimates have until recently relied mostly on data generated by surveillance systems based on a selected number of sentinel antenatal clinics (UNAIDS–World Health Organization, 2003). On the basis of 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 (Brookmeyer, 2010). Antenatal clinic data contain several sources of bias. First, they exclude men and

Address for correspondence: Elisabetta De Cao, Department of Economics, Econometrics and Finance, University of Groningen, Nettelbosje 2, 9747 AE Groningen, The Netherlands.
E-mail: e.de.cao@rug.nl

© 2013 Royal Statistical Society 0964–1998/14/177587

are representative only of sexually active women who are pregnant and attend a clinic. Second, they may provide biased estimates even for the subpopulation of pregnant women because of selective location of the clinics, mostly concentrated in urban areas. As a result, estimates of HIV prevalence based on antenatal clinic attendance may be substantively biased upwards (Gouws *et al.*, 2008; Montana *et al.*, 2008; Reniers and Eaton, 2009).

In recent years, several population-based surveys have begun, including modules that collect biomarkers that are useful to test for HIV, such as blood samples or saliva swabs. These surveys are an important new source of data because they accurately measure HIV status and, unlike surveys based on antenatal clinic attendance are not restricted to a selected subpopulation. Estimates of HIV prevalence obtained from these surveys are, in general, considerably lower than those obtained from antenatal clinic attendance data (Gouws *et al.*, 2008; Montana *et al.*, 2008). On the basis of 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), they may be affected by a different but not necessarily less severe source of bias, namely missing data on respondents' HIV status, mainly due to refusal to take the HIV test or to temporary absence or migration of those targeted for interview.

Approaches that discard cases with missing HIV status (complete-case analysis) implicitly rely on the assumption that data are missing completely at random (Rubin, 1976) or, if there are covariates, on the weaker assumption that non-response is independent of the outcome of interest given the covariates. If the covariates are completely observed, then this assumption is equivalent to missingness at random (MAR). Imputation or weighting techniques are frequently used to produce estimates of HIV prevalence. However, if the missing data mechanism depends on true HIV status, then the missingness completely at random (MCAR) and MAR assumptions are violated and methods based on these assumptions are likely to produce biased estimates of HIV prevalence.

In fact, there is evidence that people refusing to be tested have higher risk of being HIV infected (Reniers and Eaton, 2009). This risk has also been found to be higher for those who are not interviewed because of migration (Marston *et al.*, 2008; Crampin *et al.*, 2003; Obare, 2010). Anglewicz (2012) analysed 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 found that migrants are likely to report a higher number of sexual partners and are more likely to be HIV infected. 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. In these cases, HIV prevalence estimates based on the MAR assumption may be severely biased and the analyst should explicitly acknowledge the possibility that data are missing not at random (MNAR).

Recently, a few studies have employed the approach that was pioneered by Heckman (1979) to estimate HIV prevalence under the assumption of data MNAR by modelling survey non-response as a function of unobservable factors that also affect HIV status (Lachaud, 2007; Reniers and Eaton, 2009; Bärnighausen *et al.*, 2011). This approach combines a description of the missing data process with strong parametric assumptions about the distribution of the unobservables in the model, such as joint normality. To identify the model parameters credibly, it also requires exclusion restrictions, namely variables that help to explain the missing data process but not the outcome of interest. For example, Bärnighausen *et al.* (2011) used data from the Zambia Demographic and Health Survey, where 28% of men did not participate in HIV testing, and found that the estimate of the prevalence of HIV in males is only 12% when based on imputed data but it goes up to 21% when using a Heckman-type sample selection model. Similarly to other studies (Nicoletti and Peracchi, 2005), their exclusion restrictions consist of characteristics of the interview process, as they help to predict survey participation but have arguably no direct effect on HIV status.

Note that, whereas the MCAR assumption can be tested against specific MAR models and is often rejected by the data, the MAR assumption cannot be tested against the MNAR alternative because one can always find models in each class that fit the observed data equally well (Molenberghs *et al.*, 2008).

In this paper, we allow the data to be MNAR but, instead of adopting a specific model for the non-response mechanism, we ask what can be learned about HIV prevalence without imposing strong untestable assumptions, such as those commonly made in sample selection models. Following Manski (1995, 2003) and Horowitz and Manski (1998), we switch the focus away from point identification to partial identification. The second approach explicitly recognizes ambiguity by identifying the set of values, or identification region, to which the parameter of interest (HIV prevalence in our case) must necessarily belong given the available data and the assumptions maintained. If these assumptions are sufficiently strong, the identification region collapses to a single point and the parameter of interest is point identified.

We first use the cross-sectional evidence alone for partial identification of HIV prevalence. We then exploit the availability of panel data and the absorbing nature of HIV infection to narrow the width of the initial identification region. Although additional assumptions, such as instrumental variable (IV) and monotone instrumental variable (MIV) restrictions, may be used to narrow the width of the identification region further, our main contribution is to show the power of combining substantive information about the HIV process with the longitudinal nature of the available data. One advantage of the partial identification approach is that practitioners can examine the credibility of point estimates obtained under alternative assumptions by checking whether they lie within the identification region (Nicoletti, 2010). In particular, we consider point estimates obtained by using the complete-case approach, propensity score weighting and a Heckman-type estimator.

Our data are from the 2004, 2006 and 2008 waves of the Malawi Diffusion and Ideational Change Project (MDICP), which is a longitudinal survey of the population of rural Malawi. Malawi is one of the African countries that is most affected by the HIV epidemic, with acquired immune deficiency syndrome as the leading cause of death among adults (United Nations General Assembly Special Session, 2010). The complete-case estimate of the national HIV prevalence rate, based on the 2004 Malawi Demographic and Health Survey (MDHS), is 11.8% for people aged 15–49 years. As for most countries in sub-Saharan Africa, where HIV is mainly transmitted via heterosexual contact, HIV prevalence is estimated to be higher for women (13.3%, against 10.2% for men) and in urban areas (17.1%, against 10.8% in rural areas). We take 2004 as our baseline because the basic demographic characteristics of the 2004 MDICP are very similar to those of the 2004 MDHS for rural Malawi. Although the MDICP is not representative of the entire Malawian population (urban and rural), it has the key advantage over the MDHS of being a longitudinal survey, and not just a repeated cross-section. It is important to note, however, that the two data sets are not directly comparable because the target population is different. In particular, unlike the MDICP, the MDHS for rural Malawi includes periurban areas, namely areas that are immediately around urban settlements. Hence, a straight comparison of the 2004 HIV prevalences obtained by using the MDICP or MDHS would be inappropriate.

The remainder of this paper is organized 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 width of the initial identification region on the basis of empirical evidence alone. It also discusses how to use plausible IV and MIV restrictions to narrow the identification region further. Section 4 presents our empirical results, broken down by region, gender and cohort. Finally, Section 5 offers some conclusions.

The data that are analysed in the paper and the program that was used to analyse them can be obtained from

```
http://wileyonlinelibrary.com/journal/rss-datasets
```

2. The Malawi Diffusion and Ideational Change Project

The MDICP is a longitudinal survey that has been conducted in rural Malawi every 2 years since 1998. The survey is the result of the collaboration between the University of Pennsylvania and the College of Medicine and Chancellor College at the University of Malawi. The resulting data can be freely downloaded from `http://www.malawi.pop.upenn.edu` and include the outcomes of HIV tests for the years 2004, 2006 and 2008.

The MDICP 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 centre and Rumphi in the north. The three regions are very different in terms of ethnic composition, language, religious practice, population density, literacy and prevailing social system (e.g. patrilocal or matrilocal residence). In the next section we provide a brief description of the survey design and refer to Anglewicz *et al.* (2009) and the MDICP Web site for more details.

2.1. Survey design

The first wave of the survey was carried out in 1998. Two-stage sampling was used in each of the three districts, with a total of 145 villages randomly selected in the first stage. Household lists of people reported as being normally resident in the selected villages were compiled by the MDICP team in the week before fieldwork. Then, in the second stage, a sample of eligible women was randomly selected from these lists. In total, 1541 ever-married women of child-bearing age (15–50 years) and 1198 men (most of them husbands of the married women in the sample, and the rest an oversample to compensate for an unexpectedly large number of men who were away) were interviewed. The second wave, carried out in 2001, followed-up the respondents and interviewed the spouses of respondents who were married between the first and the second wave (Watkins *et al.*, 2003).

The third wave, carried out in 2004, is important because it augmented the original sample with a random sample of about 1500 people aged 15–28 years (both married and never married) to correct for aging of the baseline sample and the fact that it was restricted to ever-married women and their husbands. Since HIV-infected people may have a lower probability of marrying, using this augmented sample reduces the risk of underestimating HIV prevalence in the target population.

The fourth (2006) and fifth (2008) waves added the spouses of newly married respondents. In addition to the spouses, the 2008 wave also included all living biological parents who lived in the same village as the respondent. This new sample of about 800 parents was based on family listings obtained from 2006 respondents. The percentage of targeted people who were successfully interviewed in a specific year (extra sample included) was 67.0% in 2004, 67.9% in 2006 and 67.4% in 2008 (Kohler *et al.*, 2009).

The survey collects extensive information on household structure, health, risk assessments, sexual relations, marriage and partnership histories, intergenerational and interfamiliar transfers, as well as income and various measures of wealth. It also collects information on village level variables, regional market prices and weather conditions. The survey instrument was translated from English into 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 biomarker module called the voluntary consulting and test (VCT) survey was 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 acquired immune deficiency syndrome related questions, and free tests for HIV and other sexually transmitted infections that were administered by nurses from outside the area. Respondents to the VCT survey are also offered pretest 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 by using an improved testing procedure consisting of a rapid response blood test. According to the available documentation, the rapid blood test has a 100% probability of detecting true positive cases (99.9% for the oral saliva test) and a high estimated accuracy (1 minus the probability of false positive results) of 99.8% (99.5% for the oral saliva test in a high risk population) (see `http://www.fda.gov` for details about the tests: OraSure and Uni-Gold Recombigen HIV). Finally, measurement error in the two types of test (oral swabs and blood test) appears to be small and purely random.

We focus on people who were interviewed in 2004, excluding new entrants in 2006 and 2008, and dropping from the sample people who were never successfully contacted. We consider 2004 as our baseline, not only because it is the first year that biomarkers are available but also because the basic demographic characteristics of the 2004 MDICP are very similar to those of the 2004 MDHS in rural areas (National Statistical Office Malawi and Opinion Research Corporation Macro, 2005; Thornton, 2008). We decided to exclude new entrants (mainly new spouses of the respondents) because they do not enter the sample randomly but are selectively chosen. Because prevalence is defined for the population of living individuals, our working sample consists of 4062 people who were alive in 2004. When computing HIV prevalence for 2006 and 2008, we exclude people who died after 2004.

2.2. *Missing data*

Because of unit and item non-response, HIV status is missing for a substantial fraction of the sample in each of the three waves that were considered. Unit non-response occurs when eligible sample units do not participate in the survey owing to failure to establish a contact or refusal to co-operate. Since the survey consists of the main and the VCT surveys, we define unit non-response as the case when both parts are missing. Item non-response occurs when responding units do not provide useful answers to particular items of the questionnaire. Because our item of interest is HIV status, we focus on item non-response to the biomarker module.

There are different patterns of unit non-response across waves. About 55% of the sample are unit respondents in all three waves; about 12% are unit respondents in 2004 but not in 2006 and 2008 (attritors after 2004); about 11% are unit respondents in 2004 and 2006 but not in 2008 (attritors after 2006); about 8% are unit respondents in 2004 and 2008 but not in 2006; the remaining 14% includes other patterns of unit non-response.

Table 1 shows the various sources of missing data. The fraction with missing HIV status is 29% in 2004 and 37% in 2006 and reaches 43% in 2008, as a result of the increase in item non-response from 15% in 2004 to 19% in 2008 and the large increase in unit non-response from 15% in 2004 to 24% in 2008.

Table 1. Distribution of types of unit respondents and non-respondents by wave

Respondents	*Results for 2004*		*Results for 2006*		*Results for 2008*	
	Frequency	*%*	*Frequency*	*%*	*Frequency*	*%*
Unit respondents						
HIV negative	2700	66.5	2408	59.3	2116	52.1
HIV positive	177	4.4	123	3.0	117	2.9
Item non-respondents						
Test refused	256	6.3	200	4.9	172	4.2
Indeterminate	14	0.3	6	0.1	1	0.0
Results lost	24	0.6	0	0.0	0	0.0
Other†	319	7.9	313	7.7	569	14.0
Unit non-respondents						
Refused	27	0.7	11	0.3	58	1.4
Moved	184	4.5	479	11.8	470	11.6
Temporarily absent	36	0.9	41	1.0	76	1.9
Hospitalized	6	0.1	5	0.1	1	0.0
Other‡	319	7.9	432	10.6	359	8.8
Dead			44	1.1	123	3.0
Total§	4062	100.0	4062	100.0	4062	100.0

†People who completed the first part of the questionnaire but not the second, e.g. because they were temporarily absent during the biomarker collection.

‡People who did not complete the questionnaire for unknown reasons or because they were too old or too sick.

§New entrants between 2006 and 2008 have been excluded.

MDICP data provide information on the reasons for unit and item non-response. The main reason for unit non-response, and for its increase across waves, is migration. Hospitalization and refusal to participate are relatively unimportant. Other reasons for unit non-response are lumped into the residual category 'other', consisting mainly of people who did not fill in the questionnaire because they were too old or too sick, or for unspecified reasons that may also include migration. People who are unit non-respondents because of migration or other reasons are assumed to be alive when computing the bounds. This assumption has a limited effect on the estimates. In fact, the mortality rate estimated on the remaining sample is low (see Table 1). Moreover, this assumption does not influence the analysis when limited to unit respondents (see Section 4.3). In general, the consequence of this assumption is to have more conservative bounds. Imposing a mortality rate that is similar to the sample estimates would result, in fact, in a reduction of the proportion of missing data.

The main reason for item non-response is refusal to be tested. Note, however, that in 2004 the MDICP had a lower test refusal rate than the MDHS in rural areas (6.3% against 21.7%).

Low refusal rates (less than 5%) are also found in the 2006 and 2008 MDCIP, where rapid response blood tests are used to eliminate the time delay between testing and test results. The results of the HIV test are indeterminate or have been lost in fewer than 1% of the cases (Table 1). Other reasons for item non-response, lumped into the category other, consist of people who completed the main survey but not the VCT survey, e.g. because they were temporarily absent. The importance of this residual category almost doubled between 2004 and 2008.

Distinguishing between the different sources of missing data is important. Ignoring missing data due to migration or test refusal may bias HIV prevalence estimates downwards (Reniers and Eaton, 2009; Obare, 2010). In contrast, 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 human immunodeficiency virus prevalence

To formalize our problem, consider a population that, at a given time t, consists of living individuals who can be in two mutually exclusive states: susceptible to HIV or infected. A susceptible individual is anybody at risk of becoming infected by the disease. HIV status of a randomly selected individual at time t may be represented by a binary random variable Y_t, which is equal to 1 if the individual is infected and equal to 0 otherwise. HIV prevalence at time t is just the probability $\pi_t = \Pr(Y_t = 1)$ that a randomly selected individual is infected.

Our aim is to construct bounds for π_t when HIV status is missing for a fraction of individuals in the population.

3.1. Bounds with cross-sectional data

We first consider the problem of bounding HIV prevalence when data are available at only a single point in time, as for a cross-section or a single wave of a panel.

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

$$\pi_t = \Pr(Y_t = 1 | D_t = 1) \Pr(D_t = 1) + \Pr(Y_t = 1 | D_t = 0) \Pr(D_t = 0), \tag{1}$$

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

$$\begin{aligned}
\mathrm{LB}_t &= \Pr(Y_t = 1 | D_t = 1) \Pr(D_t = 1) \\
&= \Pr(Y_t = 1, D_t = 1), \\
\mathrm{UB}_t &= \Pr(Y_t = 1 | D_t = 1) \Pr(D_t = 1) + \Pr(D_t = 0) \\
&= \Pr(Y_t = 1, D_t = 1) + \Pr(D_t = 0).
\end{aligned}$$

These bounds are often referred to as worst-case bounds because they use only the available data and ignore any additional information that may potentially be available.

The identification region for π_t consists of all the points in the interval between LB_t and UB_t. The width $W_t = \mathrm{UB}_t - \mathrm{LB}_t$ of this interval is equal to the non-response probability $\Pr(D_t = 0)$, which therefore represents a direct measure of the uncertainty about HIV prevalence caused by non-response (Horowitz and Manski, 1998). Without non-response, there is no uncertainty about π_t. When non-response rates are high, as in our case, uncertainty is large. An important issue, therefore, is whether additional information about the HIV process may be exploited to narrow the worst-case bounds.

3.2. Bounds with panel data

HIV infection is an absorbing state: a person who is infected at any given time has probability 1 of being infected at later times, whereas a person who is not infected at any given time has probability 0 of being infected at earlier times.

These simple considerations help to narrow the worst-case bounds when panel data are available and the HIV status of non-respondents in one wave may be observed in earlier or later waves. We shall refer to the resulting bounds as 'dynamic' because they are based on longitudinal data and exploit restrictions that are implied by the dynamics of HIV epidemic. To keep things simple we present only the results for the case of short panels with only two waves. Appendix A of the on-line supplementary materials presents extra proofs and the results for the general case of a panel with $P \geqslant 1$ waves before wave t, or $F \geqslant 1$ waves after wave t, or both.

Suppose first that the two available panel waves are at times t and $t+1$. To narrow the worst-case bounds on π_t, consider again equation (1) and note that

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

where

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

since $\Pr(Y_t=1|Y_{t+1}=0,D_{t+1}=1,D_t=0)=0$. This is because someone who is HIV infected cannot recover (become uninfected), which is why infection is an 'absorbing' state. Thus, we can rewrite equation (1) as

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

From equation (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 0 and their upper bound of 1. Setting both probabilities equal to 0 gives the lower bound

$$\mathrm{LB}_t^{(+1)}=\mathrm{LB}_t,$$

whereas setting both of them equal to 1 gives the upper bound

$$\mathrm{UB}_t^{(+1)}=\mathrm{UB}_t-\Pr(D_t=0)\{1-\Pr(Y_{t+1}=1,D_{t+1}=1|D_t=0)-\Pr(D_{t+1}=1|D_t=0)\},$$

where the term in curly 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 0 and 1. The width of the resulting identification interval for π_t is

$$W_t^{(+1)}=\mathrm{UB}_t^{(+1)}-\mathrm{LB}_t^{(+1)}=W_t-\Pr(Y_{t+1}=0,D_{t+1}=1,D_t=0)\leqslant W_t,$$

where W_t is the width of the worst-case bounds. The inequality holds because $\Pr(Y_{t+1}=0,D_{t+1}=1,D_t=0)$ is non-negative and cannot exceed $\Pr(D_t=0)$. Note that simply knowing the HIV status at time $t+1$ of people with missing HIV status at time t is not enough to narrow the worst-case bounds. In fact, among the respondents at time $t+1$, only the information about negative HIV status can be used to infer HIV status at time t exactly, so only the upper bound can be reduced relative to the worst case. Respondents at time $t+1$ who are found to be HIV infected cannot be assumed to have been HIV infected already at time t, so the lower bound is the same as in the worst case.

If the two available panel waves are at times $t-1$ and t, then we can rewrite the unknown probability in equation (1) by exploiting past rather than future information. Proceeding as before, we obtain the bounds

$$\mathrm{LB}_t^{(-1)} = \mathrm{LB}_t + \Pr(Y_{t-1}=1, D_{t-1}=1, D_t=0),$$
$$\mathrm{UB}_t^{(-1)} = \mathrm{UB}_t.$$

Note that, unlike the case when future information is used, here the upper bound is the same as in the worst case, whereas the lower bound is not smaller. 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. In contrast, past positive HIV status is informative, as a person who was HIV infected in the past remains so in the future. The width of the resulting identification interval for π_t is

$$W_t^{(-1)} = \mathrm{UB}_t^{(-1)} - \mathrm{LB}_t^{(-1)} = W_t - \Pr(Y_{t-1}=1, D_{t-1}=1, D_t=0) \leqslant W_t.$$

As shown in appendix A in the supplementary on-line materials, by using three or more waves of a panel we can further narrow the identification region for π_t.

3.3. Instrumental variable and monotone instrumental variable restrictions

The restrictions that were discussed in Section 3.2 may be combined with those implied by additional assumptions. One possibility is IV assumptions. Let Z be a random variable with values in a subset $\mathcal{Z}$ of the real line and observed for both respondents and non-respondents. Then Z is an IV if, after conditioning on a set X of observable covariates with values in $\mathcal{X}$, it helps to predict survey response but not HIV status. Formally, Z is an IV if, for any $(x,z) \in \mathcal{X} \times \mathcal{Z}$,

$$\Pr(D_t=1|X=x, Z=z) \neq \Pr(D_t=1|X=x)$$

but

$$\Pr(Y_t=1|X=x, Z=z) = \Pr(Y_t=1|X=x).$$

If Z is an IV, then we have the following bounds on π_t (Manski (1994) and Manski (2003), section 2):

$$\mathrm{UB}_{\mathrm{IV}}(x) = \inf_z \{\Pr(Y_t=1|X=x, Z=z, D_t=1)\Pr(D_t=1|X=x, Z=z) + \Pr(D_t=0|X=x, Z=z)\},$$
$$\mathrm{LB}_{\mathrm{IV}}(x) = \sup_z \{\Pr(Y_t=1|X=x, Z=z, D_t=1)\Pr(D_t=1|X=z, Z=z)\}.$$

Although finding valid IVs is generally difficult, a convincing case can be made for characteristics of the interview process (interviewer characteristics, interview mode, length of the questionnaire, etc.), because they help to predict non-response (Lepkowski and Couper, 2002; Nicoletti and Peracchi, 2005) but lack predictive power for HIV status. This is in fact the strategy that was followed by Bärnighausen *et al.* (2011) in their implementation of the Heckman selection method. In Section 4.3, we also use variables that are related to the survey process and interviewers' characteristics as IVs.

Another possibility is to impose weaker MIV restrictions. A random variable Z is an MIV if, after conditioning on a set X of observable covariates, it shifts HIV status monotonically. Formally, Z is an MIV if, for any $x \in \mathcal{X}$,

$$\Pr(Y_t=1|X=x, Z=z) \geqslant \Pr(Y_t=1|X=x, Z=z')$$

whenever $z \geqslant z'$ (or $z \leqslant z'$). If Z is an MIV, then we have following bounds on π_t (Manski and Pepper, 2000):

$$\begin{aligned}\mathrm{UB}_{\mathrm{MIV}}(x,z) &= \inf_{z' \geqslant z} \{\Pr(Y_t=1|X=x, Z=z', D_t=1)\Pr(D_t=1|X=x, Z=z') \\ &\quad + \Pr(D_t=0|X=x, Z=z')\}, \\ \mathrm{LB}_{\mathrm{MIV}}(x,z) &= \sup_{z' \leqslant z} \{\Pr(Y_t=1|X=x, Z=z', D_t=1)\Pr(D_t=1|X=z, Z=z')\}.\end{aligned}$$

The details of the IVs and MIVs that are used in this paper are shown below in Section 4.3.

4. Results

We start by presenting complete-case estimates of HIV prevalence in rural Malawi constructed from the MDICP data for 2004, 2006 and 2008 (Section 4.1). These estimates are just the sample proportions of HIV-infected people based on cases with complete information on HIV status. In Section 4.2 we present simple non-parametric estimates of the bounds that were introduced in Section 3, considering both non-respondents and unit respondents. We then focus on unit respondents (Section 4.3) and present our estimated bounds, together with point estimates obtained under alternative assumptions about the missing data process. Following Nicoletti (2010), we also examine the credibility of these point estimates by checking whether they lie inside the bounds. We refer to this procedure as ‘bounds checks’.

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 show no clear trend and are substantially lower than the 2004 MDHS estimate of 10.8% for rural Malawi, possibly because the MDICP sample does not include periurban areas, namely areas that are immediately around urban settlements (Obare *et al.*, 2009).

Since it is of interest to both researchers and policy makers to know how the HIV epidemic affects different demographic groups, we also compute estimates for subgroups defined by region, gender and birth cohort. We distinguish between four cohorts:

(a) cohort A, born 1984–1989 (aged 15–20 years in 2004);
(b) cohort B, born 1975–1983 (aged 21–29 years in 2004);
(c) cohort C, born 1965–1974 (aged 30–39 years in 2004);
(d) cohort D, born before 1965 (aged 40 years or older in 2004).

The full set of results is given in Table S.1 of appendix B in the on-line supporting materials. In particular, the estimated prevalence of HIV is very low for the youngest cohort (cohort A, born 1984–1989) in all three years (less than 4%). Among men, it is always highest (between 4% and 10%) for cohort D (those born before 1965). Among women, it is highest (between 9% and 10%) for cohort B (born 1975–1983) in 2004 and for cohort C (born 1965–1974) in 2006 and 2008. However, because the fraction of the sample with missing HIV status is very high in each wave, uncertainty about the complete-case estimates is also high. This uncertainty will be made evident by the width of the bounds that we present in the next section.

4.2. *Worst-case and dynamic bounds*

The bounds that were introduced in Section 3 are easily estimated non-parametrically by their sample counterparts. To take into account sampling variability, different approaches have been developed in the literature. One approach computes separate confidence intervals for the lower and the upper bounds (Manski *et al.*, 1992). A second approach computes confidence intervals that asymptotically cover the entire identification region with a fixed probability (Horowitz and Manski, 2000). A third approach, which we follow here, computes confidence intervals that asymptotically cover the true parameter with a fixed probability (Imbens and Manski, 2004). The basic idea behind this approach is that, if the identification region has positive width, then the true parameter can be close to one of the region's boundaries. Since we cannot know whether it is close to the lower or the upper bound, we construct one-sided confidence intervals of a given coverage level around both bounds.

Thus, we construct asymptotic α-level confidence intervals for HIV prevalence by using the following formula (formula (6), page 1850, of Imbens and Manski (2004)):

$$\mathrm{CI}_{\alpha}(\pi) = \left[\widehat{\mathrm{LB}} - \mathrm{C}_n \frac{\hat{\sigma}_{\mathrm{LB}}}{\sqrt{n}}, \widehat{\mathrm{UB}} + C_n \frac{\hat{\sigma}_{\mathrm{UB}}}{\sqrt{n}}\right], \tag{3}$$

where the suffix t has been dropped to simplify the notation, $\widehat{\mathrm{LB}}$ and $\widehat{\mathrm{UB}}$ are the sample analogues of LB and UB, $\hat{\sigma}_{\mathrm{LB}}$ and $\hat{\sigma}_{\mathrm{UB}}$ are bootstrap estimates of the asymptotic standard errors of $\widehat{\mathrm{LB}}$ and $\widehat{\mathrm{UB}}$, n is the sample size and C_n satisfies

$$\Phi\left[C_n + \sqrt{n}\frac{\hat{\pi}_{\mathrm{UB}} - \hat{\pi}_{\mathrm{LB}}}{\max\{\hat{\sigma}_{\mathrm{LB}}, \hat{\sigma}_{\mathrm{UB}}\}}\right] - \Phi(-C_n) = \alpha,$$

with Φ the cumulative distribution function of the standard normal distribution. To take the MDICP clustered sampling design into account, $\hat{\sigma}_{\mathrm{LB}}$ and $\hat{\sigma}_{\mathrm{UB}}$ are estimated by using a two-stage bootstrap procedure that randomly selects villages in the first stage and then individuals within the selected villages in the second stage. Villages (in the first stage) and individuals (in the second stage) are sampled with replacement by using the function `sample` in the package `base` of the software R.

Fig. 1 displays graphically our worst-case and dynamic bounds on HIV prevalence in rural Malawi, along with the complete-case estimates. The lower and upper bounds in Fig. 1 are point estimates of the bounds in Sections 3.1 and 3.2, and do not take sampling variability into account. In fact, as can be seen in Table 2, sampling variability adds very little to the width of the identification interval.

Using worst-case bounds, the identification interval is between 4.4% and 33.5% in 2004, between 3.1% and 40.1% in 2006, and between 3% and 46.3% in 2008 (see also Table 2). Note that the width of these intervals increases over time following the pattern of missing data.

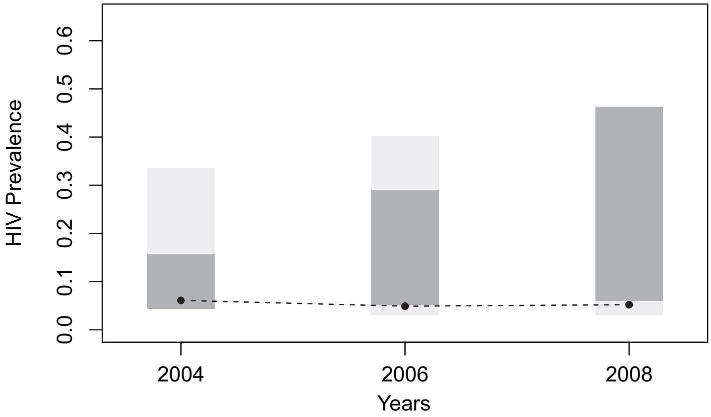


Fig. 1. Estimates of HIV prevalence for the whole sample by survey year (top-ups excluded): □, worst-case bounds; ■, dynamic bounds; ●, MCAR

Table 2. Bounds for the whole sample and by region

Year	*Region*	*Bound type*	*L*†	*U*‡	*W*§	*Lower CI*§§	*Upper CI*§§
2004	All	Worst case	0.044	0.335	0.291	0.043	0.335
		Dynamic	0.044	0.158	0.114	0.043	0.158
	North	Worst case	0.033	0.261	0.228	0.032	0.261
		Dynamic	0.033	0.116	0.083	0.032	0.117
	Centre	Worst case	0.041	0.426	0.385	0.041	0.427
		Dynamic	0.041	0.180	0.139	0.041	0.180
	South	Worst case	0.056	0.314	0.258	0.055	0.315
		Dynamic	0.056	0.174	0.118	0.055	0.175
2006	All	Worst case	0.031	0.401	0.370	0.031	0.401
		Dynamic	0.052	0.290	0.238	0.052	0.290
	North	Worst case	0.027	0.337	0.310	0.026	0.338
		Dynamic	0.038	0.251	0.213	0.038	0.252
	Centre	Worst case	0.027	0.415	0.388	0.026	0.416
		Dynamic	0.044	0.269	0.225	0.043	0.270
	South	Worst case	0.038	0.445	0.407	0.038	0.446
		Dynamic	0.073	0.346	0.273	0.073	0.347
2008	All	Worst case	0.030	0.463	0.433	0.030	0.463
		Dynamic	0.060	0.463	0.403	0.060	0.463
	North	Worst case	0.032	0.445	0.413	0.032	0.446
		Dynamic	0.048	0.445	0.397	0.048	0.446
	Centre	Worst case	0.022	0.412	0.39	0.022	0.413
		Dynamic	0.048	0.412	0.364	0.048	0.413
	South	Worst case	0.035	0.529	0.494	0.035	0.530
		Dynamic	0.082	0.529	0.447	0.082	0.530

†Point estimate of the lower bound.
‡Point estimate of the upper bound.
§§Interval width.
§Lower and upper limits of $\mathrm{CI}_{\alpha}(\pi)$.

Using dynamic bounds, the identification interval is between 4.4% and 15.8% in 2004, between 5.2% and 29% in 2006, and between 6% and 46.3% in 2008. Thus, for 2004 and 2006, we have a sizable reduction in the uncertainty about HIV prevalence compared with the worst-case bounds (amounting to a reduction in their width by about 17.7 percentage points in 2004 and 13.2 percentage points in 2006), although uncertainty remains substantial. For 2008, the reduction is instead very limited (only 3 percentage points). This pattern reflects the number of waves that were available before and after the year at which 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 remains unchanged. In 2006, both past and future information about HIV status helps to reduce the uncertainty, resulting in a decrease in the upper bound and an increase of the lower bound. In 2008, since no subsequent wave of the panel is available, only past information about HIV status helps to reduce the uncertainty, resulting in a small increase in the lower bound with the upper bound unchanged. Note that, although the complete-case estimates are always very close to the lower bound of the identification region, in 2008 they appear to be implausibly low since they fall below the lower limit of the dynamic bounds. This is a warning that estimates based on the MCAR assumption may be downward biased.

Table 2 shows the estimated bounds for rural Malawi as a whole (‘All’) and for the three administrative regions of the country: north, centre and south. According to the complete-case

Table 3. Bounds by gender and birth cohort

Year	*Cohort*	*Bounds type*	*Results for men*					*Results for women*				
			L†	*U*‡	*W*§	*Lower CI*§§	*Upper CI*§§	*L*†	*U*‡	*W*§	*Lower CI*§§	*Upper CI*§§
2004	A	Worst case	0.002	0.225	0.223	0.002	0.228	0.011	0.314	0.304	0.010	0.317
		Dynamic	0.002	0.069	0.067	0.002	0.071	0.011	0.139	0.129	0.010	0.141
	B	Worst case	0.021	0.275	0.254	0.021	0.278	0.062	0.380	0.318	0.062	0.382
		Dynamic	0.021	0.107	0.086	0.021	0.109	0.062	0.184	0.121	0.062	0.185
	C	Worst case	0.038	0.405	0.367	0.037	0.406	0.060	0.328	0.268	0.059	0.331
		Dynamic	0.038	0.178	0.141	0.037	0.178	0.060	0.166	0.106	0.059	0.168
	D	Worst case	0.067	0.36	0.294	0.066	0.362	0.061	0.295	0.234	0.060	0.296
		Dynamic	0.067	0.184	0.117	0.066	0.185	0.061	0.125	0.064	0.060	0.126
2006	A	Worst case	0.000	0.408	0.408	0.000	0.410	0.011	0.484	0.474	0.010	0.487
		Dynamic	0.002	0.275	0.273	0.002	0.278	0.017	0.342	0.326	0.016	0.345
	B	Worst case	0.008	0.435	0.426	0.008	0.438	0.043	0.424	0.381	0.042	0.426
		Dynamic	0.022	0.323	0.301	0.021	0.326	0.079	0.297	0.218	0.078	0.299
	C	Worst case	0.023	0.356	0.332	0.022	0.359	0.078	0.339	0.261	0.077	0.341
		Dynamic	0.042	0.265	0.223	0.040	0.268	0.097	0.244	0.148	0.095	0.246
	D	Worst case	0.038	0.358	0.320	0.037	0.360	0.029	0.321	0.293	0.028	0.323
		Dynamic	0.063	0.273	0.210	0.063	0.275	0.059	0.202	0.143	0.058	0.204
2008	A	Worst case	0.005	0.535	0.530	0.005	0.538	0.019	0.569	0.550	0.018	0.572
		Dynamic	0.008	0.535	0.528	0.007	0.538	0.030	0.569	0.539	0.029	0.572
	B	Worst case	0.020	0.524	0.504	0.019	0.527	0.043	0.420	0.377	0.043	0.422
		Dynamic	0.037	0.524	0.487	0.035	0.527	0.091	0.420	0.330	0.089	0.422
	C	Worst case	0.024	0.429	0.405	0.023	0.432	0.066	0.392	0.326	0.065	0.395
		Dynamic	0.044	0.429	0.385	0.043	0.432	0.103	0.392	0.289	0.101	0.394
	D	Worst case	0.027	0.439	0.412	0.026	0.441	0.026	0.346	0.320	0.025	0.348
		Dynamic	0.066	0.439	0.373	0.065	0.441	0.055	0.346	0.290	0.054	0.348

†Point estimate of the lower bound.
‡Point estimate of the upper bound.
§Interval width.
§§Lower and upper limits of $\mathrm{CI}_\alpha(\pi)$.

estimates from the MDHS, southern Malawi is the region where HIV prevalence is highest, followed by the centre and the north. Although the dynamic bounds are much narrower than the worst-case bounds, and the lower bound for the south is always considerably higher than for the other regions, the bounds overlap and do not allow a ranking of the regions in terms of HIV prevalence. Table 3 also reports the confidence intervals for HIV prevalence. The lower and upper limits of these confidence intervals are always very close to the point estimates of the lower and upper bound of the identification region, suggesting that sampling uncertainty can be neglected. Because estimates of the identification regions overlap, we make no attempt at drawing inference about differences in HIV prevalence over time or across sociodemographic groups.

Table 3 shows that the dynamic bounds are much narrower than the worst-case bounds also for subgroups that are characterized by gender and birth cohort, especially in 2004 and 2006. For example, for men of cohort C (1965–1974) and for women of cohort B (1975–1983) the width of the identification region in 2004 is narrowed by about 20 percentage points when estimated by using the dynamic bounds. Nonetheless, the identification regions remain too wide to allow us to establish a rank by gender. Table S.1 of appendix B in the on-line supplementary material reports the bounds, indicating with a star the point estimates which are implausible because they are outside the identification region. We note that almost all the MCAR point estimates in 2006 and 2008 are implausibly low because they are below the lower limit of the dynamic bounds.

4.3. Unit respondents only

Since most variables are missing for the majority of unit non-respondents, it is difficult to think of any variable that could be used as an IV or an MIV, so we restrict the analysis to unit respondents.

4.3.1. *Imposing instrumental variable and monotone instrumental variable restrictions*

The IVs that we consider are all related to the interview process. They include gender differences between the interviewer and the interviewee (same or different gender), the interviewer's age (below age 23 years or aged over 23 years) and experience (none or some), and the month of the first interview attempt (May–June or July–August). As suggested by Nicoletti and Peracchi (2005), variables that are related to the interview process can convincingly be used as instruments because they are unlikely to have a direct effect on the outcome of interest (HIV status in our case) but are important predictors of non-response. For example, having an experienced interviewer or an interviewer of the same gender as the interviewee tends to lower refusal rates. Further, the timing of the first interview attempt affects the probability of finding the interviewees at home, especially if these are men who must follow the cycle of economic activity in the countryside.

As an MIV, we consider the number of sexual partners that each respondent had up to the year of the interview. This is a valid MIV under the plausible assumption that the probability of being HIV infected does not fall as the number of sexual partners increases. Table 4 presents a summary of the IVs and MIVs that we consider.

Whereas all four IVs are available in 2004 and 2006, only the interview month is available in 2008. For this reason, and because the interview month is the IV that usually produces the

Table 4. Summary statistics for our IVs and monotone MIVs for unit respondents

	Results for 2004		*Results for 2006*		*Results for 2008*†	
	Frequency	*%*	*Frequency*	*%*	*Frequency*	*%*
IV						
Interviewer's gender						
Same	1350	49.0	1405	62.8		
Different	1408	51.0	831	37.2		
Interviewer's experience						
None	1214	44.0	1087	48.6		
Some	1544	56.0	1149	51.4		
Interviewer's age						
Below age 23 years	1112	40.3	1111	49.7		
Aged 23 years or older	1646	59.7	1125	50.3		
Interview month						
May–June	1804	65.4	1255	56.1	1250	44.3
July–August	954	34.6	981	43.9	1575	55.7
MIV						
Number of sexual partners						
0–1	1142	41.4	773	34.6	944	33.4
2	671	24.3	595	26.6	659	23.3
3	365	13.2	358	16.0	448	15.9
$\geqslant 4$	580	21.0	510	22.8	774	27.4

†In 2008, no information was collected on interviewer's gender, experience or age.

narrowest bounds, we present results based only on this variable. The full set of results by year, gender and cohort, and for all instruments, can be found in Tables S.2–S.7 of appendix B in the on-line supplementary material.

In the remainder of this section, we consider as benchmark the dynamic bounds for HIV prevalence estimated without imposing IV or MIV restrictions. In 2004, the identification region is the interval between 4.9% and 12.4% in the benchmark case, the interval between 4.9% and 10% when using the interview month as an IV and the interval between 5.1% and 11.6% when using our MIV. In 2006, the identification region is the interval between 4.3% and 15.1% in the benchmark case, the interval between 4.5% and 13.1% when using the interview month as an IV and the interval between 4.3% and 15.1% when using our MIV. In 2008, the identification region is the interval between 5.1% and 28.9% in the benchmark case, the interval between 5.8% and 24.3% when using the interview month as an IV and the interval between 5.3% and 28.7% 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 about 2 percentage points in 2004 and 2006, and by 5.3 percentage points in 2008. In contrast, using the number of sexual partners as an MIV is of little help in narrowing the identification region.

Figs 2 and 3 show the dynamic bounds on HIV prevalence by survey year, separately by gender and birth cohort, along with the point estimates based on the MCAR (complete-case estimates), MAR and MNAR assumptions. Reported results use as IV the month of the first interview attempt, as this variable is available for each year and is generally the most effective in reducing the width of the identification region. For example, in 2004, using the month of interview as IV usually reduces the bounds' widths by 2–3 percentage points compared with the benchmark bounds. Fig. 3 shows that the MIV restriction now seems to be effective in reducing the width of the identification interval, although this varies by gender and cohort (see Tables S.2–S.7 in the on-line supporting information). Using an IV or MIV restriction, we can obtain quite narrow bounds for some demographic groups. For example, the MIV bound for males in 2004 of cohort B (1975–1983) is (0.032; 0.050) and the IV bound in 2004 for females of cohort C (1965–1974) is (0.074; 0.092).

4.3.2. Estimates under the missingness completely at random and missingness at random assumption

As before, the MCAR estimate of HIV prevalence is the sample proportion of infected individuals, ignoring those with missing HIV status. We estimate HIV prevalence under the MAR assumption by using the propensity score weighting method. This corresponds to weighted maximum likelihood estimation of a probit model, where the weights are equal to the inverse of the probability of observing HIV status given a set of covariates which includes age, gender, ethnic group, region of residence, marital status and level of education of the respondent. Detailed estimation results are provided in Table S.8 of the on-line supporting material. Note that the point estimates for the MCAR and MAR assumptions in Figs 2 and 3 are very similar and are usually very close to the lower bounds. Our bounds checks show that these estimates are often implausibly low because they fall below the lower bound. Out of the 24 cases considered (eight demographic groups for 3 years), this happens eight times for the MCAR estimates and seven times for the MAR estimates.

4.3.3. Heckman selection model

We also estimate HIV prevalence by using a Heckman-type selection model similar to that used by Bärnighausen *et al.* (2011). Although this approach has the advantage of providing a point

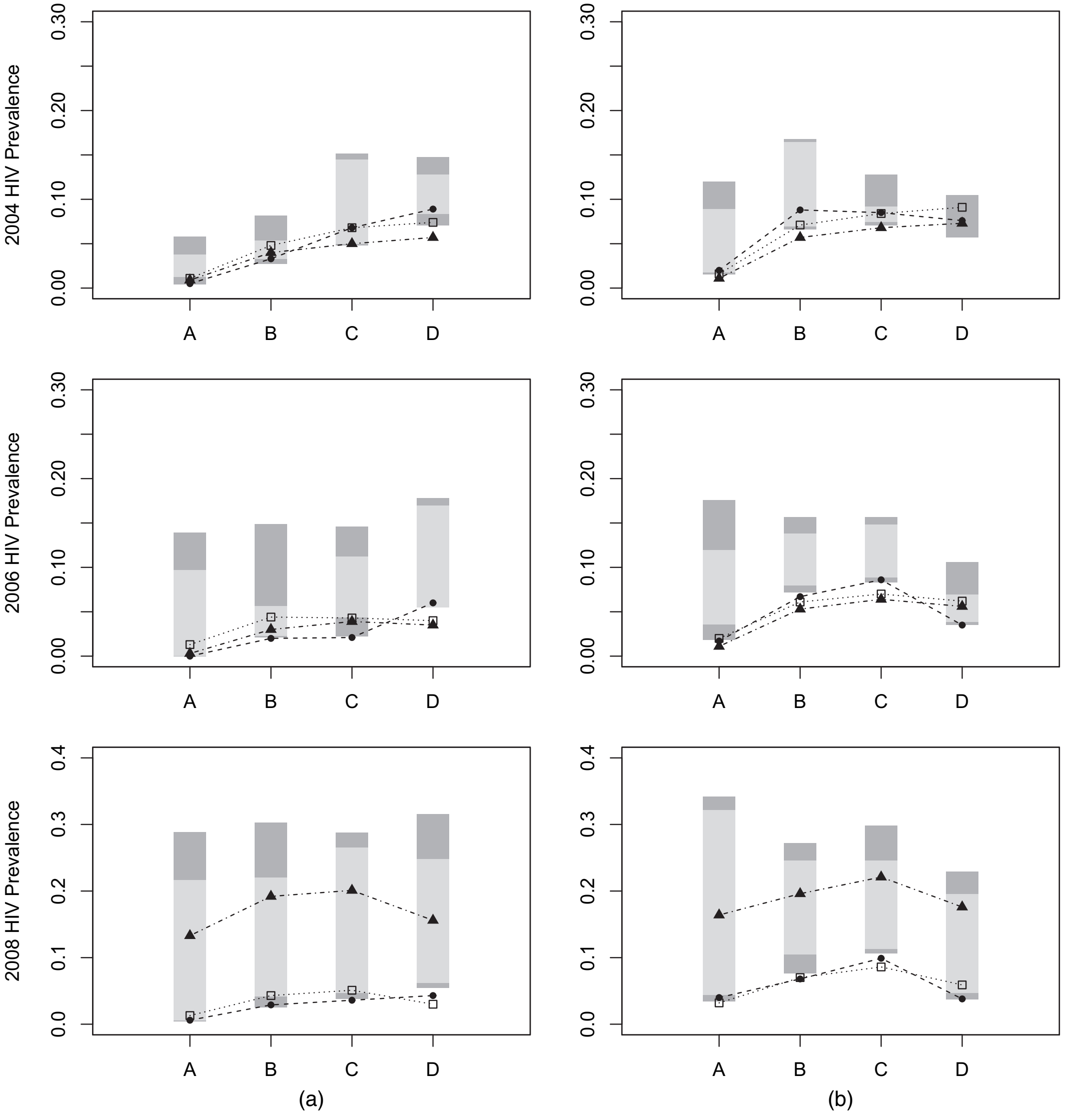


Fig. 2. HIV prevalence for unit respondents by year and cohort for (a) men and (b) women: estimates based on MCAR (●), MAR (□) and Heckman (▲) assumptions and dynamic bounds in the benchmark case (■) and dynamic bounds with IV restriction (▨) (those in cohort A were born in 1984–1989, in cohort B in 1975–1983, in cohort C in 1965–1974 and in cohort D before 1965; the bounds estimated by using the interview month as IV have a negative width in 2004 for women of cohort D and are not reported)

estimate instead of an interval of values, it requires parametric assumptions on the distribution of the unobservables in the model and an exclusion restriction, namely a variable that helps to explain the missing data process but not the outcome of interest. Our exclusion restriction is the month of the first interview attempt, which we used as an IV in the previous section. The covariates in the model are also the same as those used for the estimates under MAR. Detailed estimation results are presented in Table S.8 of the on-line supporting material.

Like the point estimates that were obtained under MCAR and MAR, the Heckman point estimates shown in Figs 2 and 3 are often very close to the lower bound but sometimes (eight cases

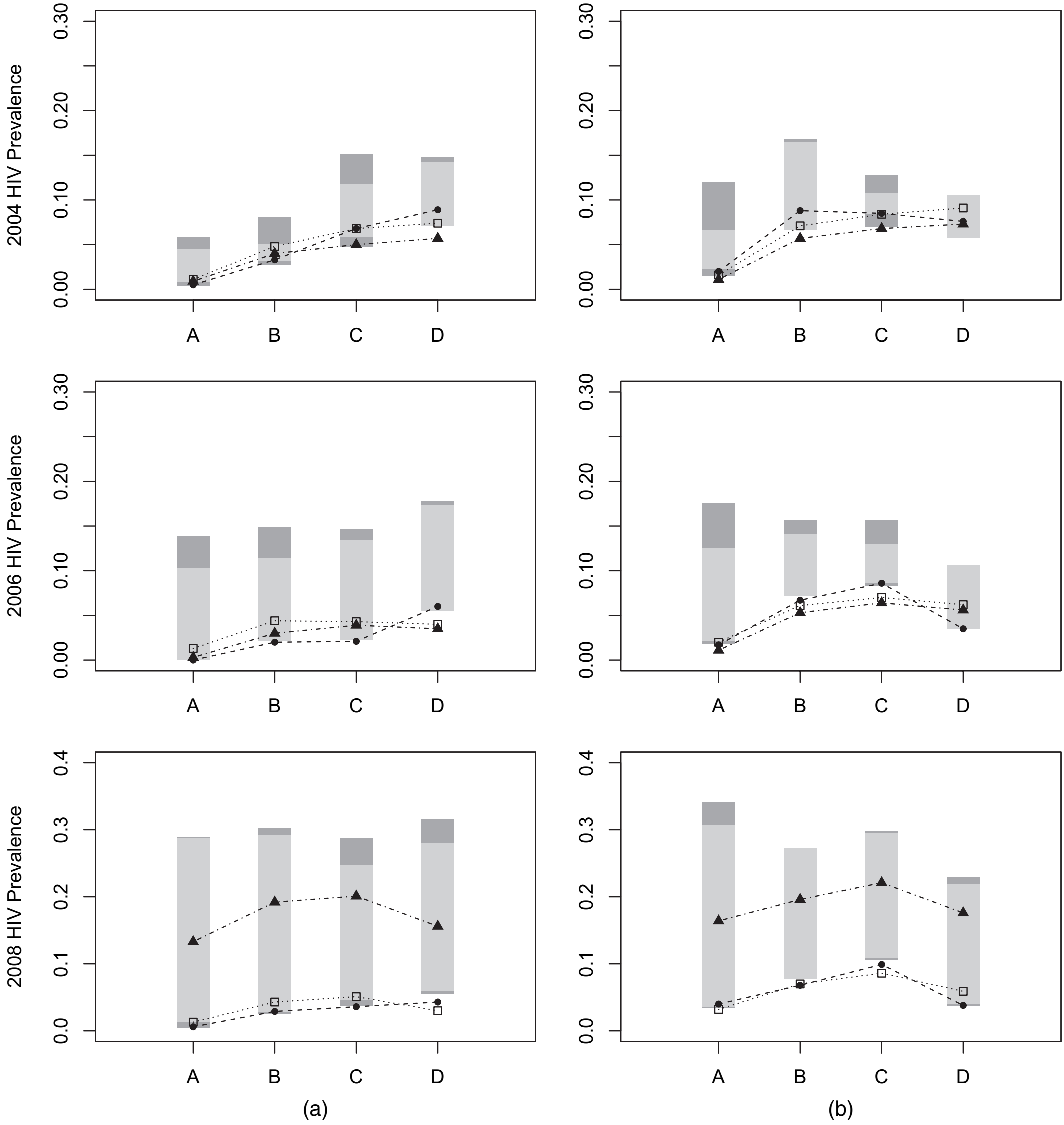


Fig. 3. HIV prevalence for unit respondents by year and cohort for (a) men and (b) women: estimates based on MCAR (●), MAR (□) and Heckman (▲) assumptions and dynamic bounds in the benchmark case (■) and dynamic bounds with MIV restriction (□) (those in cohort A were born in 1984–1989, in cohort B in 1975–1983, in cohort C in 1965–1974 and in cohort D before 1965; the bounds estimated by using the interview month as MIV have a negative width in 2004 for women of cohort D and are not reported)

out of 24) they fall below the lower limit of the identification region. This is perhaps not surprising, as these estimates crucially depend on the validity of the model assumptions.

5. Discussion

Credible estimates of HIV prevalence are critical for policy makers. Although estimates that are obtained from population-based surveys are often considered as the gold standard, they are affected by non-ignorable missing data problems, which in turn translate into substantial uncertainty about HIV prevalence in the population.

Panel data are typically used to estimate HIV incidence rates, but they can also be used to estimate HIV prevalence at different points in time for the same population. Our paper uses a bounding approach to assess what can be learned from this type of data. Our main contribution is to show how worst-case bounds, based only on sample information and often distressingly wide, can be narrowed by exploiting the longitudinal nature of the data and the absorbing nature of HIV infection.

The identifying power of panel data comes from the fact that the HIV status of current non-respondents may be observed in other waves. 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 non-respondents in the current wave have been found to be HIV infected in past waves. Thus, 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.

Ignoring the missing data problem and using only the complete cases give point estimates of HIV prevalence that are very close to our lower bounds. These estimate may be too optimistic because the data alone do not rule out the possibility that HIV prevalence is much higher. We also find that the estimates under the MCAR and Heckman assumptions fall below the lower bound of the identification region in eight out of the 24 cases considered, whereas for the MAR estimates this happens six times. We conclude from these bounds checks that, in our data, the point estimators considered do not always give plausible estimates of HIV prevalence. Thus, bounds checks are a useful reminder of the role that is played by strong and often untestable assumptions in supplementing the relatively weak information provided by the data. As argued by Manski (2011), acknowledging ambiguity reduces the danger of feigning certitude.

Our approach is easy to implement and does not require assumptions about the nature of the missing data mechanism. 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. Estimated bounds provide a range of plausible values for the parameter of interest that does not rely on strong assumptions. The bounds can be combined with point estimators to check the plausibility of the assumptions on which the latter are based.

We conclude with three remarks. First, it is important to keep non-response rates low, and to consider unit and item non-response separately when analysing the data. Second, including in the data information on the interview process is important because it can be used as a source of IVs. Third, if the data are MNAR, then an effort should be made at interviewing a subset of the non-respondents. In longitudinal surveys, in particular, it pays to collect data on people who, for different reasons, did not participate in previous waves.

Acknowledgements

We thank the Associate Editor and two referees for valuable criticism and suggestions. We also thank Philip Anglewicz and Hans-Peter Kohler for providing the data from the MDICP and for support, and participants at the Italian Congress of Econometrics and Empirical Economics 2011 (Pisa, Italy, January 2011), the Understanding Society–British Household Panel Survey conference 2011 (Colchester, UK, June 2011), the International Statistical Institute conference 2011 (Dublin, Ireland, August 2011) and the Population Association of America Annual Meeting 2012 (San Francisco, USA, May 2012) for their comments. The MDICP has been funded by Eunice Kennedy Shriver National Institute of Child Health and Human Development grants NICHD R01HD053781 (‘Consequences of high morbidity and mortality in a low-income country’), NICHD R01HD044228 (‘AIDS/HIV risk, marriage and sexual relations in Malawi’) and NICHD R01HD/MH41713 (‘Gender, conversational networks and dealing with STDs’).

References

Anglewicz, P. (2012) Migration, marital change, and HIV infection in Malawi. *Demography*, **49**, 239–265.

Anglewicz, P., Adams, J., Obare, F., Kohler, H. P. and Watkins, S. (2009) The Malawi and diffusion ideational change project 2004-2006: data collection, data quality, and analysis of attrition. *Demogr. Res.*, **20**, 503–540.

Antoniou, T., Zagorski, B., Loutfy, M. R., Strike, C. and Glazier, R. H. (2011) Validation of case-finding algorithms derived from administrative data for identifying adults living with Human Immunodeficiency Virus infection. *PLOS ONE*, **6**, no. 6.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S. and Canning, D. (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**, 27–35.

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

Brookmeyer, R. (2010) Measuring the HIV/AIDS epidemic: approaches and challenges. *Epidem. Rev.*, **32**, 26–37.

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

Garcia-Calleja, J., Gouws, E. and Ghys, P. (2006) National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates. *Sexlly Transmttd Infectns*, **82**, suppl III, iii64–iii70.

Gouws, E., Mishra, V. and Fowler, T. B. (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. *Sexlly Transmttd Infectns*, **84**, suppl. I, i17–i23.

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

Horowitz, J. L. and Manski, C. F. (1998) Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputation. *J. Econmetr.*, **84**, 37–58.

Horowitz, J. and Manski, C. F. (2000) Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Statist. Ass.*, **95**, 77–84.

Imbens, G. W. and Manski, C. F. (2004) Confidence intervals for partially identified parameters. *Econometrica*, **72**, 1845–1857.

Kohler, H. P., Taulo, F., Masanjala, W., Watkins, S. C. and Behrman, J. R. (2009) Summary of data collection 1998-2008. *Malawi Longitudinal Study of Families and Health Newslett.*, no 1. (Available from `http://www.ssc.upenn.edu/~hpkohler/malawi/2009-02-mlsfh-datacollection.pdf`.)

Lachaud, J. P. (2007) Hiv prevalence and poverty in Africa: micro- and macro-econometric evidences applied to Burkina Faso. *J. Hlth Econ.*, **26**, 483–504.

Lepkowski, J. M. and Couper, M. P. (2002) Non-response in longitudinal household surveys. In *Survey Nonresponse* (eds R. M. Groves, D. Dillman, J. Eltinge and R. Little), pp. 259–272. New York: Wiley.

Manski, C. F. (1989) Anatomy of the selection problem. *J. Hum. Resour.*, **24**, 343–360.

Manski, C. F. (1994) The selection problem. In *Advances in Econometrics: 6th Wrld Congr.* (ed. C. Sims), pp. 143–170. New York: Cambridge University Press.

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

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

Manski, C. F. (2011) Policy analysis with incredible certitude. *Econ. J.*, **121**, F261–F289.

Manski, C. F. and Pepper, J. (2000) Monotone instrumental variables with an application to the returns to schooling. *Econometrica*, **68**, 997–1010.

Manski, C. F., Sandefur, G., McLanahan, S. and Powers, D. (1992) An alternative estimate of the effect of family structure during adolescence on high school graduation. *J. Am. Statist. Ass.*, **87**, 25–37.

Marston, M., Harriss, K. and Slaymaker, E. (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. *Sexlly Transmttd Infectns*, **84**, i71–i77.

Martin-Herz, S., Shetty, A., Bassett, M., Ley, C., Mhazo, M., Moyo, S., Herz, A. and Katzenstein, D. (2006) Perceived risks and benefits of HIV testing, and predictors of acceptance of HIV counseling and testing among pregnant women in Zimbabwe. *Int. J. Sexlly Transmttd Dis. AIDS*, **17**, 835–841.

Mishra, V., Barrere, B., Hong, R. and Khan, S. (2008) Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexlly Transmttd Infectns*, **84**, suppl. I, i63–i70.

Molenberghs, G., Beunckens, C., Sotto, C. and Kenward, M. G. (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Statist. Soc.* B, **70**, 371–388.

Montana, L., Mishra, V. and Hong, R. (2008) Measuring the HIV/AIDS epidemic: approaches and challenges. *Sexlly Transmttd Infectns*, **84**, i78–i84.

National Statistical Office Malawi and Opinion Research Corporation Macro (2005) *Malawi Demographic and Health Survey 2004*. Calverton: National Statistical Office and Opinion Research Corporation Macro.

Nicoletti, C. (2010) Poverty analysis with missing data: alternative estimators compared. *Empir. Econ.*, **38**, 1–22.

Nicoletti, C. and Peracchi, F. (2005) Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *J. R. Statist. Soc.* A, **168**, 763–781.

Obare, F. (2010) Nonresponse in repeat population-based voluntary counseling and testing for HIV in rural Malawi. *Demography*, **47**, 651–665.

Obare, F., Fleming, P., Anglewicz, R., Thornton, R., Martinson, F., Kapatuka, A., Poulin, M., Watkins, S. and Kohler, H. (2009) Acceptance of repeat population-based voluntary counselling and testing for HIV in rural Malawi. *Sexlly Transmttd Infectns*, **85**, 139–144.

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

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

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

Thornton, R. (2008) The demand for, and impact of, learning HIV status. *Am. Econ. Rev.*, **98**, 1829–1863.

UNAIDS–World Health Organization (2003) *Guidelines for Conducting HIV Sentinel Serosurveys among Pregnant Women and Other Groups*. UNAIDS, Geneva.

United Nations General Assembly Special Session (2010) Malawi HIV and AIDS Monitoring and Evaluation Report: 2008-2009. *Technical Report.* United Nations, New York.

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

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Web-based supporting materials for “Using panel data to partially identify HIV prevalence when HIV status is not missing to random”’.