The Effect of Non-Response on Population-Based HIV Prevalence Estimates: The Case of Rural Malawi

Francis Obare, Population Studies Center, University of Pennsylvania[1]

Introduction

> *"All population-based surveys, including local studies, have non-response bias, and this may affect the validity of the HIV prevalence estimates…The main challenges are to obtain representative sampling of all […] people, sound testing procedures and good response rates" (WHO and UNAIDS 2003: 6 & 9).*

Population-based surveys that include the collection of biological and clinical data, *biomarkers*, are increasingly being conducted in developing countries to obtain population-based health statistics. Various factors have contributed to this development. For instance, advances in medical technology that offer simple, rapid and relatively inexpensive test devices have made it cheaper to conduct population-based biomarker studies (Giles et al. 1999; Boerma et al. 2001; WHO and UNAIDS 2003). In addition, the health facility data have been inadequate for assessing the health status of the population owing to low utilization of health services (Fisher et al. 1996; Boerma et al 2001). Lastly, the HIV pandemic has increased the need for prevalence data that are unbiased and representative of the whole population to permit a more detailed evaluation of the magnitude and distribution of the disease (Fylkesnes et al. 1998; WHO and UNAIDS 2003). In contrast, previous country estimates of HIV prevalence have been derived from pregnant women attending selected antenatal clinics (ANC), which are then extrapolated to the entire population. But this method may be biased by the selective location of the clinics that carry out testing, self-selection by those who attend the clinics, and the algorithm used to extrapolate the data to the national adult population (Boerma et al.

[1] 3718 Locust Walk, 239 McNeil, Philadelphia, PA 19104; Tel: 215-573-2621; Fax: 215-898-2124 E-mail: fonyango@pop.upenn.edu.

2003; WHO and UNAIDS 2003; Allen 2006). I return to these issues in more detail in the Data section.

The validity of estimates derived from population-based biomarker surveys, however, must also be scrutinized. There are challenges posed by non-response, for example due to refusal or absence, but also in longitudinal studies, due to death and out-migration, that can significantly bias downwards the estimates. Other challenges include logistical difficulties of testing in the field rather than a clinic setting, the cost associated with conducting a survey, ethical issues, and sample representativeness (Fisher et al. 1996; Boerma et al. 2001; WHO and UNAIDS 2003)[2]. The logistical challenges include proper methods of collecting specimen that ensure the safety of the survey teams and the laboratory personnel, maintaining the cold chain[3] from the field to the laboratory, putting in place effective testing procedures, and correct disposal of specimens and waste once the testing is done (Boerma et al. 2001; Orroth et al. 2003). The success of the surveys also requires substantial investments in terms of time and money and in some circumstances may require governmental support (Allen 2006), factors that have additional logistical implications. Ethical issues include obtaining informed consent from the study participants, ensuring confidentiality of all those tested, communicating the results to the study participants, and facilitating the provision of treatment or resources (e.g. transport costs to the nearest treatment facility) to those in need of treatment (WHO and UNAIDS 2003). The sampling procedure used may affect the geographical coverage,

[2] Some of these issues may be relevant to the ANC setting as well, with the exception of obtaining informed consent, since this is usually not required.
[3] The cold chain is the process of maintaining the proper specimen temperature from the field to the laboratory.

hence the sample representativeness. In the following analysis, I will focus on the issue of non-response.

But do these various potential influences on the validity of prevalence data result in systematic differences between estimates derived from population-based studies and those extrapolated from ANC surveillance data? Although there are only a few large-scale population-based biomarker studies, so far the answer is yes: typically the former produce *lower* estimates of HIV prevalence than the latter. This is shown in Table 1 which compares HIV prevalence from antenatal clinics with those from the Demographic and Health Surveys (DHS) for selected sub-Saharan African countries. The response rates for HIV testing and for survey interviews, respectively defined in terms of the percent of eligible persons who were tested for HIV or interviewed, are also given. How do we account for the systematically lower estimates from population-based studies? On the one hand, issues related to the location of clinics, the segment of the population involved, and the algorithm used for extrapolation may account for the higher HIV prevalence observed in the ANC data[4]. On the other hand, selectivity arising from non-response could account for the lower HIV prevalence rates observed in the population-based surveys (Fylkesnes et al. 1998; Glynn et al. 2001; and Saphonn et al. 2002, for example, have acknowledged this possibility; see also Boerma et al. 2003; WHO and UNAIDS 2003). This could be the case, for instance, if survey respondents who thought they were HIV positive were disproportionately likely to be non-respondents.

<Table 1 about here>

[4] This problem has, however, been recognized in the literature and methods of adjustment have been proposed to extrapolate ANC data to the general population (see for instance Changalucha et al. 2002; Fabiani et al. 2003; Gregson et al. 2002; Zaba et al. 2000; also Orroth et al. 2003 for sexually transmitted disease prevalence).

In 2004, the Malawi Diffusion and Ideational Change Project (MDICP)- a longitudinal study in three rural sites in Malawi- provided the opportunity for survey respondents to be tested for HIV, as well as non-HIV sexually transmitted infections (STIs). Consistent with the DHS studies, HIV prevalence was found to be substantially lower in the MDICP sample in two of the project sites than the estimates provided by the Malawi National AIDS Commission. Because non-response is important in general, and likely to be particularly important in a longitudinal study due to attrition by death and out-migration, my objective is therefore to examine the effect of non-response to HIV tests on prevalence estimates derived from this study. First, I explore whether and how non-response due to refusal, temporary absence, death or out-migration biased the MDICP HIV prevalence estimates. Second, I compare the MDICP estimates with those from the ANCs designated by the Malawi National AIDS Commission to represent the rural areas of each district with and without taking non-response into account. Finally, I use a probit sample selection model to examine if selection due to non-response exerted a significant downward bias in the MDICP prevalence rates.

An important question at this point is why such an exercise is necessary. In the first place, HIV/AIDS, just like any other public health problem, requires reliable statistics to monitor the progress of the disease, to plan HIV prevention programmes, and to assess the impact of interventions (Fisher et al. 1996; Zaba et al. 2000; Glynn et al. 2001; cf Changalucha et al. 2002). At the very least, this can partly explain the amount of effort that has been invested in developing adjustments for HIV prevalence rates obtained from ANC data. This also leads to the second justification for this paper, that is, methodological relevance. In particular, exploring whether non-response is a significant source of bias in the population-based HIV prevalence estimates could help determine if such estimates also need adjustments. Alternatively, it could help determine if adjustments are necessary in future survey designs to ensure higher response rates.

The third justification for examining non-response is what I refer to as “the power of statistics”. Statistics have considerable rhetorical power in making arguments. We use statistics to tell the broader story of the subject matter, to generate interest in the subject, and to provide evidence for debate on the issue at hand (Miller 2004). Thus, when the same national agency or two different national agencies come up with discrepant HIV prevalence rates, this is likely to generate intense debate. For example, anecdotal evidence in the case of Kenya indicates that after the release of the DHS results, the government was put under extreme pressure by the civil society and politicians to explain the discrepant HIV prevalence rates.

Background of the study

Malawi is divided into three administrative regions (North, South and Central). Within these regions, there are a total of 27 districts (National Statistics Office and ORC Macro 2001). With a population close to 12 million by mid-2004, about 86 percent lived in the rural areas (Population Reference Bureau 2004). The national adult HIV prevalence rate in 2003 was estimated at 14 percent, based on data from pregnant women attending antenatal clinics (Republic of Malawi 2003). Prevalence was higher in the urban and semi-urban areas than in the rural areas (about 22 percent and 21 percent versus 15 percent in the rural areas) and in the South (about 24 percent) than in the North (20 percent) or Central (16 percent) (Republic of Malawi 2003). Attendance at antenatal clinics was high both in the rural and urban areas: 92 percent of women in rural areas who had a live birth within five years before the 2000 DHS survey reported attending an antenatal clinic (National Statistics Office and ORC Macro 2001). However, the 2003 ANC prevalence data come from only 19 clinics that constituted the sentinel surveillance system. Of the five sentinel sites in the North, two were designated as rural, another two as semi-urban, and the remainder as urban. In each of the remaining two regions, three sites were designated as rural, another three as semi-urban, and one as urban (Republic of Malawi 2003). Thus, estimates of rural prevalence for 86 percent of the population are based on only 8 antenatal clinics.

Although population-based biomarker surveys may test a higher proportion of the population than are tested in ANCs, non-response may nonetheless lead to a significant bias in the estimates. In particular, if those who are absent at the time of the population-based study, or those who refuse to be tested, are more likely to be HIV positive, prevalence is likely to be biased downward, and thus lower than at ANC clinics. There are general reasons for survey non-response, as well as particular reasons when a survey is also testing for HIV. According to the framework by Groves and Couper (1998), participation in a household survey is influenced by the social environment, the household characteristics, the survey design, and interviewer characteristics. More relevant for this study is the finding of Boerma et al. (2003) that the process of obtaining consent was partly responsible for the high refusal rate in a population-based HIV study in South Africa.

The DHS data for a number of sub-Saharan African countries show that refusal as a source of non-response was greater than absence at the time of the survey and testing (see Table 2). Significantly more men than women refused the HIV tests in three

(Burkina Faso, Ghana and Tanzania) of the six countries for which data are available. Similarly, refusal rates were significantly higher in the urban than in the rural areas except in Zambia. The same results hold for temporary absence: higher among men than women, and higher in the urban than rural areas.

<Table 2 about here>

There are suggestions that refusal may be associated with higher risk of HIV infection (e.g. by Boerma et al. 2003; WHO and UNAIDS 2003). Respondents know their own sexual history, their health status, and that HIV is sexually transmitted; thus, those who are HIV positive are likely to perceive that their risk of being positive is high and prefer not to learn that they are correct. Absence is also likely to be associated with increased risk of HIV infection (e.g. Crampin et al. 2003). In the rural areas, those who are absent are more likely to be mobile (for example, labor migrants, job-seekers, traders and business people); in the urban areas absence can be attributed to wage labor or job-seeking. On the other hand, those of lower socio-economic status have been found to cooperate more in surveys (Groves and Couper 1998), which might explain higher refusal rates in urban compared to rural areas given that most urban areas in much of the developing world tend to be better off economically than rural areas.

Finally, in a longitudinal study such as the MDICP, which interviewed the same rural sample in 1998, 2001 and 2004, attrition through death and out-migration between survey waves are additional sources of non-response. Death is, of course, more likely to be associated with those at the highest risk of HIV infection. The implications of out-migration are likely to be similar to those of temporary absence discussed above because it is also an aspect of mobility. Thus, to the extent that non-respondents during testing constituted the highest risk group, the observed HIV prevalence in the MDICP study may be biased downward, hence the need to examine the extent of such bias.

Data

This paper is based on the MDICP data, a longitudinal study in rural Malawi that is part of the Social Networks Project of the Population Studies Center, University of Pennsylvania. Its general aim is to examine the role of social networks in changing attitudes and behavior regarding family size, family planning and HIV/ AIDS in Malawi (see http://malawi.pop.upenn.edu for further details). It is conducted in three rural sites in three distinctive districts selected from each of the three regions in the country. These are Rumphi District in the Northern region, Mchinji District in the Central Region, and Balaka District in the Southern region (below, these will be referred to as North, Center, and South respectively). Though the sampling design was not meant to be representative of the national rural population of Malawi, the sample characteristics have been shown to closely match the characteristics of the rural population of the 1996 Malawi Demographic and Health Survey (Watkins et al. 2003).

The first and second waves of the project (MDICP-1 and MDICP-2) were carried out in 1998 and 2001 respectively and involved only survey data collection. The third wave (MDICP-3) was conducted between March and August 2004 and had two components: the survey component and the STI component. Besides re-interviewing respondents from previous rounds, MDICP-3 also included new husbands to women interviewed in previous rounds and a sample of adolescents aged 15-24 years. Except for the new husbands and adolescents, eligibility for HIV test for adults in the present study is defined in terms of at least one successful interview in any of the previous rounds.

Based on this criterion, 4,075 respondents were eligible for HIV test in 2004. Out of these, 3,291 respondents (about 81 percent) were successfully contacted for the HIV test and 2,988 respondents were tested. This represents 73 percent of all eligible respondents and about 91 percent of those successfully contacted. Distribution of eligible respondents by site shows that 36 percent were from the South, 30 percent from the Center, and 34 percent from the North. Female respondents make up slightly more than half of the analysis sample (about 53 percent) while adolescents, defined in the present study as those aged 15-19 years, constitute about 16 percent.

The STI component involved the collection of biomarkers for HIV and three treatable STIs (Chlamydia, gonorrhea and trichomoniasis for females, Chlamydia and gonorrhea for males). A team of trained nurses was responsible for collecting the specimens. They usually visited respondents about two to three days after the visit by the survey team[5]. The process of collecting specimens involved pre-test counseling of respondents, administering of a brief questionnaire, obtaining the respondent's informed consent, and if consent was granted, collecting specimens[6]. Respondents could grant an interview for the questionnaire but refuse to give specimens for either or both tests. For HIV testing, saliva samples were used. For STI testing, urine samples were collected from male respondents and self-administered vaginal swabs from females.

[5] An exception was the North where, for logistical reasons, about half of the respondents were visited by the nurses' team before being interviewed by the survey team.

[6] The specimens were then refrigerated over a night or two before being transferred to the laboratory run by the University of North Carolina (UNC) at Chapel Hill based in Lilongwe Central Hospital where the analysis was done. Linked to the process of specimen collection was a randomized experiment focusing on incentives for voluntary counseling and testing (VCT) uptake. Conditional on accepting to give specimens, the respondent was given an opportunity to randomly choose an amount of money written on bottle tops put in plastic bags. The amount picked was then recorded on a voucher which the respondent was to present when he/ she came for his/ her STI or HIV test results. This was followed by post-test counseling of the clients, whether infected or not, so as to avoid any suspicion that only those who were found to be infected were being given post-test counseling (see Thornton 2005 for more details).

In what follows, I compare HIV prevalence estimates from the MDICP data with the 2003 testing of pregnant women in ANC surveillance sites in the rural areas of each of the three districts covered by the MDICP. It is thus necessary to point out that not only are there several sources of bias in the MDICP data, as discussed above, but also to reiterate that the ANC estimates may be biased due to selective location of the clinics, self-selection by ANC attendees, and the algorithm used to extrapolate the data to the adult population. Selective location of the clinics arises from the disproportionate location of the ANCs that carry out testing in urban and peri-urban areas which have high HIV prevalence compared to rural areas. Self-selection by those who attend the clinics is attributable to the fact that only pregnant women who make their first-time visits to the ANCs are involved. They may represent a select group of the sexually active population, especially younger and fecund women. Moreover, the algorithm for extrapolating from pregnant women to the general population involves fitting epidemic curves to the sentinel data based on various assumptions about the intercensal population growth rates, the sex ratio, and the relationship between ANC and population prevalence[7]. In the case of Malawi, the curve fitting was at times done manually when the curves failed to fit the data (Republic of Malawi 2004).

Furthermore, there was a deliberate attempt in 2003 to increase the rural sample size for the ANC data but almost half of the women sampled (49 per cent) were from semi-urban sites (Republic of Malawi 2003). This might have led to the inclusion in the rural sample of more semi-urban women, particularly in the North, which had fewer sites due to a smaller population size relative to the other regions (Republic of Malawi 2003).

[7] The assumption is that prevalence among pregnant women attending the clinics is similar to that among all adults in the general population.

In addition to the factors discussed above, these could be potential sources of upward bias in the rural prevalence from the ANC data.

Methods

I use two approaches to estimate the extent to which selective non-response, including refusal, biased the MDICP estimates of HIV prevalence. First, I conduct a sensitivity analysis, in which I make various assumptions about HIV prevalence among MDICP sample members who were not tested. In essence, I am asking the following questions: first, would the MDICP prevalence have been significantly different from what was observed if we had tested the non-respondents and found that they had the *same* prevalence as that observed for the adjacent rural ANC attendees? Second, how would the MDICP prevalence rate obtained through this assumption be different from that estimated by the ANC surveillance system? If we would have to assume an implausibly high prevalence among non-respondents, perhaps because they think they are infected but do not want to know, it would be unlikely that the MDICP estimates are biased.

For the first approach, I begin by estimating a new MDICP prevalence rate, based on the assumption that non-respondents have the same HIV prevalence as the pregnant women tested at the antenatal clinics that represent rural areas of three districts in which the MDICP study was conducted[8]. Subsequently, a one-sample test of proportion is performed to determine the associated probability of observing this new rate among those

[8] For ease of presentation, a ratio of the assumed prevalence among non-respondents to the observed prevalence among those who were tested by the MDICP is defined in terms of relative risk for HIV infection. This ranges from 0.8 in the Center where the MDICP estimates are slightly higher than the ANC rate to 3.4 for men in the North. In between, we have relative risks of 1.8 for women and 2.0 for men in the South, and 2.5 for women in the North. Since the published ANC rates are for women, I obtain the ANC prevalence for men by assuming a female-to-male prevalence ratio of 1.2 to 1. This is the ratio that UNAIDS uses in HIV/AIDS projections for generalized epidemics that have been on for more than ten years.

who were tested. The expectation is that if non-response is a source of bias, we should expect to see significant differences between the observed and the estimated HIV prevalence rates at any level of assumed prevalence among non-respondents.

The main sources of non-response for HIV tests were refusal, temporary absence, death, out-migration, and a final category of 'other', which included outcomes like 'too sick/ hospitalized' and 'divorced/ widowed'. The analysis is done sequentially: first, I assume that refusal was the only source of non-response; subsequently, temporary absence, death, out-migration, and 'other' are included in that order. Separate analyses are done for each site and for males and females. I also compare the estimated HIV prevalence that takes into account non-response and the prevalence among women attending antenatal clinic in adjacent rural areas of the three districts. A two-sample test of proportion is used in these comparisons[9]. The purpose is to determine if the population-based estimates would be similar to the ANC estimates had the project tested every eligible respondent and found the assumed prevalence rate among non-respondents.

The second part of the analysis involves estimating a probit regression model with sample selection (Van de Ven and Van Praag 1981) to determine if selection was a significant source of bias in the MDICP HIV prevalence rates. The underlying model is of the following form:

$$Y^{*}{}_{i} = \beta X_{i} + \varepsilon_{1i} \tag{1}$$

Equation (1) is the latent equation where $Y^{*}{}_{i}$ is the unobserved HIV status of individual i, β is a (k × 1) vector of unknown parameters, X_i is a (k × 1) vector of exogenous variables

[9] I assume that the variance, and hence the standard deviation, for male ANC HIV prevalence rate is the same as that for females for purposes of testing for the significance of differences between MDICP and ANC male HIV prevalence rates.

associated with HIV status, and the disturbance term, ε_{1i}, is assumed to be normally distributed with a mean of 0 and a standard deviation of 1 i.e. $\varepsilon_{1i} \sim N(0, 1)$. However, HIV status is observed for only those who agreed to participate in the testing so that:

$$Y_i = Y^*_i \text{ if } P^*_i > 0 \qquad (2)$$

where Y_i is the observed HIV status, and P^*_i is the unobserved propensity to participate in testing. If non-respondents were more likely to be HIV positive, estimates of β in equation (1) based on the sub-sample that was tested would be inconsistent.

The propensity to participate in the study, P^*_i, in equation (2) can be modeled separately and is commonly estimated by means of a binary regression model (Van de Ven and Van Praag 1981; Winship and Mare 1992) of the form:

$$P^*_i = \alpha Z_i + \varepsilon_{2i} \qquad (3)$$

where α is a (k × 1) vector of unknown parameters, Z_i is a (k × 1) vector of exogenous variables associated with participation in HIV testing, and the disturbance term, ε_{2i}, is also assumed to be normally distributed with a mean of 0 and a standard deviation of 1 i.e. $\varepsilon_{2i} \sim N(0, 1)$. The inconsistency of the estimates of β based on the sub-sample that was tested for HIV arises when the correlation, ρ, between ε_{1i} (in equation [1]) and ε_{2i} (in equation [3]) is not equal to zero. This is analogous to the omitted variable bias (Heckman 1979; Winship and Mare 1992) in which the conditional mean of ε_{1i} given X_i and $P^*_i > 0$ is omitted from the regression. We correct for this potential bias by introducing this conditional mean in the regression equation to obtain:

$$Y_i = \beta X_i + E[\varepsilon_{1i} \mid X_i, P^*_i > 0] + \varepsilon_{1i}$$

$$= \beta X_i + \rho\lambda_i + \varepsilon_{1i} \qquad (4)$$

where $\rho = \text{corr}(\varepsilon_{1i}, \varepsilon_{2i})$ and $\rho\lambda_i = E[\varepsilon_{1i}|X_i, P^*_i > 0]$. Since Y_i is dichotomous, we estimate equation (4) by means of a probit selection model. If selection is a significant source of bias, we should expect ρ to be significantly different from zero.

Two probit selection models predicting the probability of being HIV positive conditional on participation in testing are estimated. The first model is based on significant predictors of status and participation determined through ordinary probit models. This means that variables can enter into both the status and participation functions. The second model includes in the status function only the variables that are hypothesized to be related to HIV status but not both status and participation. Likewise only variables that are hypothesized to be related to participation but not both participation and status are included in the participation function. The purpose is to explore whether the two approaches lead to different conclusions. This is because the original Heckman (1979) sample selection model has been shown to be sensitive to the choice of variables included the selection function (Briggs 2004).

The potential predictors of HIV status include: *age*, *sex*, *site*; whether the respondent *had stayed outside the district for six months or more since age 15*; if (ever) married, *the number of times the respondent had been married*, or *whether the spouse usually stayed outside the village*; and for the sexually active, whether they had ever *used or were using abstinence or condoms*. Age and sex are defined as dichotomies i.e. adolescents (aged 15-19 years) versus adults, and males versus females respectively. Site refers to the three MDICP study sites of the North, Center and South. Whether the respondent had stayed outside the district for six months or more since age 15 and whether the spouse (for the married sample) usually stayed outside the village are indicators of mobility which has been associated with increased risk of HIV infection (e.g. by Crampin et al. 2003). The number of unions increases the chances of HIV infection while the use of condoms or abstinence reduces those chances. The potential predictors of participation in the test include *age*, *sex*, *site*, *education level*, *household size*, *level of worry about HIV infection*, *previous test*, whether the respondent *had stayed outside the district for six months or more since age 15*; and for the married sample: *whether the spouse usually stayed outside the village*, or *the respondent suspected the spouse of infidelity*. Characteristics for non-respondents are taken as at the time at which they were last interviewed i.e. in 1998 or 2001.

Results

HIV prevalence rates

Table 3 gives the MDICP and ANC HIV prevalence rates in the three districts by selected background characteristics. Prevalence was highest in the South (8.4%) and lowest in the North (4.8%). The regional pattern slightly differs from that observed from antenatal clinic data where prevalence was also highest in the South and but lowest in the Center. Sex differences, however, reflect the patterns observed from ANC data i.e. prevalence was higher among females (7.9%) than among males (5.6%) ($p<0.01$). There is no significant difference in prevalence between the South and the Center. But the differences between the North and the South, and between the North and the Center are statistically significant ($p<0.01$ and $p<0.05$ respectively).

<Table 3 about here>

Prevalence rates from the adjacent rural ANC sites are given in the lower panel of Table 3. The MDICP prevalence rates for the North and the South are significantly lower ($p<0.001$) than the ANC rates, a pattern that is observed even if we compare prevalence rates among females alone. In the Center, the MDICP prevalence rate is slightly higher than the ANC rate but the difference is not statistically significant.

Response and non-response rates

I begin with basic description, in order to see if there is any reason to expect that the patterns and the influence of non-response may differ by sex or across the three sites. Table 4 shows the response and non-response rates, based on all eligible respondents, by site and sex. While it appears that a slightly higher proportion of eligible respondents were tested in the South and Center than in the North, these differences are not statistically significant. The refusal rate was however significantly higher in the South and Center ($p<0.05$ in both cases) than in the North. There was, however, no significant difference in refusal rates between men and women.

<Table 4 about here>

Out-migration was the major source of non-response in the North and the South. It was more of a problem in the North than in the other sites. It was also significantly higher among women than among men in the North, but the opposite was the case in the South and Center, although these differences (between men and women in the South and Center) are not statistically significant.

Effect of non-response on prevalence estimates

The purpose of this section is to answer some ‘what if’ questions. For instance, what would the MDICP prevalence be if the percentage of non-respondents HIV positive was the same as the observed rate among ANC attendees? And would this rate be significantly different from what was actually observed? Table 5 shows the results of this exercise by site and by sex. Panel A of Table 5 gives the estimated HIV prevalence rates that would be obtained assuming that *refusal* was the only source of non-response and that those who refused had the same prevalence as that observed among ANC attendees. As it turns out, the recalculated prevalence rates based on these assumptions are not significantly different from the observed rates among those who were tested (both men and women).

<Table 5 about here>

Panel B of Table 5 considers the situation in which both *refusal* and *temporary absence* were the only sources of non-response. The recalculated HIV prevalence rates under this assumption for various levels of relative risk for HIV infection among non-respondents lead to qualitatively similar conclusions as for *refusal* alone. However, *death* does make a difference: when we include *death* as a source of non-response, there are significant differences between the recalculated and the observed MDICP rates (Panel C of Table 5). This is true for women in the South and Center but only when we assume a relative risk of 3.4, which corresponds to prevalence rates for non-respondents that are higher than the ANC rates observed in the adjacent sites.

Out-migration is added as an additional source of non-response in the results shown in Panel D of Table 5. For the North, we find some significant differences between the recalculated and the observed prevalence rates, for both men and women, when non-respondents are assumed to have the same prevalence rate as that observed in the adjacent ANC site. The observed rate for the MDICP male sample is under-estimated by about 2.1 percentage points and that for the female sample by about 2.3 percentage points. For the South and Center, significant differences are obtained by assuming prevalence rates among non-respondents that are higher than those observed in the adjacent ANC sites. At least for these two sites, the recalculated prevalence rates would not be significantly different from the observed rates even if we assume that non-respondents had the same prevalence as that observed in the adjacent ANC sites.

Comparison with the ANC prevalence rates

In this section, I consider all sources of non-response, including the 'other' category, and examine how different the recalculated rates are from the ANC rates for each site. As with the preceding analyses, the results given in Table 6 confirm that we would need to assume that all non-respondents (both men and women and in all sites) had a higher risk of HIV infection than that observed in the ANC sites to obtain rates that do not differ significantly from the ANC rates. I also made comparisons between the recalculated and the observed MDICP rates when all sources of non-response are considered (not shown). Such comparisons yielded similar results to those in the last panel of Table 5.

<Table 6 about here>

Because the ANC data are obtained from pregnant women, I further compared the prevalence rates among MDICP female respondents who reported that they were pregnant at the time of the study with the ANC rates without taking non-response into account. The assumption is that pregnant women in the MDICP study were likely to attend the clinics though the survey did not ask about it. Figure 1 shows the results of this comparison by site. Whereas the figures for the South and Center are very similar, in the North they are not. This could be a further indication that the ANC sample for the North might have included a significant number of women from peri-urban sites thereby resulting in substantial upward bias in prevalence.

<Figure 1 about here>

HIV status, testing and selection bias

The results from the probit sample selection models are given in Table 7. The first model is based on the significant predictors of status and participation respectively. The second model includes only those factors that are hypothesized to be associated with either status or participation but not both. In this case, *age*, *sex*, and *site* were included only in the status function as there were no significant differences in participation rates based on these factors. The results show that adolescents (aged 15-19 years, married and unmarried) were significantly less likely ($p<0.01$) to be HIV positive compared to adults. Model 1 shows that women were significantly more likely ($p=0.045$) to be HIV positive than men but this ceases to be the case in Model 2. Those whose partners usually stayed outside the village were also significantly more likely ($p<0.01$) to be HIV positive than those whose partners usually resided in the village[10]. Similarly, those who had been married multiple times were significantly more likely ($p<0.01$) to be HIV positive than those who had been married just once.

[10] In the preliminary probit models estimated to determine significant factors for inclusion in Model 1, those who had stayed outside the district for six months and more since age 15 were significantly more likely to be HIV positive ($\beta=0.166$; standard error=0.078; $p<0.05$) than those who had not. There was also significant difference in the likelihood of infection between the North and the South. But these variables were omitted from the status function in Model 1 because overall tests showed that they were statistically insignificant. Education level was also not significantly associated with status.

<Table 7 about here>

On the other hand, education level and having been previously tested are significantly associated with participation in testing. Compared to those with no education, persons with some education (primary, and secondary and above) were significantly *less* likely ($p<0.01$) to participate in the test, whereas those who had previously been tested were significantly *more* likely ($p<0.01$) to participate. The significant associations between participation in testing and household size, worry about catching AIDS, and staying outside the district for six months or more since age 15 seem to be largely driven by those responding “don’t know”, “can’t remember” or “missing”. Similarly, the significance of suspicion of the spouse’s infidelity appears to be largely a function of other differences than that between those who knew or suspected their spouses and those who did not know or did not suspect their spouses.

The two models give correlations of $\rho = 0.09$ with $p=0.559$ (Model 1) and $\rho = 0.05$ with $p=0.745$ (Model 2). The results lead to two conclusions. First, the two model specifications show that selection did *not* exert a significant downward bias in the MDICP HIV prevalence estimates. Second, the results are stable across the models and therefore not sensitive to model specification. This is also confirmed by Figures 2 and 3, which compare the observed HIV prevalence with the predicted prevalence for men (Figures 2a and 3a) and women (Figures 2b and 3b) from the two models. In all the sites and for both men and women, the predicted prevalence from the two models is not significantly different from the observed prevalence.

<Figures 2a, 2b, 3a and 3b about here>

Discussion and conclusion

In this paper, I examined the extent of bias in population-based HIV prevalence rates due to non-response using data from the Malawi Diffusion and Ideational Change Project (MDICP) collected in rural Malawi in 2004. The data show that prevalence in two of the three sites was significantly lower than that from the antenatal clinics designated by the National AIDS Commission to represent rural areas for the districts. The lower prevalence in population-based studies has been observed elsewhere. Non-response is an obvious potential source of bias in population-based testing, and would significantly bias the observed prevalence downwards if non-respondents were at higher risks of HIV infection than those who participated in testing. It is thus not surprising that even as country estimates of HIV prevalence are being adjusted to reflect rates obtained from population-based studies (see for instance UNAIDS 2004), there is also a call for caution that these rates may not be the gold standard owing to low response rates (e.g. by Boerma et al. 2003; WHO and UNAIDS 2003).

Out-migration was the major source of non-response in the present study, a pattern that is likely due to the longitudinal nature of the study. The percent of eligible respondents who had moved out of the study sites was highest in the North. The seasonality of employment in tobacco plantations in the North might provide one potential explanation for this. One might argue that people would be more likely to move during low season when tobacco is already harvested and there is not much work. But the gender pattern in out-migration observed in the North (with more women than men having moved) coupled with the fact that tobacco-growing is mainly a male-dominated task implies that the differential marriage patterns might be part of the explanation. The

North is largely a patrilineal and patrilocal society, the South is mostly matrilineal and matrilocal, and the Center is characterized by mixed marriage patterns (Zulu and Chepngeno 2003). Studies have shown that one of the strategies that rural Malawian women are using to avoid exposure to HIV/AIDS is divorce (Watkins 2004; Reniers 2005). The differential marriage patterns imply that in the South, men are more likely to move in case of marriage or divorce while in the North, it is the women (see for instance, Reniers 2003).

A simple sensitivity analysis showed that in one district, the North, there seems to be bias from assuming the ANC rate for non-respondents, but in the other districts, there does not seem to be a bias. Why might the North be different? First, there could be a genuine downward bias in the observed MDICP HIV prevalence rate in the North. But a comparison of the observed prevalence with the predicted prevalence from the sample selection models shows that there is no significant difference between the two rates. This leads to the second and more plausible explanation, i.e. that it could be due to differences in the population characteristics of MDICP and ANC sites. As noted before, a deliberate attempt in 2003 to increase the rural sample size for the ANC data ended up including almost half of the women (49 percent) from peri-urban sites (Republic of Malawi 2003). Since the North had fewer sites due to a smaller population size compared to the other regions, the rural sample might have included the most women from peri-urban sites. Assuming a peri-urban prevalence for non-respondents in a rural setting is likely to bias the results.

A comparison between the MDICP and ANC HIV prevalence rates taking non-response into account shows that, for the South and the North, we need to assume substantially higher prevalence among non-respondents than that observed in the antenatal clinics to obtain comparable rates. The assumption that non-respondents had the same prevalence as that observed in the antenatal clinics otherwise gives significantly lower estimates than the ANC rates. The implication is that while the MDICP rates may understate the true HIV prevalence in these two sites, it is most likely that the ANC rates significantly overstate it. This is consistent with other studies which have suggested that population-based studies may capture the rural HIV prevalence better than ANC data owing to the location of fewer clinics in rural areas (e.g. Boerma et al 2003).

In conclusion, the most important result of this study is that selection due to non-response does not appear to exert a significant downward bias in the population-based data used here. A similar finding has been noted for Kenya by Bignami-Van Assche et al. (2005). Perhaps this finding does confirm the suggestion that population-based surveys provide better quality HIV prevalence data for rural populations than do ANC data.

Acknowledgements

The data used in this study was collected through NIH/NICHD grants RO1-HD372-276, RO1-HD41713, and RO1 HD044228-01. The study benefited from valuable comments from Susan Watkins, Hans-Peter Kohler, Herb Smith, Georges Reniers and members of the 2003 cohort of the Demography program at the University of Pennsylvania.

References

Allen, Tim. 2006. “AIDS and Evidence: Interrogating some Ugandan Myths.” *Journal of Biosocial Science* 38: 7-28.

Bignami-Van Assche, Simona, Joshua A. Salomon, and Christopher J.L. Murray. 2005. *Evidence from National Population-Based Surveys on Bias in Antenatal Clinic-Based Estimates of HIV Prevalence*. Paper presented at the 2005 Meeting of the Population Association of America, Philadelphia, March 31-April 2.

Boerma, J. Ties, Peter D. Ghys and Neff Walker. 2003. “Estimates of HIV-1 prevalence from national population-based survey as a new gold standard.” *The Lancet* 362:1929-1931.

Boerma, J. Ties, Elizabeth Holt and Robert Black. 2001. “Measurement of Biomarkers in Surveys in Developing Countries: Opportunities and Problems.” *Population and Development Review* 27(2): 303-314.

Briggs, Derek C. 2004. “Causal Inference and the Heckman Model.” *Journal of Educational and Behavioral Statistics* 29(4): 397-420.

Cellule de Planification et de Statistique du Ministère de la Santé (CPS/MS), Direction Nationale de la Statistique et de IˈInformatique (DNSI) et ORC Macro. 2002. *Enquête Démographique et de Santé au Mali 2001*. Calverton, Maryland, USA: CPS/MS, DNSI et ORC Macro.

Central Bureau of Statistics (CBS) [Kenya], Ministry of Health (MOH) [Kenya], and ORC Macro. 2004. *Kenya Demographic and Health Survey 2003*. Calverton, Maryland: CBS, MOH, and ORC Macro.

Central Statistical Office [Zambia], Central Board of Health [Zambia], and ORC Macro. 2003. *Zambia Demographic and Health Survey 2001-2002*. Calverton, Maryland, USA: Central Statistical Office, Central Board of Health, and ORC Macro.

Changalucha, John, Heiner Grosskurth, Wambura Mwita, James Todd, David Ross, Philippe Mayaud, Abdul Mahamoud, Arnoud Klokke, Frank Mosha, Richard Hayes and David Mabey. 2002. “Comparison of HIV prevalences in community-based and antenatal clinic surveys in rural Mwanza, Tanzania.” *AIDS* 16: 661-665.

Crampin, Amelia C., Judith R. Glynn, Bagrey M.M. Ngwira, Frank D. Mwaungulu, Jörg M. Pönnighaus, David K. Warndorff and Paul E.M. Fine. 2003. “Trends and measurement of HIV prevalence in northern Malawi.” *AIDS* 17: 1817-1825.

Fabiani, Massimo, Knut Fylkesnes, Barbara Nattabi, Emingtone O. Ayella and Silvia Declich. 2003. “Evaluating two adjustment methods to extrapolate HIV prevalence from pregnant women to the general female population in sub-Saharan Africa”. *AIDS* 17:399-405.

Fisher, Gail, Gregory Pappas and Magdalena Limb. 1996. “Prospects, Problems, and Prerequisites for National Health Examination Surveys in Developing Countries.” *Social Science and Medicine* 42(12): 1639-1650.

Fylkesnes, Knut, Zacchaeus Ndhlovu, Kelvin Kasumba, Rosemary Mubanga Musonda and Moses Sichone. 1998. “Studying the dynamics of the HIV epidemic: population-based data compared with sentinel surveillance in Zambia.” *AIDS* 12: 1227-1234.

Ghana Statistical Service (GSS), Noguchi Memorial Institute for Medical Research (NMIMR), and ORC Macro. 2004. *Ghana Demographic and Health Survey 2003*. Calverton, Maryland, USA: GSS, NMIMR, and ORC Macro.

Giles, Ralph E., Keith R. Perry and John V. Parry. 1999. "Simple/Rapid Test Devices for Anti-HIV Screening: Do They Come UP to the Task?" *Journal of Medical Virology* 59: 104-109.

Glynn, Judith R., Anne Buvé, Michel Caraël, Rosemary M. Musonda, Maina Kahindo, Isaac Macauley, Francis Tembo, Léopold Zekeng and the Study Group on Heterogeneity of HIV Epidemics in African Cities. 2001. "Factors influencing the difference in HIV prevalence between antenatal clinic and general population in sub-Saharan Africa." *AIDS* 15: 1717-1725.

Gregson, Simon, Nicola Terceira, Memory Kakowa, Peter R. Mason, Roy M. Anderson, Stephen K. Chandiwana and Michel Caraël. 2002. "Study of bias in antenatal clinic HIV-1 surveillance data in a high contraceptive prevalence population in sub-Saharan Africa." *AIDS* 16: 643-652.

Groves, Robert M. and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-162.

Institut National de la Statistique et de la Démographie (INSD) et ORC Macro. 2004. *Enquête Démographique et de Santé du Burkina Faso 2003*. Calverton, Maryland, USA: INSD et ORC Macro.

Institut National de la Statistique (INS) et ORC Macro. 2004. *Enquête Démographique et de Santé du Camroun 2004*. Calverton, Maryland, USA: INS et ORC Macro.

Miller, Jane E. 2004. *The Chicago Guide to Writing About Numbers*. Chicago: The University of Chicago Press.

National AIDS Commission (NAC). 2004. *HIV/AIDS in Malawi: 2003 Estimates and Implications*. Malawi: National AIDS Commission.

National Statistics Office [Malawi] and ORC Macro. 2001. *Malawi Demographic and Health Survey 2000*. Zomba, Malawi and Calverton, Maryland, USA: National Statistics Office and ORC Macro.

Nyblade, Laura, Jane Menken, Maria J. Wawer, Nelson K. Sewankambo, David Serwadda, Frederick Makumbi, Tom Lutalo and Ron H. Gray. 2001. "Population-Based HIV Testing and Counseling in Rural Uganda: Participation and Risk

Characteristics." *Journal of Acquired Immune Deficiency Syndrome* 28(5): 463-470.

Orroth, K.K., E.L. Korenromp, R.G. White, J. Changalucha, S.J. de Vlas, R.H. Gray, P. Hughes, A. Kamali, A. Ojwiya, D. Serwadda, M.J. Wawer, R.J. Hayes and H. Grosskurth. 2003. "Comparison of STD prevalences in the Mwanza, Rakai, and Masaka trial populations: the role of selection bias and diagnostic errors." *Sexually Transmitted Infections* 79: 98-105.

Population Reference Bureau. 2004. *2004 World Population Data Sheet*. Washington, DC: Population Reference Bureau.

Reniers, Georges. 2005. *Marital Strategies for Managing Exposure to HIV in Rural Malawi*. Paper presented at the Annual Meeting of the Population Association of America, Philadelphia, March 31-April 2.

______________. 2003. "Divorce and remarriage in rural Malawi." *Demographic Research Special Collection 1*. Max-Planck Institute for Demographic Research, Rostock, Germany.

Republic of Malawi [National AIDS Commission]. 2004. *Malawi National HIV/AIDS Estimates 2003: Technical Report*.

Republic of Malawi [Ministry of Health and Population]. 2003. *HIV Sentinel Surveillance Report 2003*. National AIDS Commission.

Saphonn, Vonthanak, Leng Bun Hor, Sun Penh Ly, Samrith Chhuon, Tobi Saidel and Roger Detels. 2002. "How well do antenatal clinic (ANC) attendees represent the general population? A comparison of HIV prevalence from ANC sentinel surveillance sites with a population-based survey of women aged 15-49 in Cambodia." *International Journal of Epidemiology* 31: 449-455.

Tanzania Commission for AIDS (TACAIDS), National Bureau of Statistics (NBS), and ORC Macro. 2005. *Tanzania HIV/AIDS Indicator Survey 2003-04*. Calverton, Maryland, USA: TACAIDS, NBS, and ORC Macro.

Thornton, Rebecca. 2005. *The Demand for and Impact of Learning HIV Status: Evidence from a Field Experiment*. MIMEO Harvard.

UNAIDS 2004. *2004 Report on the Global HIV/AIDS Epidemic*. Joint United Nations Programme on HIV/AIDS (UNAIDS).

Van de Ven, Wynand P.M.M. and M.S. van Praag. 1981. "The Demand for Deductibles in Private Health Insurance: A Probit Model with Sample Selection." *Journal of Econometrics* 17: 229-252.

Watkins, Susan Cotts. 2004. “Navigating the AIDS Epidemic in Rural Malawi.” *Population and Development Review* 30(4): 673-705.

Watkins, Susan C., Eliya M. Zulu, Hans-Peter Kohler and Jere R. Behrman. 2003. “Introduction to Social Interactions and HIV/AIDS in Rural Africa.” *Demographic Research Special Collection 1*. Max-Planck Institute for Demographic Research, Rostock, Germany.

Winship, Christopher and Robert D. Mare. 1992. “Models for Sample Selection.” *Annual Review of Sociology* 18: 327-350.

World Health Organization (WHO) and UNAIDS. 2003. *Reconciling Antenatal Clinic-Based Surveillance and Population-Based Survey Estimates of HIV Prevalence in sub-Saharan Africa*.

Zaba, Basia W., Lucy M. Carpenter, J. Ties Boerma, Simon Gregson, Jessica Nakiyingi and Mark Urassa. 2000. “Adjusting ante-natal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa.” *AIDS* 14: 2741-2750.

Table 1: Comparison of antenatal clinic-based and population-based HIV prevalence with corresponding response rates (for HIV testing and survey interviews) for selected sub-Saharan African countries.

Country/ Year	ANC-based HIV prevalence	Population-based HIV prevalence	Response rate for HIV testing[a]	Response rate for survey interviews[a]
Burkina Faso (2003)	4.2%	1.8%	89%	95%
Cameroon (2003-04)	6.9	5.5	91	94
Ghana (2003)	3.1	2.2	85	95
Kenya (2003)	8.0	6.7	73	91
Mali (2001)	1.9	1.7	81	92
South Africa (2001-02)	20.1	15.6	62	74
Tanzania (2003-04)	8.8	7.0	81	94
Zambia (2001-02)	21.5	15.6	77	95

Notes: [a]Response rates appertain to the percent of eligible respondents who were tested for HIV or interviewed in the surveys; ANC- antenatal clinic.

Sources: Boerma et al. 2003; CBS, MOH, & ORC Macro 2004; Central Statistical Office, Central Board of Health, and ORC Macro 2003; CPS/MS, DNSI et ORC Macro 2002; GSS, NMIMR & ORC Macro 2004; INS et ORC Macro 2004; INSD et ORC Macro 2004; TACAIDS, NBS and ORC Macro 2005; WHO & UNAIDS 2003; UNAIDS 2004.

Table 2: Percent refusing and temporarily absent for HIV test by sex and by urban-rural residence for selected sub-Saharan African countries, DHS.

	Percent refusing HIV test						
	Sex			Urban-rural residence			
Country/Year	Males	Females	Sig. test[a]	Urban	Rural	Sig. test[a]	Total
Burkina Faso (2003)	6.6	4.4	**	13.5	2.6	**	5.4
Cameroon (2003-04)	5.6	5.4	NS	8.7	2.2	**	5.5
Ghana (2003)	10.7	5.7	**	10.6	6.3	**	8.1
Kenya (2003)	13.0	14.4	NS	17.8	11.5	**	13.7
Tanzania (2003-04)	13.9	12.3	**	19.9	10.4	**	13.0
Zambia (2001-02)	14.8	15.4	NS	15.3	15.1	NS	15.1
	Percent temporarily absent for HIV test						
Burkina Faso (2003)	4.8	1.9	**	4.9	2.6	**	3.2
Cameroon (2003-04)	3.7	1.7	**	3.4	2.0	**	2.7
Ghana (2003)	7.2	3.3	**	6.1	4.6	**	5.2
Kenya (2003)	12.2	6.0	**	15.4	5.7	**	9.1
Tanzania (2003-04)	8.7	4.1	**	8.5	5.3	**	6.2
Zambia (2001-02)	8.1	3.0	**	8.1	4.2	**	5.5

Notes: [a]Significant test of difference; **p<0.01; *p<0.05; NS- not significant.

Sources: CBS, MOH, & ORC Macro 2004; Central Statistical Office, Central Board of Health, and ORC Macro 2003; GSS, NMIMR & ORC Macro 2004; INS et ORC Macro 2004; INSD et ORC Macro 2004; TACAIDS, NBS and ORC Macro 2005.

Table 3: Percent HIV positive by selected background characteristics in three rural sites in Malawi, MDICP 2004 and Antenatal Clinic (ANC) 2003.

	MDICP 2004			
Characteristic	Male prevalence	Female prevalence	Both males and females	Number of cases tested
Site				
South (Balaka)	6.9	9.7	8.4	1073
Center (Mchinji)	6.3	8.1	7.3	904
North (Rumphi)	3.6	5.8	4.8	1011
Age group				
Adolescents (15-19 years)	0.4	1.0	0.7	582
Adults (20+ years)	8.0	9.5	8.8	2406
Total	**5.6**	**7.9**	**6.8**	**2988**
Number of cases	**1379**	**1609**	**2988**	
	Antenatal Clinic (ANC) 2003			
Gawanani (South)	14.2[a]	17.0		206[b]
Kamboni (Center)	5.6[a]	6.7		238[b]
Mbalachanda (North)	12.1[a]	14.5		193[b]
All rural sites		**14.5**		**1627**

Notes: [a]ANC prevalence rates for men are obtained by assuming a female-to-male HIV prevalence ratio of 1.2 to 1; [b]Number of cases refer to women.
Sources: MDICP 2004; Republic of Malawi (2003).

Table 4: Response rates for HIV test among all eligible respondents by site and sex in rural Malawi, MDICP 2004.

	Response rates for HIV test (percent)								
	South			Center			North		
Outcome	Males	Females	Total	Males	Females	Total	Males	Females	Total
Tested	71.7	76.1	74.0	73.2	74.4	73.9	73.5	71.1	72.2
Refused	8.3	7.9	8.1	7.1	9.0	8.2	6.9	5.2	6.0
Absent	2.5	1.1	1.7	3.0	1.2	2.0	3.2	2.0	2.6
Moved	13.1	10.2	11.6	2.9	2.6	2.7	12.0	16.5	14.4
Dead	2.2	2.1	2.1	2.9	2.4	2.6	2.0	2.3	2.1
Other	2.3	2.6	2.5	10.9	10.4	10.6	2.3	3.1	2.7
Cases (N)	**688**	**762**	**1450**	**560**	**664**	**1224**	**648**	**753**	**1401**

Note: Percentages may not add up to exactly 100 in some cases due to round-off error.

Table 5: Comparison of observed and estimated MDICP HIV prevalence rates under different assumptions of relative risks for HIV infection among non-respondents, MDICP 2004.

Assumed relative risk for non-respondents[a]	Estimated HIV prevalence rates assuming refusal is the only source of non-response (Panel A)					
	Males			Females		
	South	Center	North	South	Center	North
0.8	6.8	**6.2**	3.5	9.5	**7.9**	5.7
1.8	7.5	6.8	3.8	**10.4**	8.8	6.1
2.0	**7.6**	6.9	3.9	10.6	9.0	6.2
2.5	8.0	7.2	4.0	11.0	9.4	**6.4**
3.4	8.6	7.7	**4.3**	11.8	10.2	6.7
	Estimated HIV prevalence rates assuming refusal and absence as the only sources of non-response (Panel B)					
0.8	6.7	**6.2**	3.5	9.5	**7.9**	5.7
1.8	7.6	7.0	3.9	**10.5**	8.9	6.2
2.0	**7.8**	7.1	4.0	10.7	9.1	6.3
2.5	8.2	7.5	4.2	11.2	9.6	**6.6**
3.4	9.1	8.2	**4.6**	12.1	10.4	7.1
	Estimated HIV prevalence rates assuming refusal, absence and death as the only sources of non-response (Panel C)					
0.8	6.7	**6.1**	3.5	9.4	**7.9**	5.7
1.8	7.7	7.1	4.0	**10.6**	9.0	6.3
2.0	**8.0**	7.3	4.1	10.9	9.3	6.5
2.5	8.5	7.8	4.3	11.5	9.9	**6.8**
3.4	9.4	8.6	**4.8**	12.6*	10.9*	7.4
	Estimated HIV prevalence rates assuming refusal, absence, death and movement as the sources of non-response (Panel D)					
0.8	6.5	**6.1**	3.4	9.2	**7.8**	5.5
1.8	8.4	7.2	4.3	**11.4**	9.2	7.0
2.0	**8.7**	7.5	4.5	11.8	9.5	7.3
2.5	9.7*	8.0	4.9	12.8*	10.2	**8.1***
3.4	11.3**	9.0	**5.7***	14.7**	11.4*	9.5**

Notes: [a]Relative risk is defined as the ratio of the expected percentage of non-respondents HIV positive to the observed percent HIV positive among those tested; Figures in bold are the prevalence rates that were obtained by assuming that non-respondents had the same prevalence as that observed among antenatal clinic attendees in the adjacent rural ANC site; **p<0.01; *p<0.05.

Table 6: Comparison of MDICP and antenatal clinic (ANC) HIV prevalence rates under different assumptions of relative risks for HIV infection among all non-respondents, ANC 2003 and MDICP 2004.

Assumed relative risk for non-respondents[a]	Estimated HIV prevalence rates (percent) assuming ANC rates for all non-respondents (Males[b])					
	South		Center		North	
	Estimated percent	Difference from ANC	Estimated percent	Difference from ANC	Estimated percent	Difference from ANC
0.8	6.5	(-) **	**6.0**	(+) NS	3.4	(-) **
1.8	8.5	(-) *	7.7	(+) NS	4.3	(-) **
2.0	**8.9**	(-) *	8.0	(+) NS	4.5	(-) **
2.5	9.8	(-) NS	8.9	(+) NS	5.0	(-) **
3.4	11.6	(-) NS	10.4	(+) *	**5.9**	(-) **
	Estimated HIV prevalence rates (percent) assuming ANC rates for all non-respondents (Females)					
0.8	9.2	(-) **	**7.7**	(+) NS	5.5	(-) **
1.8	**11.5**	(-) *	9.8	(+) NS	7.1	(-) **
2.0	12.0	(-) NS	10.2	(+) NS	7.5	(-) **
2.5	13.1	(-) NS	11.2	(+) *	**8.3**	(-) **
3.4	15.2	(-) NS	13.1	(+) **	9.8	(-) NS

Notes: [a]Relative risk is defined as the ratio of the expected percentage of non-respondents HIV positive to the observed percent HIV positive among those tested; [b]Male ANC rate is computed by assuming a female to male HIV prevalence ratio of 1.2 to 1; the tests assume that the standard error for the proportion of males HIV positive is the same as that of females for the ANC data; Figures in bold are the prevalence rates that were obtained by assuming that non-respondents had the same prevalence as that observed among antenatal clinic attendees in the adjacent rural ANC site; **p<0.01; *p<0.05; NS: not significant; (+): positive difference; (-): negative difference.

Table 7: Results of the probit selection models predicting HIV status conditional on participation in testing, MDICP 2004.

Variables	Model 1	Model 2
	HIV status	
Age group (REF=Adults 20+ years)		
Adolescents (15-19 years)	-0.811** (0.240)	-0.759** (0.238)
Sex (Females=1)	0.162* (0.081)	0.148 (0.084)
Site (REF=South)		
Center	--	-0.041 (0.089)
North	--	-0.206* (0.094)
Partner usually stays outside village (REF=No)		
Yes	0.383** (0.123)	0.386** (0.121)
Not married†	-1.133 (0.638)	-1.015 (0.621)
Missing	-0.549 (0.388)	-0.469 (0.401)
Number of times married (REF=Married once)		
Never married	1.085 (0.638)	1.085 (0.618)
Married more than once	0.437** (0.079)	0.415** (0.081)
Missing	0.627 (0.368)	0.590 (0.364)
Ever used condoms/abstinence (REF=No)		
Never had sex	--	-0.459 (0.369)
Yes	--	-0.038 (0.104)
Missing	--	-0.093 (0.135)
	Participation in testing	
Highest education level (REF=No education)		
Primary education	-0.251** (0.074)	-0.258** (0.074)
Secondary and above	-0.323** (0.099)	-0.327** (0.097)
Other/missing	-1.051** (0.164)	-1.053** (0.163)
Household size (REF=Below median, <6)		
Median and over (6+)	0.051 (0.062)	0.053 (0.062)
Missing	-0.388** (0.078)	-0.362** (0.077)
Stayed outside district 6+ months (REF=No)		
Yes	-0.094 (0.055)	-0.099 (0.055)
Can't remember/missing	2.374** (0.208)	2.289** (0.240)
Partner usually stays outside village (REF=No)		
Yes	-0.384** (0.089)	--
Not married†	0.143 (0.082)	--
Missing	-0.229 (0.160)	--
Worried about catching AIDS (REF=No)		
Yes	0.083 (0.054)	0.090 (0.055)
Don't know/missing	-1.538** (0.087)	-1.545** (0.160)
Suspects spouse of infidelity (REF=No/DK)		
Not married†	--	0.273** (0.105)
Knows/suspects	--	-0.012 (0.057)
Missing	--	-0.109 (0.220)

Table 7 (cont'd)

Variables	Model 1	Model 2
Previously tested for HIV (REF=No)		
Yes	0.408** (0.084)	0.396** (0.084)
Missing	-1.695** (0.087)	-1.714** (0.087)
ρ (rho)	0.09 (0.155)	0.05 (0.162)
LR test of independence of equations	χ^2=0.34; p=0.559	χ^2=0.11; p=0.745

Notes: REF-Reference category; DK- don't know; LR- likelihood ratio test; †Not married refers to never married, divorced, separated, and widowed; robust standard errors (in parentheses) were estimated; ρ = corr(ε_{1i}, ε_{2i}); *p<0.05; **p<0.01.

Figure 1: HIV prevalence among pregnant mothers, ANC 2003 and MDICP 2004

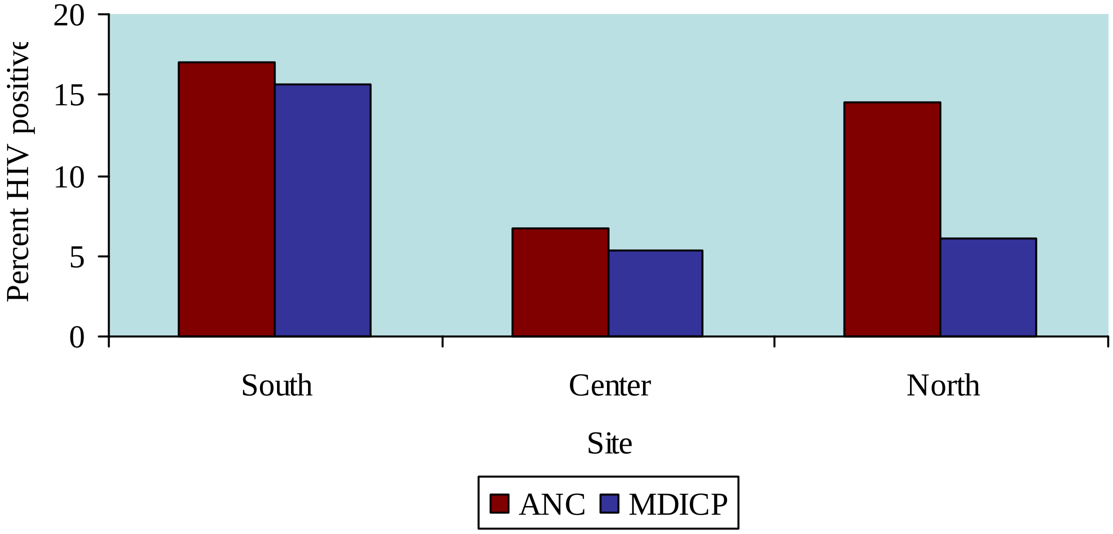


Figure 2a: Observed versus predicted HIV prevalence: men, first model

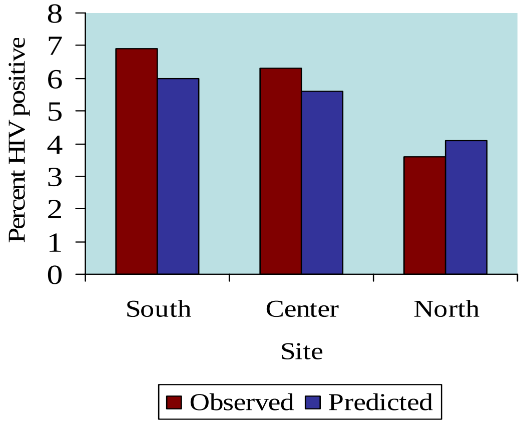
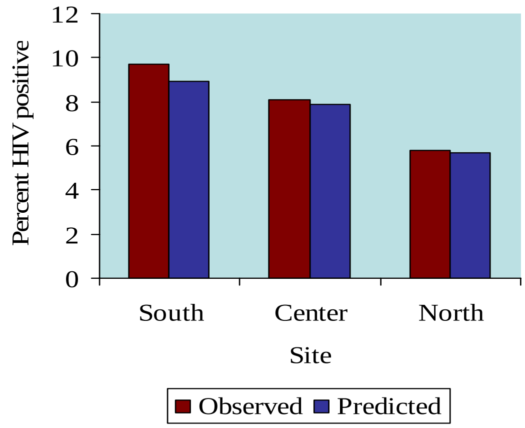


Figure 2b: Observed versus predicted HIV prevalence: women, first model



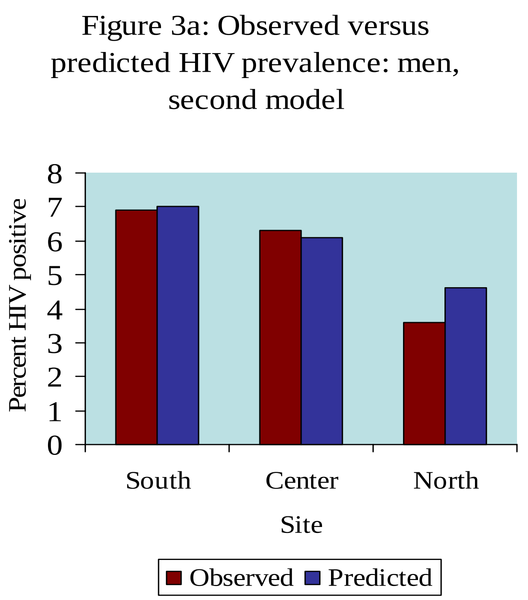


Figure 3a: Observed versus predicted HIV prevalence: men, second model
Percent HIV positive
8
7
6
5
4
3
2
1
0
South
Center
North
Site
Observed
Predicted

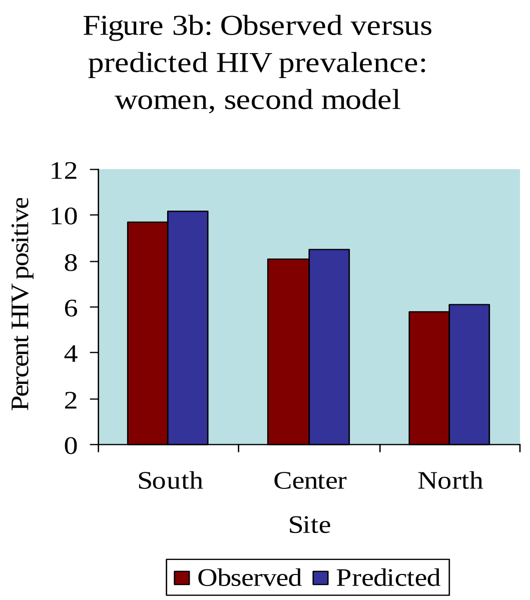


Figure 3b: Observed versus predicted HIV prevalence: women, second model
Percent HIV positive
12
10
8
6
4
2
0
South
Center
North
Site
Observed
Predicted