HIV prevalence estimation in the presence of missing data: A bounding approach with panel data

Arpino, Bruno
Bocconi University, Department of Decision Sciences
Via Roentgen 1
20136 Milan, Italy
E-mail: bruno.arpino@unibocconi.it

De Cao, Elisabetta
Bocconi University, Dondena Center
Via Roentgen 1
20136 Milan, Italy
E-mail: elisabetta.decao@unibocconi.it

Peracchi, Franco
Tor Vergata University, Department SEFEMEQ
Via Columbia, 2
I-00133 Rome, Italy
E-mail: franco.peracchi@uniroma2.it

Introduction

The prevalence of HIV in a population is defined as the proportion of people who are infected. Having reliable estimates of the HIV prevalence is essential for policy makers in order to plan control programs and interventions. Since the mid-1980s, the mainstay for monitoring the HIV epidemic has been facility-based sentinel surveillance data. In most cases, estimates of HIV prevalence have been derived from pregnant women attending antenatal clinics (ANC) (2). ANC data have several sources of bias. First, they are only representative of pregnant women who are sexually active, and exclude men. Second, they may provide biased estimates even for the sub-population of pregnant women because of the selective location of the clinics, that are more concentrated in urban areas.

In recent years, many large-scale national surveys have started to include biomarker modules to collect information on HIV serostatus. These biometric surveys are an important new source of data because they accurately measure HIV status and are not restricted to a selected sub-population, as it is the case with ANC-based surveys. Estimates of HIV prevalence derived from biometric surveys are, in general, considerably lower than those based on ANC data (4; 8). Based on these new results, UNAIDS corrected downward HIV prevalence estimates in several countries (2).

Even tough population-based surveys are now considered the "gold standard" to monitor the HIV epidemic (3; 7; 4; 1) these data can be affected by severe source of bias due to missing data on the respondents' HIV status.

The aim of our paper is to study what can be learned about HIV prevalence when data are subject to a nonignorable missing data mechanism. Our approach avoids strong untestable assumptions and switches the focus away from point identification, which typically relies on a combination of strong requirement on the data and strong assumptions about the model, to partial identification (6). The idea is to use empirical evidence alone to identify a region of credible values for the parameter of interest, and then study the identifying power of plausible assumptions to narrow the width of this region. We adopt the partial identification approach to the estimation of HIV prevalence and discuss its use in the context of panel data.

Data

We use data from the Malawi Diffusion and Ideational Change Project (MDICP), a longitudinal survey conducted every two years in rural Malawi. Malawi is one of the countries mostly affected by the HIV epidemic. The national HIV prevalence rate, based on the 2004 Malawi Demographic and Health Survey (MDHS), is equal to 11.8% for adults aged 15–49. The MDHS is a nationally representative cross-sectional survey, and no new data have been collected before 2010. We can instead use the MDICP, that it is a longitudinal survey, to estimate the prevalence for the years 2004, 2006 and 2008 in rural Malawi.

The MDICP survey has been carried out in three of the 28 Malawian districts, one for each of the three administrative regions in the country: Rumphi in the North, Balaka in the South and Mchinji in the Center. The first wave of the survey was carried out in 1998, interviewing 1541 ever-married women of childbearing age and 1198 men. In 2001, the second round of the survey followed-up the same respondents and also interviewed the new spouses of respondents who got married between the two survey rounds (11). In 2004, the sample was augmented with a random sample of about 1,500 married and never-married adolescents (aged 15-28) to correct for aging of the baseline sample over time, and to introduce never-married adolescents given that the original sample was restricted to ever-married women and their husbands. The fourth (2006) and fifth (2008) rounds added the spouses of the adolescents.

The MDICP survey contains information on sexual relations, risk assessments, marriage and partnership histories, household rosters and transfers, as well as income and other measures of wealth. The survey is made up of two parts: the main survey and the biomarker survey, also called voluntary consulting and test (VCT) survey. The first part consisted of the main questionnaire, while the biomarker survey consists of a short questionnaire mainly focused on sexual behavior and questions related to AIDS, and the biomarkers collection. The second part of the survey was administered a few days after the main questionnaire.

Although the survey was not designed to be representative of the population in rural Malawi, the baseline characteristics in 2004 closely match those of the DHS conducted in Malawi in 2004 (10). Moreover, since the HIV tests were administered only in 2004, 2006 and 2008, we only use data from these three waves and we consider the population of 2004 as reference. Measurement error connected with the two types of tests used (oral swabs and blood test) is very limited and, being due to the accuracy limit of the measuring instruments, it can be considered as random. We exclude new entrants in 2006 and 2008, and we drop from the analysis people that were never successfully contacted in none of the waves. The resulting working sample consists of 4062 alive persons in 2004. Since prevalence is defined on the population of alive people, we exclude people who died after 2004.

In each wave where respondents were HIV tested, HIV test status is missing for a substantial number of respondents. First of all, it is important to distinguish between unit and item nonresponse because they can be treated differently. Unit nonresponse occurs when eligible sample units do not to participate to a survey because of failure to establish a contact or refusal to cooperate. Given that the survey is composed by the main survey and the biomarker survey, we define as unit nonresponse the case in which both parts of survey are missing (e.g., the respondent was absent when both the main questionnaire and the biomarker questionnaire took place). Item nonresponse occurs when responding units do not provide useful answers to particular items of the questionnaire. In this paper, the item of interest is the HIV test.

About 55% of the sample corresponds to unit respondents in all three waves (always purple), 12% are respondents in 2004 and nonrespondents in 2006 and 2008 (purple in 2004 and green afterwards), while about 11% of the sample are unit respondents in the first two waves and unit nonrespondents in 2008 (purple in 2004 and 2006, and green in 2008).

The table presents the different sources of missing data. The percentage of missing HIV status is very high in each wave with a peak of 42% in 2008. As argued by (9), an important reason for missing data across all waves is permanent migration. A small percentage of unit nonresponse is due to refusal to participate in the survey or being hospitalized. Other causes of unit nonresponse are lumped into the category 'other', most of them being people who did not fulfill the questionnaire for unknown reasons or because they were too old or too sick. A particular case of unit nonresponse is when people refused to get tested. In comparison to the Malawi DHS (2004), the HIV test refusal rate is relatively low. This may be due to the fact that respondents were not required to learn their results at the time of testing, or to the method of testing through saliva (10). In very few cases the HIV test did not give a clear result (indeterminate) or the test results were lost. The other causes of item nonresponse are grouped in the category 'other' that corresponds to people that fulfill the first part of the questionnaire, but not the second, for example because they were temporarily absent during the biomarker collection.

The classification of different sources of missing data is important because it affects the HIV prevalence estimation in different ways. If we ignore the missing data in the MDICP and derive the complete case estimates (under the MCAR assumption) we obtain an HIV prevalence of 6.15% in 2004, 4.86% in 2006, and 5.24% in 2008, as reported on Table . The 2004 prevalence in the MDICP sample based on the complete case estimates is substantially lower than the rural MDHS prevalence (equal to 10.8%). This might be because the MDICP sample does not include peri-urban areas.

Model

We now introduce the bounding approach to partial identification (5; 6) and then considering its extension to the case of panel data on HIV.

Consider a population that, at a given point in time t, consists of N_t living individuals who can be either susceptible to HIV or infected. HIV status of individual i is represented by the binary indicator y_{it}, which is equal to 1 if individual i is HIV positive at time t and is equal to zero otherwise. HIV prevalence at time t is given by: $\text{Prevalence}_t = \Pr(Y_t = 1) = \sum_{i=1}^{N_t} y_{it}/N_t$, where Y_t is a random variable that represents the variability of the indicator of HIV status in the population. Thus, HIV prevalence is just the proportion of HIV infected people. Our aim is to estimate $\Pr(Y_t = 1)$ from sample surveys when HIV status may be missing for some cases.

We first consider the problem of bounding HIV prevalence when data are only available at a given point in time, as in a cross-sectional survey or when the longitudinal nature of a survey is not exploited. By the law of total probability, we can write HIV prevalence at time t as

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

where D_t is a binary indicator equal to one if HIV status is known and to 0 otherwise. The missing data problem arises because the data tell us nothing about $\Pr(Y_t = 1|D_t = 0)$. On the other hand, we know that necessarily $0 \leq \Pr(Y_t = 1|D_t = 0) \leq 1$. Substituting, $\Pr(Y_t = 1|D_t = 0) = 0$ or $\Pr(Y_t = 1|D_t = 0) = 1$ in equation (1) we obtain the following lower and upper bounds on the HIV prevalence:

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

We will refer to these bounds as worst-case bounds. The identification region for $\Pr(Y_t = 1)$ lies between the lower bound and the upper bound, and its width is given by:

$$W_t = UB_t - LB_t = \Pr(D_t = 0).$$

The width of the interval of logically plausible values for HIV prevalence is equal to the nonresponse probability $\Pr(D_t = 0)$ that represents a measure of the uncertainty about the HIV prevalence caused by nonresponse.

Now consider the case when panel data are available. We know from medical research that HIV is an absorbing state: a person infected in period t has zero probability of becoming susceptible in period $t+1$, while a person susceptible in period t was also susceptible in period $t-1$ with probability one. This simple consideration helps narrowing the worst-case bounds. We report as example the case of two waves, where we have panel data at time t and $t+1$, and we want to bound the HIV prevalence at time t. The unknown probability to find is: $\Pr(Y_t = 0|D_t = 0)$. We use future HIV status to derive some information about the nonrespondents.

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

From the previous equation, we obtain the following bounds and width:

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

Having information at time $t+1$ helps in narrowing the bounds, in fact $W_{t(+1)}$ is lower than W_t obtained with cross-section data. This is because among the nonrespondents at time t we recover the HIV status for some of them at time $t+1$. However among the respondents at time $t+1$, only the information about the HIV negative status can be used to impute the missing HIV status at time t, and this reduces the upper bound. While respondents at time $t+1$ who are HIV positive cannot be assumed already HIV positive at time t. As a consequence the lower bound does not change. Vice-versa, if we have panel data at time t and $t-1$, and we want to bound the HIV prevalence at time t, past positive HIV status increases the lower bound with respect to the cross-sectional case, but the upper bound does not change. Increasing the number of future or past waves decreases the width of the identification region.

In the case of unit nonresponse we lack information, not just on HIV status, but also on other variables. In the case of item nonresponse, instead, HIV status is missing but most of the other variables, such as respondents' characteristics, socio-demographic variables, or characteristics of the data collection process, are available.

We apply instrumental variables (IV) and monotone instrumental variable (MIV) restrictions to the bounds on HIV prevalence for unit respondents. The reason being that IV and MIV are only available for unit respondents. The bounds we obtain are not directly comparable with the bounds found before, that are bounds on HIV prevalence for the entire sample. The IV considered are: age of the interviewer; difference in gender between interviewer and respondent; interviewer's experience; and month of the interview. The MIV is, instead, the number of sexual partners the respondent had till that year.

Results and conclusions

Using longitudinal data from the Malawi Diffusion and Ideational Change Project (MDICP), we find that the presence of missing data translates into a substantial uncertainty about the HIV prevalence rate in the population. Our results show that using panel data and the absorbing nature of HIV helps in shrinking the worst-case bounds. Longitudinal data are typically used to study incidence rates. However, they can also be used to estimate the prevalence at different points in time for the same population. Overall, the identification region interval produced by the worst-case bounds is

between 3.8% and 34.2% in 2004, between 4.5% and 28.9% in 2006, and between 2.4% and 46.6% in 2008 (Table **??**). The width of the interval increases over time. The identification region interval given by the dynamic bounds is between 3.8% and 15.9% in 2004, between 2.6% and 40.2% in 2006, and between 4.9% and 46.6% in 2008. Therefore, it is easy to see that the dynamic bounds, or the information about the transition from one status - susceptible - to another status - infected - helps us to reduce the worst-case scenario by about 18.2 percentage points in 2004, by 13.2 percentage points in 2006, and by 2.4 percentage points in 2008, but with more time periods we would have obtained smaller intervals. Moreover, introducing plausible instrumental and monotone instrumental variable restrictions help us in narrowing the bounds based on the unit respondents even further. If we ignore the missing data and we rely on the complete case estimates, we obtain an HIV prevalence that is very close to our lower bounds. According to our bounds, the HIV prevalence could be much higher, as a larger part of the non respondents could be infected.

However, our approach is easy to implement, it does not require any assumptions about the nature of the missing data, and it allows to obtain reliable intervals from a statistical point of view. The HIV prevalence can take any possible values between the bounds, producing intervals that could be useful from a policy prospective. The main caveat of our bounding approach is that the intervals remain too big to be able to derive conclusions about the trend in HIV prevalence. For the same reason, it is also difficult to compare subgroups of the population. Moreover, we stress the fact that it is important to well-design surveys to reduce nonresponse, either unit and item nonresponse. It is also critical to include in the data information, such as interviewer's characteristics, fieldwork procedures etc, as they can be used as instrumental variables.

Distribution of respondents across waves

		2004		**2006**		**2008**
	Freq.	Percent	Freq.	Percent	Freq.	Percent
UNIT RESPONDENTS						
HIV negative	2700	66.47	2408	59.28	2116	52.09
HIV positive	177	4.36	123	3.03	117	2.88
Items nonresponse						
Test refused	256	6.30	200	4.92	172	4.23
Indeterminate	14	0.34	6	0.15	1	0.02
Results lost	24	0.59	0	0.00	0	0.00
Other*	319	7.85	313	7.71	569	14.01
UNIT NONRESPONDENTS						
Refused	27	0.66	11	0.27	58	1.43
Moved	184	4.53	479	11.79	470	11.57
Temporarily absent	36	0.89	41	1.01	76	1.87
Hospitalized	6	0.15	5	0.12	1	0.02
Other**	319	7.85	432	10.64	359	8.84
Dead	/	/	44	1.08	123	3.03
Total	4062	100	4062	100	4062	100
HIV prevalence "Complete cases"		6.15%		4.86%		5.24%
% of HIV status missing		**29.17%**		**36.61%**		**42.00%**

The new entrants 2006/2008 are excluded.

* The category other item nonrespondents corresponds to people that fulfill the first part of the questionnaire, but not the second, for example because they were temporarily absent during the biomarker collection.

** The majority of unit nonrespondents categorized in the class other corresponds to people who did not fulfill the questionnaire for unknown reasons or because too old or too sick.

REFERENCES (RÉFERENCES)

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

[2] R. Brookmeyer. Measuring the hiv/aids epidemic: Approaches and challenges. *Epidemiologic Reviews*, 32:26–37, 2010.

[3] J. Garcia-Calleja, E. Gouws, and P. Ghys. National population based hiv prevalence surveys in sub-saharan africa: Results and implications for hiv and aids estimates. *Sexually Transmitted Infections*, 82(Suppl III):iii64iii70, 2006.

[4] E. Gouws, V. Mishra, and T.B. Fowler. 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, 2008.

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

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

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

[8] L.S. Montana, V. Mishra, and R. Hong. Measuring the hiv/aids epidemic: Approaches and challenges. *Sexually Transmitted Infections*, 84(1):i78–i84, 2008.

[9] F. Obare. Nonresponse in repeat population-based voluntary counseling and testing for hiv in rural malawi. *Demography*, 47(3):651–665, 2010.

[10] R.L. Thornton. The demand for, and impact of, learning hiv status. *American Economic Review*, 98:1829–1863, 2008.

[11] S. C. Watkins, E. M. Zulu, H. P. Kohler, and J. R. Behrman. Introduction to: Social interactions and hiv/aids in rural africa. *Demographic Research*, Special Collection 1(1):1–30, 2003.