10-2014

Bound estimator of HIV prevalence: Application to Malawi

Tomoki FUJII
Singapore Management University, tfujii@smu.edu.sg

Denis H. Y. LEUNG
Singapore Management University, denisleung@smu.edu.sg

SMU ECONOMICS & STATISTICS
WORKING PAPER SERIES

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Bound estimator of HIV prevalence: Application to Malawi

Tomoki Fujii and Denis H.Y. Leung

October 2014

Paper No. 17-2014

Bound estimator of HIV prevalence: Application to Malawi

Tomoki Fujii and Denis H.Y. Leung

School of Economics, Singapore Management University, Singapore

September 29, 2014

Abstract

Objective

To find lower and upper bounds of HIV prevalence in Malawi under mild and intuitive assumptions to assess the importance of the refusal issue in the estimation of HIV prevalence.

Methods

We derive bounds based on the following two key assumptions: (i) Among those who have never taken an HIV test before, those who refuse to take an HIV test (hereafter "refusers") have at least as much risk to be HIV positive as those who participate in the HIV test, and (ii) among the refusers, those who have a prior testing experience are at least as likely to be HIV positive as those who have no prior experience. We compute the bounds using the Malawi Demographic and Health Survey and a longitudinal data set with a HIV testing component collected in the Malawi Diffusion and Ideational Change Project disaggregated by the sex, urban/rural areas, and three regions of Malawi.

Findings

The bounds of HIV prevalence vary substantially across geographic and demographic groups. In particular, the bounds for males are tighter than those for females and the bounds for the Northern region are also tighter than those for other regions. There is no substantial difference in the width of bounds between the rural and urban populations.

Conclusion

Bounds are useful for assessing the influence of refusal bias without the need for strong assumptions. Refusal issue is less of a concern if bounds are tight. However, when bounds are wide, refusal issue may be important.

Keywords: Bias; Demographic and Health Surveys; Malawi; Missing data; Non-response; Refusals; Surveys

Introduction

In sub-Saharan Africa, home to around 23 million people living with HIV,[1] accurate measurement of HIV prevalence is essential for policy planning and resource allocation. Demographic and Health Surveys (DHS) and other national population-based surveys have served as important data sources for such measurement in the past three decades.[2,3] These surveys are useful because they contain detailed demographic and risk characteristics and possibly outcome variables of interest (e.g., HIV status) at the individual and household level. However, significant refusal rates in these surveys are often reported[4] and failure to address the refusal issue may severely undermine the reliability of estimates, because those who refuse may be systematically different from the survey participants.[5,6]

In this paper, we propose two methods to estimate lower and upper bounds of HIV prevalence in the presence of refusals (the "refusers") in a DHS survey under two mild and intuitive assumptions: (A1) For those who have never taken an HIV test before, the refusers are at least as likely to be HIV positive as the test participants; (A2) Among the refusers, those who have taken an HIV test before are at least as likely to be HIV positive as those without a previous testing experience.

The first method requires the existence of a set of supplementary longitudinal data, which permits estimation of the relative risk of HIV between participants who have never taken an HIV test before and the refusers who have taken a test before. In the second method, the researcher specifies this relative risk to study the sensitivity of the bounds to the relative risk. Thus, this method offers a practical alternative to evaluatethe significance of the refusal issue when a suitable longitudinal data set is unavailable.

Data sources

The primary data source for this study is the 2004 Malawi Demographic Health Survey (MDHS), which is a nationally-representative survey. In addition, we use a longitudinal data set collected under the Malawi Diffusion and Ideational Change Project (MDICP). To focus on refusals, we exclude all missing observations due to non-contact.[a]

MDHS

The 2004 MDHS is a two stage survey using households from 28 districts in Malawi. All women aged 15-49 in a selected household were eligible for interview. In about one in three selected households, male members of the household aged 15-54 were also surveyed and an HIV test was offered to both male and female members.[8] However, HIV test was successfully carried out only for 67 percent of the eligible

[a] A major reason for non-contact is migration, which is potentially an important issue as migrants appear to have higher HIV prevalence than non-migrants in Malawi.[7] While migration is beyond the scope of this study, the bound estimators presented below can be extended to include migrants by re-interpreting refusal as migration.

individuals with refusal accounting for the majority of missing HIV status for the non-participants.[8] The proportion of refusals was higher in the Central region (26 percent) than in the Southern or the Northern regions (21 percent and 14 percent, respectively). The refusal rate was slightly higher in urban (25 percent) than in rural areas (22 percent).

We confine the 2004 MDHS sample to those aged under 49 to make the male and female populations comparable. Following a Malawi National Statistical Office report, we omit Lilongwe district, which has an unusually high refusal rate and low observed prevalence.[8] We also exclude those who refused to answer the individual questionnaire, those whose HIV testing results are not available for reasons other than refusal (e.g., non-contact and technical problem), and those whose previous HIV testing status is unknown. As a result, we have a total of 6,343 eligible individuals (3,511 women and 2,832 men) in our sample.

MDICP

The MDICP is an ongoing longitudinal study, which includes married women and their husbands randomly drawn from 120 villages in a total of three rural districts with one district from each of the Southern, Central, and Northern regions. While the initial sample is not designed to be representative of rural Malawi, its sample characteristics closely matched those of the *rural* sample in the 1996 MDHS.[9]

We restrict our sample to those aged 15-49 who appear in both the third and fourth phases (MDICP-3 and -4) conducted in 2004 and 2006 with known previous HIV testing status and non-missing HIV test results at the time of MDICP-3. As a result, we have a total of 2,287 individuals (1,240 women and 1,047 men) in the sample.

Methods

Let D_i be an indicator variable that takes one if individual i is HIV positive and zero otherwise. The goal is to estimate $\rho \equiv P(D_i = 1)$, where individual i is drawn randomly from the population of interest. When we are interested in the HIV prevalence of a certain sub-population, we simply need to use a suitable sub-sample.

We typically estimate ρ from surveys such as DHS. However, in the presence of refusal, the data are sufficient to estimate $E[D_i \mid R_i = 0]$ where $R_i = 0[R_i = 1]$ indicates that individual i accepts [refuses] an HIV test. However, $E[D_i \mid R_i = 0]$ in general is not the same as ρ because some respondents, particularly those who already know they are (likely to be) HIV positive, may refuse to take an HIV test out of fear that their HIV status be known to others.

Since we generally know little about the reasons for refusals, we only wish to make weak and plausible assumptions. To motivate assumptions (A1) and (A2), we note the following two points: First, individuals may know the risk of HIV infection even without HIV tests because they know their behavior. Hence, those who are at a higher risk of HIV may be more likely to refuse HIV tests. Second, previous testing experience may be informative of current HIV status because, for example, those who have engaged in risky sexual behaviour may be more likely to take HIV tests out of necessity. Based on these considerations, we propose to estimate lower and upper bounds of ρ under the following assumptions:

$$P(D_i = 1 | T_i = 0, R_i = 0) \le P(D_i = 1 | T_i = 0, R_i = 1) \le P(D_i = 1 | T_i = 1, R_i = 1), \quad (1)$$

where $T_i = 1$ $[T_i = 0]$ means that the subject has [never] taken an HIV test before. The first and second inequalities in eq. (1) are respectively the mathematical restatement of assumptions (A1) and (A2).

To derive the bound estimators, we first decompose ρ as follows:

$$\begin{aligned}\rho = {} & P(D_i = 1, R_i = 0) + P(D_i = 1 | T_i = 1, R_i = 1) P(T_i = 1, R_i = 1) \\ & + P(D_i = 1 | T_i = 0, R_i = 1) P(T_i = 0, R_i = 1). \end{aligned} \quad (2)$$

By applying eq. (1) to eq. (2), we obtain the following lower bound ρ_- and upper bound ρ_+ satisfying $\rho_- \le \rho \le \rho_+$:

$$\begin{aligned}\rho_- = {} & P(D_i = 1, R_i = 0) + P(D_i = 1 | T_i = 1, R_i = 1) P(T_i = 1, R_i = 1) \\ & + P(D_i = 1 | T_i = 0, R_i = 0) P(T_i = 0, R_i = 1) \end{aligned} \quad (3)$$

$$\begin{aligned}\rho_+ = {} & P(D_i = 1, R_i = 0) + P(D_i = 1 | T_i = 1, R_i = 1) P(T_i = 1, R_i = 1) \\ & + P(D_i = 1 | T_i = 1, R_i = 1) P(T_i = 0, R_i = 1). \end{aligned} \quad (4)$$

These bounds cannot be directly calculated from DHS, because $P(D_i = 1 | T_i = 1, R_i = 1)$ is unknown in general. However, with a suitable longitudinal data set, it may be possible to estimate this quantity under some additional assumptions. Below, we first develop a method when such a longitudinal data set is available. We then consider a practical solution in the absence of a longitudinal data set.

Method 1: When auxiliary longitudinal data is available

In our empirical example, we know the HIV status of the refusers in MDICP-4 from the MDICP-3 test results. However, since MDICP is not nationally representative, it would be inappropriate to estimate $P(D_i = 1 | T_i = 1, R_i = 1)$ directly from the MDICP data. Therefore, we explicitly account for the non-representativeness of the MDICP data. Let $M_i = 1$ be an indicator variable for individual i belonging to the MDICP population. Further, we assume that the relative risk of HIV between the

MDICP population and non-MDICP population is independent of refusal among those who have previously taken an HIV test. This assumption implies:

$$\begin{aligned} Z &\equiv \frac{P(D_i = 1 \mid T_i = 1, M_i = 0)}{P(D_i = 1 \mid T_i = 1, M_i = 1)} \\ &= \frac{P(D_i = 1 \mid T_i = 1, R_i = 1, M_i = 0)}{P(D_i = 1 \mid T_i = 1, R_i = 1, M_i = 1)} \\ &= \frac{P(D_i = 1 \mid T_i = 1, R_i = 0, M_i = 0)}{P(D_i = 1 \mid T_i = 1, R_i = 0, M_i = 1)}. \end{aligned} \quad (5)$$

The numerator and denominator of the last line of eq. (5) can be estimated by the proportions of HIV positive among those who have previously taken an HIV test in MDHS and MDICP data, respectively. Using eq. (5), the following holds:

$$\begin{aligned} P(D_i = 1 \mid T_i = 1, R_i = 1) &= P(D_i = 1 \mid T_i = 1, R_i = 1, M_i = 1) P(M_i = 1 \mid T_i = 1, R_i = 1) \\ &\quad + P(D_i = 1 \mid T_i = 1, R_i = 1, M_i = 0) P(M_i = 0 \mid T_i = 1, R_i = 1) \\ &= P(D_i = 1 \mid T_i = 1, R_i = 1, M_i = 1) \cdot \left[P(M_i = 1 \mid T_i = 1, R_i = 1) + Z P(M_i = 0 \mid T_i = 1, R_i = 1) \right], \end{aligned} \quad (6)$$

where $P(D_i = 1 \mid T_i = 1, R_i = 1, M_i = 1)$ can be estimated by the proportion of HIV-positive individuals among the refusers in MDICP-4. Note that everyone in the MDICP sample has taken an HIV test in MDICP-3.

We can interpret $P(M_i = 0 \mid T_i = 1, R_i = 1)$ and $P(M_i = 1 \mid T_i = 1, R_i = 1)$ as the urban and rural population shares, respectively, of Malawi among the individuals with $T_i = 1$ and $R_i = 1$. We estimate them by the urban and rural shares of the sample weights, respectively, among those refusers with a prior testing experience in the MDHS data. Once we have an estimate of $P(D_i = 1 \mid T_i = 1, R_i = 1)$, all the remaining terms in eq. (4) can be estimated from the MDHS data.[b]

To obtain ρ_-, we additionally need to compute $P(D_i = 1 \mid T_i = 0, R_i = 0)$. Similar to the derivation of Z, define:

$$Z' = \frac{P(D_i = 1 \mid T_i = 0, R_i = 0, M_i = 0)}{P(D_i = 1 \mid T_i = 0, R_i = 0, M_i = 1)}. \quad (7)$$

As with Z, the denominator and numerator of Z' can be estimated from the MDHS and MDICP data, respectively, using the proportion of HIV-positive among the non-refusers with no previous HIV-testing experience. This leads to:

[b] We use the MDHS sample weights to calculate $P(M_i \mid T_i, R_i)$ in eqs. (6) and (8) below. For the rest, we chose not to apply the weights in our main results as we do not have corresponding weights in the MDICP data. In the Appendix, we consider alternative weighting schemes and show that our main results remain unaffected by the choice of weights.

$$P(D_i = 1 \mid T_i = 0, R_i = 0) = P(D_i = 1 \mid T_i = 0, R_i = 0, M_i = 1) \cdot \big[P(M_i = 1 \mid T_i = 0, R_i = 0) + Z' P(M_i = 0 \mid T_i = 0, R_i = 0)\big]. \quad (8)$$

Note here that eq. (1) is an assumption that has to be empirically validated. When the assumption is violated, the lower bound is not guaranteed to be smaller than the upper bound.

Method 2: When auxiliary longitudinal data is unavailable

We now turn to the case where a longitudinal data set is not available. In this case, we are generally unable to calculate the bounds based on eqs. (6) and (8). However, it is still possible to evaluate the influence of refusals. To see this point, define the relative risk $k(\geq 1)$ of HIV between non-refusers with no previous testing experience and refusers with previous testing experience:

$$k = \frac{P(D_i = 1 \mid T_i = 1, R_i = 1)}{P(D_i = 1 \mid T_i = 0, R_i = 0)},$$

which is the ratio of the right-hand-side to the left-hand-side in eq. (1). Given k, we can use the following expressions of ρ_- and ρ_+, which can be estimated from MDHS:

$$\rho_- = P(D_i = 1, R_i = 0) + P(D_i = 1 \mid T_i = 0, R_i = 0)[kP(T_i = 1, R_i = 1) + P(T_i = 0, R_i = 1)] \quad (9)$$

$$\rho_+ = P(D_i = 1, R_i = 0) + P(D_i = 1 \mid T_i = 0, R_i = 0)k[P(T_i = 1, R_i = 1) + P(T_i = 0, R_i = 1)] \quad (10)$$

These expressions show that the width of the bounds, $\rho_+ - \rho_-$, is driven by three factors: $P(D_i = 1 \mid T_i = 0, R_i = 0)$, $P(T_i = 0, R_i = 1)$, and k. Therefore, when both $P(D_i = 1 \mid T_i = 0, R_i = 0)$ and $P(T_i = 0, R_i = 1)$ are small, even for a conservative value of k, the bounds are relatively tight.

When $k = 1$, the two inequalities in eq. (1) are held with equality and the following holds by eqs. (2), (9), and (10):

$$\rho = \rho_- = \rho_+ = P(D_i = 1 \mid R_i = 0)P(R_i = 0) + P(D_i = 1 \mid T_i = 0, R_i = 0)P(R_i = 1). \quad (11)$$

Notice that these bounds are different from the complete-case estimator even when $k = 1$, because the complete-case estimator is consistent for $P(D_i = 1 \mid R_i = 0)$ which is the same as eq. (11) if and only if $P(D_i = 1 \mid T_i = 0, R_i = 0) = P(D_i = 1 \mid R_i = 0)$. This difference can also be seen from the fact that the complete-case estimator does not make use of information from the previous testing experience.

Results

In theory, the foregoing derivations allow us to apply the methods to any sub-population of interest. However, when using eqs. (3) and (4) with Method 1, we must ensure that the bounds are empirically

consistent with eq. (1). This issue is especially important when the relevant sub-sample is small and a few positive HIV cases can significantly influence the results. To avoid this problem, we let k to depend only on the sex of individuals while other quantities in eqs. (6) and (8) such as $P(D_i = 1, R_i = 0)$, $P(T_i = 1, R_i = 0)$, and $P(T_i = 1, R_i = 1)$ are allowed to depend both on the location of residence as well.

Table 1: Method 1 estimates of ρ_- and ρ_+.

		Male		Female		Total	
		ρ_-	ρ_+	ρ_-	ρ_+	ρ_-	ρ_+
North	Rural	0.0393	0.0418	0.0726	0.1043	0.0565	0.0741
	Urban	0.1300	0.1304	0.1987	0.2214	0.1645	0.1761
	Total	0.0609	0.0630	0.1055	0.1349	0.0839	0.1000
Central	Rural	0.0662	0.0701	0.0947	0.1359	0.0808	0.1038
	Urban	0.1009	0.1038	0.1331	0.1574	0.1167	0.1301
	Total	0.0679	0.0717	0.0964	0.1369	0.0824	0.1049
South	Rural	0.1140	0.1182	0.1762	0.2148	0.1465	0.1688
	Urban	0.1561	0.1602	0.2321	0.2716	0.1936	0.2152
	Total	0.1213	0.1255	0.1835	0.2222	0.1535	0.1757
Total	Rural	0.0863	0.0902	0.1345	0.1732	0.1112	0.1331
	Urban	0.1431	0.1462	0.2100	0.2423	0.1761	0.1937
	Total	0.0943	0.0981	0.1438	0.1817	0.1197	0.1410

The estimates based on Method 1 are given in Table 1. A few notable patterns emerge from this table. First, there are sizeable geographic variations in HIV prevalence in Malawi with the Southern region having a substantially higher prevalence than Northern and Central regions. Second, in each of these three regions, urban prevalence is substantially higher than rural prevalence. In fact, the rural upper bound is lower than the urban lower bound for both males and females in all regions with the exception of female prevalence in the Central region. Third, the tightness of the bounds varies across regions. The bounds in the Northern region are tighter than those in the Central and Southern regions. These three points indicate that the policies to tackle HIV would need to take into account the geographic differences in HIV prevalence.

Fourth, there is also a wide gap between male and female HIV prevalences. Male upper bound is lower than female lower bound in all locations we considered. Further, the bounds for males are generally much tighter than those for females. This is in part because men generally have lower HIV prevalence. In particular, the HIV prevalence for the participants with no previous HIV testing (i.e., $P(D_i \mid T_i = 0, R_i = 0)$) is 0.0951 for males and 0.1464 females. The empirically obtained values of k for males ($k = 1.282$) is also lower than that for females ($k = 3.297$). These two factors contribute to

the tighter bounds for males. On the other hand, $P(T_i = 0, R_i = 1)$ for males is slightly larger than that for females (0.1999 for males and 0.1903 for females).

Even when an auxiliary longitudinal data set like MDICP is unavailable, it is possible to evaluate how serious the refusal issue may be by varying the values of k within a reasonable range and thereby checking the sensitivity of the bounds with respect to k using Method 2. Note that the left-hand-side of eq. (1) is guaranteed to be no larger than the right-hand-side for any $k(\geq 1)$ in Method 2.

Table 2: Method 2 estimates of ρ_- and ρ_+.

Value of k			k=1	k=2		k=3		k=4	
Bound			$\rho_- = \rho_+$	ρ_-	ρ_+	ρ_-	ρ_+	ρ_-	ρ_+
Male	North	Rural	0.0337	0.0345	0.0394	0.0354	0.0451	0.0363	0.0508
		Urban	0.1317	0.1317	0.1348	0.1317	0.1378	0.1317	0.1408
		Total	0.0596	0.0608	0.0672	0.0619	0.0749	0.0630	0.0825
	Central	Rural	0.0642	0.0661	0.0788	0.0680	0.0934	0.0699	0.1079
		Urban	0.1159	0.1224	0.1449	0.1288	0.1739	0.1353	0.2029
		Total	0.0668	0.0688	0.0821	0.0709	0.0975	0.0729	0.1128
	South	Rural	0.1257	0.1291	0.1549	0.1326	0.1841	0.1360	0.2133
		Urban	0.1914	0.2064	0.2484	0.2214	0.3054	0.2364	0.3624
		Total	0.1361	0.1410	0.1692	0.1459	0.2023	0.1507	0.2354
	Total	Rural	0.0897	0.0921	0.1094	0.0946	0.1292	0.0971	0.1489
		Urban	0.1655	0.1752	0.2036	0.1848	0.2416	0.1944	0.2797
		Total	0.1002	0.1034	0.1224	0.1065	0.1445	0.1097	0.1667
Female	North	Rural	0.0633	0.0644	0.0735	0.0656	0.0838	0.0667	0.0940
		Urban	0.2123	0.2203	0.2460	0.2283	0.2797	0.2364	0.3134
		Total	0.1026	0.1049	0.1193	0.1073	0.1361	0.1097	0.1528
	Central	Rural	0.0894	0.0917	0.1097	0.0941	0.1301	0.0964	0.1504
		Urban	0.1353	0.1353	0.1482	0.1353	0.1611	0.1353	0.1740
		Total	0.0915	0.0938	0.1116	0.0960	0.1317	0.0983	0.1519
	South	Rural	0.1934	0.1996	0.2367	0.2058	0.2800	0.2119	0.3233
		Urban	0.2680	0.2838	0.3377	0.2995	0.4075	0.3153	0.4772
		Total	0.2026	0.2097	0.2488	0.2169	0.2950	0.2240	0.3411
	Total	Rural	0.1397	0.1436	0.1699	0.1475	0.2002	0.1514	0.2304
		Urban	0.2312	0.2414	0.2791	0.2516	0.3270	0.2619	0.3749
		Total	0.1509	0.1554	0.1832	0.1599	0.2156	0.1644	0.2480

In Table 2, we report the lower and upper bounds for $k \in \{1,2,3,4\}$ calculated solely from the MDHS data set. While the choice of these values is subjective, it could serve as rule-of-thumb figures to use in Africa given the empirical estimtates of *k* used in Method 1.

Table 2 shows that both the lower and upper bounds tend to increase as k goes up. However, when no one with a previous testing experience refuses to participate in the sample, the lower bound does not vary with k. This is indeed the case for males in the rural Northern region and females in the urban

Central region.

The table also shows that the tightness of the bounds varies substantially between the sexes, between urban and rural areas, and across regions. The bounds are reasonably tight when k is less than 2. However, when $k = 4$, the width of the bounds can be as large as 16.2 percentage points (females in the urban Southern region) and as small as 0.9 percentage points (males in the urban Northern region). Overall, the results from Methods 1 and 2 indicate that we need to exercise greater caution when interpreting female HIV prevalence.

Discussion

Existing studies on refusal bias in the estimation of HIV prevalence typically either provide some evidence of the existence of the bias or try to correct for the bias by making some (often strong) behavioral assumptions about the subjects. In this paper, we have instead derived plausible lower and upper bounds for HIV prevalence under mild and intuitive assumptions by exploiting a longitudinal data set in addition to the DHS data set. This study complements the results of an earlier report on the potential bias due to refusal/absence using the MDICP data[10] by showing the significance of the refusal issue in the estimation of the national HIV prevalence based on the MDHS data set.

We find that the prevalence bounds are fairly tight for males and also close to the complete case estimator. Since these bounds are created under very mild conditions, the refusal bias in the complete-case MDHS estimates for males is likely to be small. On the other hand, the bounds for females are much wider. Based on Method 1, the upper bound for women (0.1817) being over three percentage points above the complete-case estimate (0.1521) (Further details available in the Appendix). This may be because women are more likely to suffer from and thus sensitive to the stigma associated with HIV/AIDS. This in turn suggests that the refusers, especially those with a previous testing experience, may be systematically different from the non-refusers. The results based on Method 2 shows that the significance of refusal can be meaningfully evaluated by varying the value k within a plausible range.

Our results also provide encouraging evidence that longitudinal data sets can be fruitfully used to supplement DHS data for drawing inferences, even when the longitudinal study is not nationally representative. This point is important because longitudinal surveys in such settings are often based on more stable populations in rural communities and not necessarily nationally representative (for example, longitudinal studies of malaria in Africa have been typically conducted in rural communities). However, they differ from urban counterparts both in access to treatment and in demographic and socioeconomic characteristics and inferences drawn from these studies cannot be directly extrapolated to the general population. We addressed this issue by explicitly taking into account the differences between the MDICP

and non-MDICP populations.

Given the increasing number of longitudinal health studies,[11, 12, 13, 14] it is likely that we will be able to obtain better empirical estimates of k over time, which in turn would allow us to derive more reliable bounds in areas where relevant longitudinal survey is unavailable.

References

1 WHO, Geneva. *Global HIV/AIDS response: epidemic update and health sector progress towards universal access: progress report 2011*, 2011.

2 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*, 362:1929–1931, 2003.

3 J.M. Garcia-Calleja, E. Gouws, and P. D. Ghys. National population-based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates. *Sexually Transmitted Infections*, 82(Suppl 3):iii64–iii70, 2006.

4 J. Larmarange, R. Vallo, S. Yaro, P. Msellati, N. Meda, and B. Ferry. Estimating effect of non response on hiv prevalence estimates from demographic and health surveys. CePeD Working paper 2009-03, Centre Population et Développement, Paris, 2009.

5 M. Marston, K. Harriss, and E. Slaymaker. Non-response 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 (Suppl 1):i71–i77, 2008.

6 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 1):i63–i70, 2008.

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

8 National Statistical Office and ORC Macro. *Malawi Demographic and Health Survey 2004*. National Statistical Office and ORC Macro, 2005.

9 S. C. Watkins, E. M. Zulu, H.-P. Kohler, and J. R. Behrman. Introduction to social interactions and HIV/AIDS in rural Africa. In *Demographic Research Special Collection 1*. Max-Planck Institute for Demographic Research, Rostock, Germany, 2003.

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

11 G. Andargie, Y. Berhane, A. Worku, and Y. Kebede. Predictors of perinatal mortality in rural population of Northwest Ethiopia: a prospective longitudinal study. *BMC Public Health*, 13:168, 2013.

12 M. Ainsworth and I. Semali. The impact of adult deaths on children's health in Northwestern Tanzania. Working Paper 2266, World Bank, Washington, DC, USA, 2000.

13 J. .C. Davis, T. D. Clark, S. K. Kemble, N. Talemwa, D. Njama-Meya, S. G. Staedke1, and G. Dorsey.

Longitudinal study of urban malaria in a cohort of Ugandan children: description of study site, census and recruitment. *Malaria Journal*, 5:18, 2006.

14 A. R. Quisumbing. Food aid and child nutrition in rural Ethiopia. Discussion paper 158, International Food Policy Research Institute, Washington, DC, USA, 2003.

Acknowledgement
We acknowledge ORC Macro for granting us access to the DHS. We thank the Population Studies Center, University of Pennsylvania for providing us with the MDICP data. In particular, we gratefully acknowledge the help of Dr. Philip Anglewicz for sending us the data and documentations for MDICP-3 and MDICP-4.

Appendix: Additional tables

Tables 3 and 4 are the same as Tables 1 and 2 except that we apply the MDHS sample weights in the calculation of joint probabilities in eqs. (3) and (4) estimated directly from the MDHS (i.e., $P(D_i = 1, R_i = 0)$, $P(T_i = 1, R_i = 1)$, and $P(T_i = 0, R_i = 1)$). The results are generally similar to the unweighted counterpart.

In Table 5, we use the 2008 Population and Housing Census data downloaded from the National Statistical Office website[c] to estimate $P(M_i \mid T_i, R_i)$ used in eqs. (6) and (8) under the additinal assumption that M_i is independent of (T_i, R_i). We do this exercise because the census population estimates are based on the actual household visits and thus likely to be more accurate than the estimates based on the MDHS sample weights. However, the drawback of the census is that we need the independence assumption as we do not have observations of T_i and R_i in the census.

When the census-based estimates of $P(M_i \mid T_i, R_i)$ is used, the gap in the value of k between male and female is slightly larger with $k = 1.055$ for males and $k = 3.324$ for females. However, as the comparison of Table 5 with Table 1 shows, the use of census-based estimates of $P(M_i \mid T_i, R_i)$ does not alter the results much overall.

Finally, we report the complete-case estimates of the HIV prevalence in Table 6.

Table 3: Method 1 estimates of ρ_- and ρ_+ with MDHS sample weights.

		Male		Female		Total	
		ρ_-	ρ_+	ρ_-	ρ_+	ρ_-	ρ_+
North	Rural	0.0363	0.0381	0.0826	0.1144	0.0602	0.0775
	Urban	0.1522	0.1526	0.2436	0.2624	0.1981	0.2077
	Total	0.0508	0.0523	0.1141	0.1434	0.0833	0.0992
Central	Rural	0.0606	0.0632	0.0971	0.1397	0.0793	0.1023
	Urban	0.1209	0.1228	0.1653	0.1957	0.1427	0.1586
	Total	0.0637	0.0662	0.1002	0.1422	0.0823	0.1049
South	Rural	0.1200	0.1225	0.1833	0.2198	0.1531	0.1734
	Urban	0.1574	0.1598	0.2076	0.2484	0.1822	0.2036
	Total	0.1253	0.1279	0.1860	0.2230	0.1568	0.1772
Total	Rural	0.0863	0.0888	0.1405	0.1785	0.1143	0.1352
	Urban	0.1504	0.1524	0.2112	0.2444	0.1804	0.1978
	Total	0.0932	0.0956	0.1476	0.1852	0.1212	0.1416

[c] http://www.nsomalawi.mw/images/stories/data_on_line/demography/census_2008/Main Report/Statistical tables/Population Size and Composition.xls

Table 4: Method 2 estimates of ρ_- and ρ_+ with MDHS sample weights.

Value of k			1.0	2.0		3.0		4.0	
Bound			$\rho_- = \rho_+$	ρ_-	ρ_+	ρ_-	ρ_+	ρ_-	ρ_+
Male	North	Rural	0.0293	0.0301	0.0344	0.0309	0.0395	0.0317	0.0446
		Urban	0.1570	0.1570	0.1641	0.1570	0.1711	0.1570	0.1781
		Total	0.0478	0.0490	0.0558	0.0502	0.0638	0.0514	0.0717
	Central	Rural	0.0570	0.0589	0.0710	0.0609	0.0850	0.0628	0.0990
		Urban	0.1454	0.1554	0.1854	0.1654	0.2253	0.1754	0.2653
		Total	0.0615	0.0637	0.0769	0.0659	0.0922	0.0681	0.1075
	South	Rural	0.1321	0.1356	0.1615	0.1392	0.1909	0.1427	0.2204
		Urban	0.1907	0.2059	0.2444	0.2210	0.2981	0.2362	0.3518
		Total	0.1398	0.1445	0.1720	0.1493	0.2042	0.1540	0.2365
	Total	Rural	0.0893	0.0919	0.1091	0.0945	0.1289	0.0971	0.1487
		Urban	0.1787	0.1910	0.2239	0.2032	0.2690	0.2155	0.3142
		Total	0.0986	0.1019	0.1208	0.1051	0.1431	0.1083	0.1653
Female	North	Rural	0.0760	0.0773	0.0889	0.0786	0.1019	0.0799	0.1148
		Urban	0.2638	0.2725	0.3008	0.2812	0.3379	0.2900	0.3749
		Total	0.1148	0.1172	0.1345	0.1196	0.1542	0.1219	0.1739
	Central	Rural	0.0926	0.0950	0.1148	0.0975	0.1370	0.0999	0.1591
		Urban	0.1729	0.1729	0.1942	0.1729	0.2155	0.1729	0.2368
		Total	0.0965	0.0989	0.1188	0.1013	0.1412	0.1037	0.1636
	South	Rural	0.2011	0.2072	0.2440	0.2134	0.2870	0.2195	0.3300
		Urban	0.2381	0.2472	0.2977	0.2564	0.3572	0.2655	0.4168
		Total	0.2049	0.2113	0.2495	0.2178	0.2940	0.2242	0.3386
	Total	Rural	0.1473	0.1513	0.1790	0.1553	0.2108	0.1593	0.2426
		Urban	0.2357	0.2429	0.2836	0.2501	0.3315	0.2572	0.3794
		Total	0.1563	0.1606	0.1898	0.1649	0.2233	0.1692	0.2569

Table 5: Method 1 estimates using the census estimate of urban and rural population shares.

		Male		Female		Total	
		$\hat{\rho}_-$	$\hat{\rho}_+$	$\hat{\rho}_-$	$\hat{\rho}_+$	$\hat{\rho}_-$	$\hat{\rho}_+$
North	Rural	0.0402	0.0407	0.0745	0.1095	0.0579	0.0762
	Urban	0.1302	0.1303	0.2006	0.2258	0.1656	0.1782
	Total	0.0617	0.0621	0.1074	0.1399	0.0852	0.1021
Central	Rural	0.0676	0.0684	0.0972	0.1428	0.0827	0.1065
	Urban	0.1018	0.1024	0.1340	0.1610	0.1176	0.1312
	Total	0.0692	0.0701	0.0988	0.1436	0.0843	0.1075
South	Rural	0.1155	0.1165	0.1787	0.2214	0.1486	0.1714
	Urban	0.1572	0.1581	0.2354	0.2790	0.1958	0.2178
	Total	0.1228	0.1237	0.1861	0.2289	0.1556	0.1783
Total	Rural	0.0877	0.0885	0.1369	0.1797	0.1131	0.1357
	Urban	0.1440	0.1447	0.2126	0.2483	0.1779	0.1958
	Total	0.0956	0.0965	0.1462	0.1881	0.1216	0.1435

Table 6: Complete-case estimates.

		Male	Female	Total
North	Rural	0.0333	0.0646	0.0495
	Urban	0.1313	0.2101	0.1708
	Total	0.0596	0.1036	0.0823
Central	Rural	0.0651	0.0901	0.0779
	Urban	0.1081	0.1395	0.1235
	Total	0.0673	0.0926	0.0802
South	Rural	0.1292	0.1941	0.1631
	Urban	0.1921	0.2667	0.2289
	Total	0.1395	0.2032	0.1725
Total	Rural	0.0914	0.1408	0.1170
	Urban	0.1629	0.2310	0.1965
	Total	0.1017	0.1521	0.1276