# Using panel data to partially identify HIV prevalence when HIV status is missing not 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' and EIEF, Rome, Italy*

**Abstract.** Population-based surveys are considered the "gold standard" to estimate HIV prevalence but typically suffer of 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 imposed by methods that point identify HIV prevalence.

*Keywords:* HIV prevalence; Malawi Diffusion and Ideational Change Project data; Non-ignorable nonresponse; Panel data; Partial identification.

## 1. Introduction

The prevalence of 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 HIV prevalence are 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).

†*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

ANC data contain several sources of bias. First, they exclude men and are only representative of sexually active women who are pregnant and go to the clinics. Second, they may provide biased estimates even for the sub-population of pregnant women because of 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, several population-based surveys began including modules that collect various biomarkers, such as blood samples or saliva swabs useful to test for HIV. These 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 obtained from these surveys are, in general, considerably lower than those obtained from 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), 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 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) or, if there are covariates, on the weaker assumption that non-response is independent of the outcome of interest given the covariates (missing at random, MAR) (Wooldridge, 2007; Little and Rubin, 1987; Rubin, 1989). Under the MAR assumption, imputation or weighting techniques are frequently used to produce consistent estimates of HIV prevalence. However, if the missing data mechanism depends on true HIV status, then the 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) analyses 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 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 pioneered by Heckman (1979) to estimate HIV prevalence under the MNAR assumption by letting survey nonresponse depend on 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 avoid identification via functional form assumptions, it also requires exclusion restrictions, namely variables that help explain the missing data process but not the outcome of interest. For example, Bärnighausen et al. (2011) use data from the Zambia Demographic and Health Survey, where 28% of men did not participate in HIV testing, and find that the estimate of male HIV prevalence 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, which help predict survey participation but have arguably no direct effect on HIV status.

Notice that, while 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, we ask what can be learned about HIV prevalence without the need of strong untestable assumptions, such as those commonly made in sample selection models. Following Horowitz and Manski (1998) and Manski (1995, 2003), we 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. This approach explicitly recognizes ambiguity by identifying the set of values (the identification region) to which the parameter of interest (HIV prevalence in our case) must necessarily belong given the available data and the maintained assumptions. If the maintained assumptions are sufficiently strong, the identification region collapses to a single point and the parameter of interest is point identified.

We first use the empirical evidence alone to partially identify HIV prevalence. We then exploit the availability of panel data and the absorbing nature of HIV infection to narrow the width of this initial identification 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 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 re-

gion (Nicoletti, 2010). In particular, we consider point estimates obtained using the complete-case approach, propensity score weighting, and a Heckman-type estimator.

Our data are from the Malawi Diffusion and Ideational Change Project (MDICP), a longitudinal survey of the population of rural Malawi. Starting from 2004, a biomarker module has been added to the main survey allowing estimation of HIV prevalence. Malawi is one of the African countries most affected by the HIV epidemic and AIDS is the leading cause of death among adults (UNGASS, 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. 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). 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, its biomarker module is available for 2004, 2006 and 2008, and not just for 2004 as the case for the MDHS.

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 width of the initial identification region based on empirical evidence alone. It also discusses how to use plausible IV and MIV restrictions to further narrow the identification region. Section 4 presents our empirical results, broken down 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 every two years since 1998 in rural Malawi. 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 data can be freely downloaded from the following website: `http://www.malawi.pop.upenn.edu`, and include the outcomes of HIV tests for the years 2004, 2006 and 2008.

### *2.1. The MDICP survey*

The 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 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).

Here we only provide a brief description of the survey design and refer to the MDICP website for more detail.

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. Then, in the second stage, a sample of eligible women was randomly selected from the list of people normally resident in those villages.

In total, 1,541 ever-married women of childbearing age and 1,198 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 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. 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 resided in the same village as the respondent. This new sample of about 800 parents was based on family listings obtained from 2006 respondents. The overall survey response rate (the percentage of targeted people who were successfully interviewed) was 78.6% in 1998, 72.1% in 2001, 67.0% in 2004, 67.9% in 2006 and 67.4% in 2008.

The survey collects extensive information on household structure, health, risk assessments, sexual relations, marriage and partnership histories, intergenerational and inter-familiar 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 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. According to the available documentation, this test has a 100% probability of detecting true positives and a very small probability of false positives. Because measurement error in the two types of tests (oral swabs and blood test) appears to be limited, and due only

to the accuracy of the measuring instruments, we ignore the problem.

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. We consider the 2004 as our baseline, not only because biomarkers are available only from this year, but also because the basic demographic characteristics of the 2004 MDICP are very similar to those of the 2004 MDHS (National Statistical Office (NSO) Malawi and ORC 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 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. Two cases are possible. One is when the main and the VCT surveys are both missing due to failure to establish a contact or refusal to cooperate (unit nonresponse). The other is when HIV status is not available for a responding unit (item nonresponse).

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 but not in 2006 and 2008, about 11% are unit respondents in 2004 and 2006 but not in 2008, about 8% are unit respondents in 2004 and 2008 but not in 2006, while the remaining 14% include other patterns of unit nonresponse.

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

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. Notice, however, that in 2004 the MDICP has a lower refusal rate than the MDHS in rural areas (6.3% against 21.7%). Thornton (2008) argues that this may be due to the testing method (oral swabs) and the fact that the MDICP does not require respondents to learn their results at the time of testing. Low refusal rates (less than 5%) are also found in the 2006 and 2008 MDCIP. In very few cases the results of the HIV test are indeterminate or have been lost. Other reasons for

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

| | 2004 | | 2006 | | 2008 | |
|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq. | % |
| **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 nonrespondents** | | | | | | |
| Test refused | 256 | 6.3 | 200 | 4.9 | 172 | 4.2 |
| Indeterminate | 14 | .3 | 6 | .1 | 1 | .0 |
| Results lost | 24 | .6 | 0 | .0 | 0 | .0 |
| Other† | 319 | 7.9 | 313 | 7.7 | 569 | 14.0 |
| **Unit nonrespondents** | | | | | | |
| Refused | 27 | .7 | 11 | .3 | 58 | 1.4 |
| Moved | 184 | 4.5 | 479 | 11.8 | 470 | 11.6 |
| Temporarily absent | 36 | .9 | 41 | 1.0 | 76 | 1.9 |
| Hospitalized | 6 | .1 | 5 | .1 | 1 | .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, for example because they were temporarily absent during the biomarker collection.
‡People who did not complete the questionnaire for unknown reasons or because too old or too sick.
§New entrants between 2006 and 2008 are excluded.

item nonresponse, lumped into the category 'other', consist 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.

Distinguishing between the different sources of missing data is important. Ignoring missing data due to migration or test refusal may bias HIV prevalence estimates 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 living individuals who can be susceptible to HIV or infected. A susceptible individual is a member of the population who is at risk of becoming infected by the disease. HIV status of a randomly selected individual at time $t$ is represented

by the binary random variable $Y_t$, equal to one if the individual is infected and equal to zero 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 $\pi_t$ when HIV status is missing for a fraction of individuals in the population. As argued in the previous section, measurement error is negligible in our data 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 $\pi_t$ caused by the 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 single point in time, as in a cross-section or when the longitudinal dimension of a panel 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 $\Pr(Y_t = 1|D_t = 0)$, the prevalence of HIV among people with missing HIV status. 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 $\pi_t$

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

These bounds are often referred to as worst-case bounds because they only use the available data and are therefore wider than those obtained by imposing additional restrictions.

The identification region for $\pi_t$ consists of all the points in the interval between $LB_t$ and $UB_t$. The width $W_t = UB_t - LB_t$ of this interval 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 $\pi_t$. When nonresponse rates are high, as in our case, uncertainty is large. An important issue, therefore, is whether there exists additional information about the HIV process which may be exploited 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 nonrespondent in one wave may be observed in earlier or later waves. We will refer to the resulting bounds as 'dynamic' because they are based on longitudinal data and exploit restrictions implied by the dynamics of HIV epidemic. To keep things simple and in line with the data that we use, we only present the results for the case of short panels with two or three waves. Appendix A of the on-line supplementary materials 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 $\pi_t$, 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 $\Pr(Y_t = 1|Y_{t+1} = 0, D_{t+1} = 1, D_t = 0) = 0$. This is because someone who is HIV infected cannot recover (become uninfected), which is why infection is an "absorbing" state. 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 $\pi_t$ 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. The width of the resulting identification interval for $\pi_t$ is

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

where $W_t$ is the width of the worst-case bounds. Because $\Pr(Y_{t+1} = 0, D_{t+1} = 1, D_t = 0)$ is non-negative and cannot exceed $\Pr(D_t = 0)$, it follows that $0 \leq W_t^{(+1)} \leq W_t$.

Notice 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 exactly infer HIV status at time $t$, 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 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

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

and we used the fact that $\Pr(Y_t = 1|D_t = 0, D_{t-1} = 1, Y_{t-1} = 1) = 1$ 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 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. On the other hand, 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 $\pi_t$ is

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

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

Using three waves of a panel, we can further narrow the identification region for $\pi_t$. Suppose that, in addition to wave $t$, we use the two adjacent waves at times $t-1$ and $t+1$. Then it follows immediately from our previous results that

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

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 $\pi_t$ 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, it decreases because of a combination of the two effects. Increasing the number of available waves further decreases the uncertainty due to missing data, as shown in Appendix A in the supplementary on-line materials.

### *3.3. IV and MIV restrictions*

To further narrow the identification region for $\pi_t$, the restrictions discussed in Section 3.2 may be combined with those implied by additional assumptions.

One possibility are instrumental variable (IV) assumptions (Manski, 1994, 2003). A random variable $Z$ with values in a subset $\mathcal{Z}$ of the real line is an IV if, after conditioning on a set $X$ of observable covariates with values in $\mathcal{X}$, it helps predict survey response but not HIV status. Thus, $Z$ 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 $\pi_t$ (Manski, 1994, 2003):

$$\begin{aligned} UB_{IV}(x) = \inf_z \{ &\Pr(Y_t = 1|X = x, Z = z, D_t = 1) \Pr(D_t = 1|X = x, Z = z) + \\ &+ \Pr(D_t = 0|X = x, Z = z)\}, \\ LB_{IV}(x) = \sup_z \{ &\Pr(Y_t = 1|X = x, Z = z, D_t = 1) \Pr(D_t = 1|X = z, Z = z)\}. \end{aligned}$$

Although finding valid instrumental variables 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 predict nonresponse (Lepkowski and Couper, 2002; Nicoletti and Peracchi, 2005) but lack predictive power for HIV status. This is in fact the strategy followed by Bärnighausen et al. (2011) in their implementation of Heckman selection method.

Since IV restrictions may be controversial, another possibility is to impose weaker monotone instrumental variable (MIV) restrictions (Manski and Pepper, 2000). A random variable $Z$ is a MIV if, after conditioning on a set $X$ of observable covariates, it shifts HIV status monotonically. Formally, $Z$ is a MIV if, for any $x \in \mathcal{X}$,

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

whenever $z \geq z'$ (or $z \leq z'$). If $Z$ is a MIV, then we have following bounds on $\pi_t$ (Manski and Pepper, 2000):

$$\begin{aligned} UB_{MIV}(x, z) = \inf_{z' \geq z} \{ &\Pr(Y_t = 1|X = x, Z = z', D_t = 1) \Pr(D_t = 1|X = x, Z = z') + \\ &+ \Pr(D_t = 0|X = x, Z = z')\}, \\ LB_{MIV}(x, z) = \sup_{z' \leq z} \{ &\Pr(Y_t = 1|X = x, Z = z', D_t = 1) \Pr(D_t = 1|X = z, Z = z')\}. \end{aligned}$$

## 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 considering both non-respondents and unit respondents (Section 4.1). Our complete-case estimates are the sample proportions of HIV infected people based on cases with complete information on HIV status, ignoring covariates. We then focus on unit respondents (Section 4.2) and present estimated bounds, together with point estimates obtained under alternative assumptions about the missing data process. Following Nicoletti (2010), we examine the credibility of these point estimates by checking whether they lie inside the bounds. We refer to this procedure as "bounds checks".

Since it is of interest to both researchers and policy-makers to know how the HIV epidemic affects different demographic groups, we present estimates for the

whole population of rural Malawi 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. Nonrespondents and unit respondents*

#### *4.1.1. 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 peri-urban areas (Obare et al., 2009).

The full set of results from complete-case estimation are given in Table S.1 of Appendix B in the on-line supporting materials. In particular, estimated HIV prevalence 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 (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 we present in next section.

#### *4.1.2. Worst-case and dynamic bounds*

The bounds introduced in Section 3 are easily estimated non-parametrically by their sample counterparts. To take into account sampling variability, different confidence intervals have been developed in the literature. One approach computes separate confidence intervals for the lower and the upper bounds (Manski et al., 1992). Another approach computes confidence intervals that asymptotically cover the entire identification region with a fixed probability (Horowitz and Manski, 2000). A third approach computes confidence intervals that asymptotically cover the true value of the parameter with a fixed probability (Imbens and Manski, 2004).

Here we follow the third approach and construct the following asymptotic $\alpha$-level confidence intervals for HIV prevalence

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

where the suffix $t$ has been dropped to simplify notation, $\widehat{LB}$ and $\widehat{UB}$ are the sample analogues of $LB$ and $UP$, $\hat{\sigma}_{LB}$ and $\hat{\sigma}_{UB}$ are bootstrap estimates of the

asymptotic standard errors of $\widehat{LB}$ and $\widehat{UB}$, $n$ is the sample size, and $C_n$ satisfies

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

with $\Phi$ the cumulative distribution function of the standard normal distribution. To take the MDICP sampling design into account, $\hat{\sigma}_{LB}$ and $\hat{\sigma}_{UB}$ are estimated 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.

Figure 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 the figure 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). Notice that the width of these intervals increases over time following the pattern of missing data.

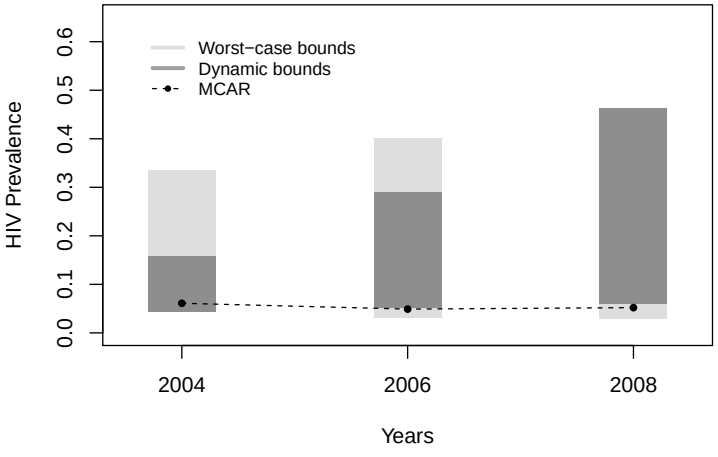


**Figure 1.** Estimates of HIV prevalence for the whole sample by survey year.

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 sizeable reductions of the uncertainty about

HIV prevalence compared to the worst-case bounds (amounting to a reduction of 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 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 remains unchanged. In 2006, both past 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 past information about HIV status helps reduce the uncertainty, resulting in a small increase of the lower bound with the upper bound unchanged. Notice 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 (All) and for the three administrative regions of the country: North, Centre and South. According to the complete-case 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 socio-demographic groups.

**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 | .044 | .335 | .291 | .043 | .335 |
| | | Dynamic | .044 | .158 | .114 | .043 | .158 |
| | North | Worst-case | .033 | .261 | .228 | .032 | .261 |
| | | Dynamic | .033 | .116 | .083 | .032 | .117 |
| | Center | Worst-case | .041 | .426 | .385 | .041 | .427 |
| | | Dynamic | .041 | .180 | .139 | .041 | .180 |
| | South | Worst-case | .056 | .314 | .258 | .055 | .315 |
| | | Dynamic | .056 | .174 | .118 | .055 | .175 |
| 2006 | All | Worst-case | .031 | .401 | .370 | .031 | .401 |
| | | Dynamic | .052 | .290 | .238 | .052 | .290 |
| | North | Worst-case | .027 | .337 | .310 | .026 | .338 |
| | | Dynamic | .038 | .251 | .213 | .038 | .252 |
| | Center | Worst-case | .027 | .415 | .388 | .026 | .416 |
| | | Dynamic | .044 | .269 | .225 | .043 | .270 |
| | South | Worst-case | .038 | .445 | .407 | .038 | .446 |
| | | Dynamic | .073 | .346 | .273 | .073 | .347 |
| 2008 | All | Worst-case | .030 | .463 | .433 | .030 | .463 |
| | | Dynamic | .060 | .463 | .403 | .060 | .463 |
| | North | Worst-case | .032 | .445 | .413 | .032 | .446 |
| | | Dynamic | .048 | .445 | .397 | .048 | .446 |
| | Center | Worst-case | .022 | .412 | .39 | .022 | .413 |
| | | Dynamic | .048 | .412 | .364 | .048 | .413 |
| | South | Worst-case | .035 | .529 | .494 | .035 | .530 |
| | | Dynamic | .082 | .529 | .447 | .082 | .530 |

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

**Table 3.** Bounds by gender and birth cohort.

| | | | Men | | | | | Women | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Cohort | Bounds type | L† | U‡ | W§ | Lower CI§§ | Upper CI§§ | L† | U‡ | W§ | Lower CI§§ | Upper CI§§ |
| 2004 | A | Worst-case | .002 | .225 | .223 | .002 | .228 | .011 | .314 | .304 | .010 | .317 |
| | | Dynamic | .002 | .069 | .067 | .002 | .071 | .011 | .139 | .129 | .010 | .141 |
| | B | Worst-case | .021 | .275 | .254 | .021 | .278 | .062 | .380 | .318 | .062 | .382 |
| | | Dynamic | .021 | .107 | .086 | .021 | .109 | .062 | .184 | .121 | .062 | .185 |
| | C | Worst-case | .038 | .405 | .367 | .037 | .406 | .060 | .328 | .268 | .059 | .331 |
| | | Dynamic | .038 | .178 | .141 | .037 | .178 | .060 | .166 | .106 | .059 | .168 |
| | D | Worst-case | .067 | .36 | .294 | .066 | .362 | .061 | .295 | .234 | .060 | .296 |
| | | Dynamic | .067 | .184 | .117 | .066 | .185 | .061 | .125 | .064 | .060 | .126 |
| 2006 | A | Worst-case | .000 | .408 | .408 | .000 | .410 | .011 | .484 | .474 | .010 | .487 |
| | | Dynamic | .002 | .275 | .273 | .002 | .278 | .017 | .342 | .326 | .016 | .345 |
| | B | Worst-case | .008 | .435 | .426 | .008 | .438 | .043 | .424 | .381 | .042 | .426 |
| | | Dynamic | .022 | .323 | .301 | .021 | .326 | .079 | .297 | .218 | .078 | .299 |
| | C | Worst-case | .023 | .356 | .332 | .022 | .359 | .078 | .339 | .261 | .077 | .341 |
| | | Dynamic | .042 | .265 | .223 | .040 | .268 | .097 | .244 | .148 | .095 | .246 |
| | D | Worst-case | .038 | .358 | .320 | .037 | .360 | .029 | .321 | .293 | .028 | .323 |
| | | Dynamic | .063 | .273 | .210 | .063 | .275 | .059 | .202 | .143 | .058 | .204 |
| 2008 | A | Worst-case | .005 | .535 | .530 | .005 | .538 | .019 | .569 | .550 | .018 | .572 |
| | | Dynamic | .008 | .535 | .528 | .007 | .538 | .030 | .569 | .539 | .029 | .572 |
| | B | Worst-case | .020 | .524 | .504 | .019 | .527 | .043 | .420 | .377 | .043 | .422 |
| | | Dynamic | .037 | .524 | .487 | .035 | .527 | .091 | .420 | .330 | .089 | .422 |
| | C | Worst-case | .024 | .429 | .405 | .023 | .432 | .066 | .392 | .326 | .065 | .395 |
| | | Dynamic | .044 | .429 | .385 | .043 | .432 | .103 | .392 | .289 | .101 | .394 |
| | D | Worst-case | .027 | .439 | .412 | .026 | .441 | .026 | .346 | .320 | .025 | .348 |
| | | Dynamic | .066 | .439 | .373 | .065 | .441 | .055 | .346 | .290 | .054 | .348 |

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

Table 3 shows that the dynamic bounds are much narrower than the worst-case bounds also for subgroups characterised 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 using the dynamics 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 outside the identification region. We notice that almost all the MCAR point estimates in 2006 and 2008 are implausibly low because they are lower than the lower limit of the dynamic bounds.

### *4.2. Units respondents only*

#### *4.2.1. Adding IV and MIV restrictions*

For unit nonrespondents, given the very limited information available, it is hard to think of any variable that could be used as an IV or a MIV, so we restrict the analysis to unit respondents. The instrumental variables that we consider are all related to the interview process, and include the gender difference between the interviewer and the interviewee (a binary indicator equal to 1 if their gender is the same), the interviewer's experience (a binary indicator equal to 1 for more experienced interviewers), the interviewer's age (categorised in two classes: below age 23 and aged 23 or older), and the month of the first interview attempt (two categories: May–June and July–August). As suggested by Nicoletti and Peracchi (2005), variables related to the interview process can convincingly be used as instruments because they are unlikely to have a direct effect on the variable of interest (HIV status in our case), but are important predictors of nonresponse. For example, having a more experienced interviewer or an interviewer of the same gender as the interviewee is likely to reduce refusal rates. Further, the timing of the first interview attempt is likely to affect the probability of finding the interviewees at home, especially if these are men who have to follow the cycle of economic activity in the countryside.

As a MIV, we consider the number of sexual partners 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.

While 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 works better, i.e. yields narrower bounds, we only present results based 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, our benchmark are the dynamic bounds for

**Table 4.** Summary of instrumental variables (IV) and monotone instrumental variables (MIV) for unit respondents.

| | 2004 | | 2006 | | 2008 | |
|---|---|---|---|---|---|---|
| | Freq. | % | Freq. | % | Freq.† | % |
| **IV** | | | | | | |
| Interviewer's gender | | | | | n/a | |
| Same | 1350 | 49.0 | 1405 | 62.8 | | |
| Different | 1408 | 51.0 | 831 | 37.2 | | |
| Interviewer's experience | | | | | n/a | |
| No | 1214 | 44.0 | 1087 | 48.6 | | |
| Yes | 1544 | 56.0 | 1149 | 51.4 | | |
| Interviewer's age | | | | | n/a | |
| Young | 1112 | 40.3 | 1111 | 49.7 | | |
| Old | 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 |
| >3 | 580 | 21.0 | 510 | 22.8 | 774 | 27.4 |

†n/a means not applicable because information was not collected.

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. On the other hand, using the number of sexual partners as a MIV 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 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 to the benchmark bounds. Unlike for the whole sample, 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). Using IV or MIV restriction, we are able to obtain for some demographic groups quite narrow bounds. 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).

*4.2.2. Estimates under the MCAR and MAR assumption*

As before, MCAR estimates are obtained as the sample proportion of infected individuals ignoring those with missing HIV status and the availability of covariates, that are used to get the MAR estimates. 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, with weights 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 online supporting material. Table S.8 indicates with a star the point estimates that fall outside the dynamic bounds. Note that the MCAR and MAR point estimates in Figure 2 are very similar and are usually very close to the lower bounds. Our bounds checks show that the MCAR and MAR estimates are often implausibly low because they fall below

the lower bound. Out of the 24 cases considered (8 demographic groups for 3 years), this happens 8 times for the MCAR estimates and 7 times for the MAR estimates.

*4.2.3. Heckman selection model*

We also estimate HIV prevalence using a Heckman-type selection model similar to that used by Bärnighausen et al. (2011). Our exclusion restriction, namely the variable that helps explain the missing data process but not the outcome of interest, is the month of the first interview attempt, which we used as an IV in the previous section. The covariates are the same used for the MAR estimates. Detailed estimation results are presented in Table S.8 of the online supporting material.

Like the point estimates obtained under MCAR and MAR, the Heckman point estimates shown in Figure 2 are often very close to the lower bound but sometimes (8 cases 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 assumed model. We conclude from these bounds checks that, in our data, none of the point estimators considered gives plausible estimates of HIV prevalence for all years and all demographic groups. Therefore, they should be employed carefully and our bounds checks offer an easy way to assess possible violations of the assumptions on which they are based.

## 5. Discussion

Credible estimates of HIV prevalence are critical for policy makers. Today, the gold-standard is estimates obtained from population-based surveys. These surveys 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 absorbing nature of HIV infection.

We show that the identifying power of panel data comes from the fact that the HIV status of current nonrespondents 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 nonrespondents 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.

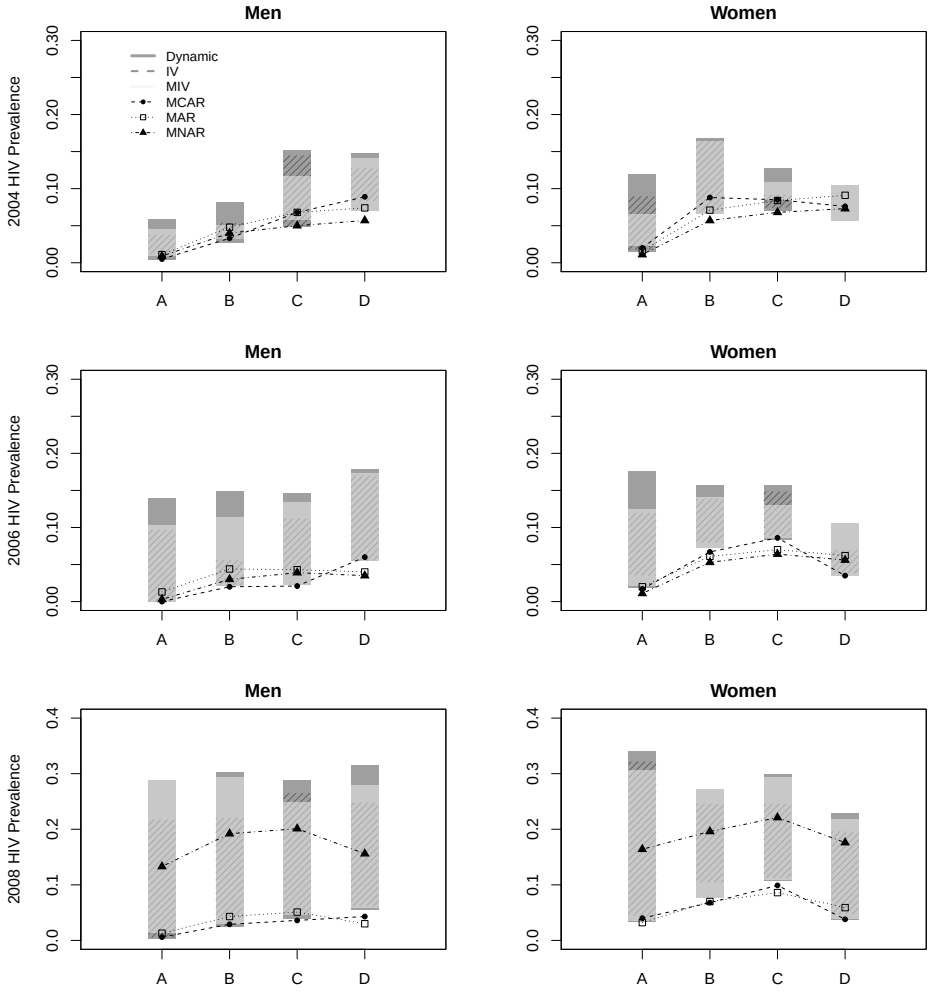


**Figure 2.** HIV prevalence for unit respondents by year, gender and cohort. The graph shows MCAR, MAR, Heckman estimates and dynamic bounds in the benchmark case under IV or MIV restrictions. Cohort A was born in 1984–1989, Cohort B in 1975–1983, Cohort C in 1965–1974 and Cohort D before 1965. The bounds estimated using the interview month as IV have negative width in 2004 for women of Cohort D and are not reported. The dynamic bounds are indicated with a dark grey area, the dynamic bounds with IV with a dashed grey area and the dynamic bounds with MIV with a light grey area.

Applying our bounds to longitudinal data from Malawi, we obtain a reduction of the width of the worst-case bounds by about 17.7 percentage points in 2004, 13.2 percentage points in 2006, and 2.3 percentage points in 2008. Despite

the smaller width, our bounds remain large, especially when unit non respondents are included. When focusing on unit respondents, imposing plausible IV and MIV restrictions helps narrow the bounds. For some of the demographic groups considered, the width of the MIV or IV identification regions is about 2 percentage points and the bounds are informative.

Ignoring the missing data problem and only using the complete cases, gives 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 found that the MCAR and Heckman estimates fall below the lower bound of the identification region in 8 out of the 24 cases considered, while for the MAR estimates this happen 6 times. These bounds checks are a useful reminder of the role 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.

Three additional conclusions may be drawn from our results. First, it is important to keep nonresponse rates low, and to consider unit and item nonresponse separately. Second, including in the data information on the interview process is important because it can be used as a source of instrumental variables. Third, if the data are MNAR, then an effort should be made at interviewing a subset of the nonrespondents. In particular, in longitudinal surveys, it pays off to collect data on people who, for different reasons, did not participate in previous waves.

## Acknowledgments

We thank an Associate Editor and a referee for valuable criticism and suggestions. We also thank Philip Anglewicz and Hans-Peter Kohler for providing the data from the Malawi Diffusion and Ideational Change Project (MDICP) and for support, and participants to the Italian Congress of Econometrics and Empirical Economics 2011 (Pisa, Italy, January 2011), the Understanding Society/BHPS 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 the following grants: ICHD R01HD053781, NICHD R01 HD044228, NICHD R01HD/MH41713.

## References


Anglewicz, P. (2012). Migration, marital change, and HIV infection in Malawi. *Demography 49*(1), 239–265.

Bärnighausen, 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. B. 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. and C. F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association 95*, 77–84.

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.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

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 macro-econometric 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, pp. 259–272. New York: Wiley: eds R.M. Groves, D. Dillman, J. Eltinge, and R. Little.

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 (Eds.), *Advances in Econometrics, Sixth World Congress*, pp. 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. (2011). Policy analysis with incredible certitude. *Economic Journal 121*, F261–F289.

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

Manski, C. F., G. Sandefur, S. McLanahan, and D. Powers (1992). An alternative estimates of the effect of family structure during adolescence on high school graduation. *Journal of the American Statistical Association 87*, 25–37.

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.

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

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

National Statistical Office (NSO) Malawi and ORC Macro (2005). *Malawi Demographic and Health Survey 2004*. Calverton, MD: NSO and ORC Macro.

Nicoletti, C. (2010). Poverty analysis with missing data: Alternative estimators compared. *Empirical Economics 38*, 1–22.

Nicoletti, C. and F. Peracchi (2005). Survey response and survey characteristics: Micro-level evidence from the European Community Household Panel. *Journal of the Royal Statistical Society, Series A 168*, 763–781.

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*, 139–144.

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.

UNGASS (2010). Malawi HIV and AIDS Monitoring and Evaluation Report: 2008-2009. Technical report, United Nation.

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.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*, 1281–1301.