2-2019

Refusal bias in HIV data from the Demographic and Health Surveys: Evaluation, critique and recommendations

Oyelola A. ADEGBOYE
James Cook University

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

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

SMU ECONOMICS & STATISTICS

Refusal bias in HIV data from the Demographic and Health Surveys:

Evaluation, critique and recommendations

Oyelola A. Adegboye, Tomoki Fujii, Denis H.Y. Leung

February 2019

Paper No. 06-2019

Refusal bias in HIV data from the Demographic and Health Surveys: Evaluation, critique and recommendations*

Oyelola A. Adegboye (corresponding author)
Australian Institute of Tropical Health & Medicine, James Cook University
E-mail: oyelola.adegboye@jcu.edu.au

Tomoki Fujii
School of Economics, Singapore Management University
E-mail: tfujii@smu.edu.sg

Denis H.Y. Leung
School of Economics, Singapore Management University
E-mail: denisleung@smu.edu.sg

February 25, 2019

*We acknowledge ORC Macro for granting us access to the MDHS data. 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. The ANC data were obtained from the 2003 Malawi National AIDS Commission report. The census data were part of the 1998 and 2008 Population and Housing Census carried out by the National Statistical Office, Government of Malawi and made available by the Minnesota Population Center (Integrated Public Use Microdata Series, International: Version 6.1 [Machine-readable database]. Minneapolis: University of Minnesota, 2011).

Refusal bias in HIV data from the Demographic and Health Surveys: Evaluation, critique and recommendations

Abstract

Non-response is a commonly encountered problem in many population-based surveys. Broadly speaking, non-response can be due to refusal or failure to contact the sample units. Although both types of non-response may lead to bias, there is much evidence to indicate that it is much easier to reduce the proportion of non-contacts than to do the same with refusals. In this article, we use data collected from a nationally-representative survey under the Demographic and Health Surveys program to study non-response due to refusals to HIV testing in Malawi. We review existing estimation methods and propose novel approaches to the estimation of HIV prevalence that adjust for refusal behaviour. We then explain the data requirement and practical implications of the conventional and proposed approaches. Finally, we provide some general recommendations for handling non-response due to refusals and we highlight the challenges in working with Demographic and Health Surveys and explore different approaches to statistical estimation in the presence of refusals. Our results show that variation in the estimated HIV prevalence across different estimators is due largely to those who already know their HIV test results. In the case of Malawi, variations in the prevalence estimates due to refusals for women are larger than those for men.

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

1 Introduction

In sub-Saharan Africa, home to around 23 million people living with HIV,HIV (2011) accurate measurement of the trends of important diseases such as HIV is essential for governments to design policies and aid programs. In the past two decades, national population-based surveys have become an important source for such measurement. Boerma et al. (2003), Garcia-Calleja et al. (2006) A major challenge in using these survey data is the potential bias from missing data created by non-response. There is much evidence that non-respondents may have patterns of outcome and/or behaviour that are very different from those of the rest of the population.Marston et al. (2008)

The problem of non-response has always been a concern for those who work with survey data. One reason why non-response has captured so much attention from researchers is because the nature of the problem is complex. It is widely acknowledged that non-response does not arise from a unitary source under a well-defined situation. Rather, the causes and processes that lead to non-response are varied and often a function of multiple factors, such as the population under study, the nature of the outcome, and the way the survey is designed and conducted. A most challenging issue is that information about the non-respondents is usually scant, making it very difficult for surveyors to determine the nature of non-response.

Non-response arises when sample units in a survey refuse to respond or when the surveyors fail to contact a sample unit. (Groves et al., 2002) Many researchers distinguish between non-contacts and refusals because the processes leading to these two types of non-response are believed to be distinct. There are good reasons for espousing this belief. For example, in the context of an HIV survey in rural Africa where the sample units are asked to participate in an HIV test, a non-contact is often the result of migration of the household or absence for work. However, a refusal may be the result of the sample unit's

knowledge of his/her HIV status. (Obare et al., 2009) Furthermore, we can argue that a non-contact is the result of a passive behaviour since a move of address is a family-based decision that is less likely to be related to the sample unit's HIV status whereas a refusal is an active decision by the sample unit not to provide information about his/her HIV status.[1] Therefore, different approaches are required to address non-contact and refusal. For example, as the study of six national surveys in the UK indicates, repeated efforts to contact the subject may be able to reduce non-contacts but the same cannot be said about refusals. (Lynn and Clarke, 2002)

While there has been a lot of attention paid to issues related to non-response, most of the attention has been directed towards surveys carried out in the developed world. (Lynn and Clarke, 2002, Hawkes and Plewis, 2006, Billet et al., 2007, Durrant and Steele, 2009, Lynn, 2012) We argue that there is a need to consider the problem separately for surveys carried out in developing countries. Our argument rests on three observations. First, in some parts of the developed world, many non-response problems can be, at least partially, resolved by linking survey data to administrative records, (Thomsen and Holmøy, 1998, Zanutto and Zaslavsky, 2002, Yucel and Zaslavsky, 2005, van den Berg et al., 2006) which is often rich in content and well documented. The same cannot be done easily in many parts of Africa and elsewhere in the developing world as such records often do not exist, are poorly archived, or outdated. Second, many researchers advocated using callbacks to reduce the non-response rate. (Stoop, 2004, Kreuter et al., 2010, Olson, 2013) While the developing world has witnessed a massive expansion of mobile phone and broadband networks, such means of contacting sampled units remain practically infeasible in impoverished areas where telephones and computers are not affordable or in sparsely populated areas without easy access to such networks. Third, it is often difficult to rule out that non-response is non-informative. In that situation, unbiased inferences are still possible by combining the survey data with information from longitudinal data in a comparable population. (Alho, 1990, Burton et al., 2006, Billet et al., 2007) In many parts of the developing world, however, the organisation of a nationally representative longitudinal study is difficult due to mobility of individuals and lack of reliable demographic records, especially in rural areas, statistical capacity, and necessary financial resources. Hence, such a strategy needs to be adapted to the conditions in the developing world.

In this paper, we study non-response due to refusals to HIV testing using data collected from a nationally-representative survey under the Demographic and Health Surveys (DHS) program. Some relevant earlier works include, Garcia-Calleja et al., Garcia-Calleja et al. (2006) who carried out a scenario study for 20 sub-Saharan countries using HIV relative risks between the non-respondents and respondents. However, they did not treat non-contacts and refusals separately. Marston et al. Marston et al. (2008) examined non-response bias in a nine-country study. They assumed non-response is non-informative and estimated the prevalence among the non-respondents by multiple imputation. (Rubin, 1987) Similarly,

[1] There is a possibility that people move because of the positive HIV status, for example, to seek for medical care in urban areas.

Mishra et al. Mishra et al. (2008) used a logistic regression to predict the HIV prevalence among the non-respondents under a non-informative non-response assumption in a twelve-country study. Hogan et al. Hogan et al. (2012) adjusted non-response bias by a selection model, (Heckman, 1979) which allows non-response to be informative but requires the existence of a valid instrumental variable that explains non-response but not the outcome. Reniers and Eaton Reniers and Eaton (2009) and Floyd et al. Floyd et al. (2013) corrected refusal bias in population surveys by using auxiliary longitudinal data. Their methods rely on the assumption that refusal behaviour in different populations are comparable. In some of the methods discussed below, we also adopt a similar assumption.

The main contribution of this paper is threefold. First, we put together existing methods and re-examine their underlying assumptions; we discuss the possible merits and demerits of each of these assumptions. Second, we introduce a few alternative novel approaches to HIV prevalence estimates that adjust for refusal behaviour and compare them to existing methods on a common platform. This comparison allows us to determine how important refusal bias may be. Third, based on thorough robustness checks against potential refusal bias, we draw lessons that could be applied elsewhere.

2 Study design and survey data

In health and population studies in Africa, the following three types of survey data are often available: national population-based surveys, sentinel surveillance surveys, and longitudinal surveys. National population based surveys are usually large scale, cross-sectional studies with the intent of drawing nationally representative samples. They collect detailed demographic characteristics and various outcomes of interest, such as health, nutrition, and land use. Sentinel surveillance surveys are useful for capturing cross-sectional data over time, such as outbreak of disease, nutritional trends, and changes in land use, at sentinel sites. The sentinel sites are typically located in the more densely populated urban areas and hence may not be representative of the general population in many developing countries, since most of them have a sizable proportion of rural population. Longitudinal surveys collect data on vital events and migration for individuals and households over time. When linked with appropriate data, such as individual demographic and behaviour information, longitudinal survey data make it possible to evaluate cause-specific impacts on outcome of interest. However, since longitudinal surveys are often carried out in smaller communities at specific locations, inferences drawn from them are unlikely to be directly applicable to the general population. We use all of these three types of surveys in Malawi for empirical illustration. We examine the relevance and implications of different approaches to the estimation of HIV prevalence.

The primary data source for this study is the 2004 Malawi Demographic and Health Survey (MDHS), which is a nationally-representative survey. All women aged 15-49 years in a selected household are

eligible for interview. In about one in three selected households, male members of the household aged 15-54 years are also surveyed and HIV testing is offered to both male and female members. We focus on those aged 49 years or below to keep the same age group for both women and men, and also make our study comparable to the MDHS report.(National Statistical Office and ORC Macro, 2005) In addition, we exclude those who refused to answer the individual questionnaire, those who consented but their HIV testing results are not available (e.g., technical problem), and those whose previous HIV testing history (i.e., whether the individual has previously taken an HIV test) is not known. We note that Lilongwe district has an unusually high refusal rate (54%) and low observed prevalence (Figure 1). In an earlier report, (National Statistical Office and ORC Macro, 2005) separate analyses were carried out with and without Lilongwe. To facilitate comparison with earlier studies, Reniers and Eaton (2009), National Statistical Office and ORC Macro (2005) we elect to include Lilongwe in the main article; the results of a parallel analysis, excluding Lilongwe, are given separately as online supplemental materials.

In addition to the MDHS data, we also use the 2003 Malawi antenatal clinics (ANC) survey data. (National AIDS Commission, 2003a) The collection of HIV data in the Malawi ANC started in 1990 and by 2003, there were 19 ANC sites in Malawi. In the 2003 ANC, HIV data were collected on nearly 8000 pregnant women, of which 20%, 49%, and 31% are in rural, semi-urban, and urban areas, respectively. (National AIDS Commission, 2003a)

Lastly, we use a dataset collected under the Malawi Diffusion and Ideational Change Project (MDICP), which consists of a series of longitudinal surveys conducted in the rural areas in three districts of Malawi, one from each of the Southern, Central, and Northern regions of Malawi. As such it is not representative of the general population of Malawi. The sample is made up of married women and their husbands in the selected households. We only use the 2004 (MDICP-3) and 2006 (MDICP-4) phases as HIV test component is available only for these phases.

3 Assumptions and methods for estimating HIV prevalence

The goal of our research is to estimate HIV prevalence in a population of interest using sample surveys (such as DHS) drawn randomly from the population. However, such surveys might suffer from non-responses due to refusals which might lead to bias. In this section, we discuss various methods for estimating HIV prevalence, including those previously used in the literature and some newly introduced in this study. We begin our analysis by first ignoring selection bias and estimate HIV prevalence by simply taking the sample proportion of HIV status based on only those who accept an HIV test.

Let D_i be an indicator variable that takes one if individual i is HIV positive and zero otherwise. The goal of our research is to identify $\pi \equiv E[D_i]$, where i is drawn randomly from the population of interest. Sometimes, we are also interested in HIV prevalence of certain sub-populations. In that case,

the parameter of interest is $E[D_i|Z_i]$, where Z_i is a variable that characterises the sub-populations, which may include the location of residence, gender, occupation, and education level, etc. However, we drop Z_i hereafter, because the same method can be used for estimating HIV prevalence in each sub-population of interest by restricting the sample used for estimation accordingly.

We typically estimate $E[D_i]$ from sample surveys such as DHS, because it is prohibitively expensive and practically infeasible to measure D_i for all individuals in the population. Let N be the total number of individuals in our MDHS sample and R_i is an indicator variable for refusal such that $R_i = 0$ indicates the individual i accepts an HIV test. Therefore, if we ignore the selection on non-refusals, $E[D_i]$ can be estimated by the complete case estimator:

$$\hat{\pi}_{CC} = \frac{\sum_{i=1}^{N}(1 - R_i)D_i}{\sum_{i=1}^{N}(1 - R_i)}. \tag{1}$$

An advantage of the estimator $\hat{\pi}_{CC}$ is that it is easy to calculate and requires no additional models. However, even if the sample is random, $\hat{\pi}_{CC}$ is only an unbiased estimator for $E[D_i|R_i = 0]$ and not for $E[D_i]$ in general. Hence, unless we have $E[D_i] = E[D_i|R_i = 0]$, $\hat{\pi}_{CC}$ is good only as an estimator of HIV prevalence of those who would agree to take an HIV test when such a test is offered.

In practice, we have no strong reason to believe *a priori* that $E[D_i] = E[D_i|R_i = 0]$ holds. To address this issue, certain additional assumptions and/or data are required. For example, assume that HIV status can be explained by a set of covariates X_i observable on every individual in the MDHS data and that there is no refusal bias.

In the current context, this method requires

$$P(D_i = 1|X_i, R_i = 0) = P(D_i = 1|X_i, R_i = 1) = P(D_i = 1|X_i). \tag{2}$$

If an unbiased estimator $\hat{D}_i$ of $P(D_i = 1|X_i)$ can be obtained from those with observed HIV status, then we can estimate the prevalence by a method equivalent to the mean score imputation (MSI) method, e.g., Pepe et al., (Pepe et al., 1994) in the missing data literature.

$$\hat{\pi}_{MSI} = \frac{\sum_i (1 - R_i)D_i + R_i\hat{D}_i}{N}, \tag{3}$$

As we pointed out earlier, the estimator $\hat{\pi}_{CC}$ is generally a biased estimator of $E[D_i]$. Another possibility is to model the probability of refusal using covariates X_i and assume that D_i is conditionally independent of refusal, given X_i (eq. (2)). To keep the presentation simple, we temporarily assume that X_i is discrete but this assumption can be relaxed. With these assumptions, we have:

$$E[D_i] = \sum_x E[D_i|X_i = x] \cdot P[X_i = x] = \sum_x E[D_i|R_i = 0, X_i = x] \cdot P[X_i = x].$$

Let $I(\cdot)$ be an indicator function (which takes one if its argument is true and zero otherwise), we can estimate $E[D_i]$ by:

$$\sum_x \left[\frac{\sum_{i=1}^N (1-R_i) D_i I(X_i = x)}{\sum_{i=1}^N (1-R_i) I(X_i = x)} \right] \left[\sum_{i=1}^N \frac{I(X_i = x)}{N} \right]. \tag{4}$$

The estimator above is unbiased if a suitable discrete covariate X_i can be found. In practice, a discrete covariate is often not sufficient to completely explain selection due to refusal. A more general estimator

$$\hat{\pi}_{IF} = \sum_{i=1}^N \frac{(1-R_i)D_i}{P(R_i = 0)} \Big/ \sum_{i=1}^N \frac{(1-R_i)}{P(R_i = 0)}, \tag{5}$$

is unbiased for $E[D_i]$. We use "IF", which stands for infeasible, to qualify this estimator because $P(R_i = 0)$ is generally unknown. If we replace $P(R_i = 0)$ by an estimator $\hat{P}(R_i = 0) \equiv \hat{P}(R_i = 0|X_i)$ in $\hat{\pi}_{IF}$ and call this estimator $\hat{\pi}_1$, then it becomes the well known inverse probability or inverse propensity score estimator. (Horvitz and Thompson, 1952) The estimator $\hat{\pi}_1$ can be viewed as a continuous version of eq. (4). Unbiasedness of $\hat{\pi}_1$ requires $P(R_i = 0) = P(R_i = 0|X_i) = P(R_i = 0|X_i, D_i)$, which is the conditional independence assumption for eq. (4).

A common strategy to come up with $\hat{P}(R_i = 0|X_i)$ is to use a parametric model, usually a logistic regression using variables that are thought to predict acceptance of an HIV test (see, for example, National Statistical Office and ORC Macro, 2005 Appendix G (National Statistical Office and ORC Macro, 2005)). However, this strategy works only if the model of acceptance is known and covariates in the model are observable.

To address refusal due to the prior knowledge of HIV status, Reneirs and Eaton Reniers and Eaton (2009) suggested a method to estimate $E[D_i]$ under the following two assumptions:

$$\begin{aligned} P(R_i = 1|D_i = 1, T_i = 0) &= P(R_i = 1|D_i = 0, T_i = 0) \\ &= P(R_i = 1|T_i = 0), \qquad (6) \\ P(D_i = 1|T_i = 1) &= P(D_i = 1), \qquad (7) \end{aligned}$$

where $T_i = 0$ means that a subject does not know his/her HIV status and $T_i = 1$ means that a subject has had an HIV test and knows the test result. The first assumption given in eq. (6) states that refusal is independent of HIV status given that the subject has never taken an HIV test before. The second assumption in eq. (7) states that being tested previously does not depend on one's HIV status. Under these assumptions, the following quadratic equation in $P(D_i = 1) \equiv E[D_i]$ can be shown to hold:

$$\begin{aligned} 0 = {} & [\{P(R_i = 0|T_i = 0)P(T_i = 0) + P(T_i = 1)\}(\Delta - 1)]P(D_i = 1)^2 + \\ & [-P(D_i = 1|R_i = 0)P(R_i = 0)(\Delta - 1) + P(R_i = 0|T_i = 0)P(T_i = 0) + \end{aligned}$$

$$\{1 - \Delta P(R_i = 1|T_i = 1)\}P(T_i = 0)]P(D_i = 1) -$$
$$P(D_i = 1|R_i = 0)P(R_i = 0), \tag{8}$$

where the relative risk of refusal Δ is defined as follows:

$$\Delta \equiv \frac{P(R_i = 1|D_i = 1, T_i = 1)}{P(R_i = 1|D_i = 0, T_i = 1)}.$$

Reniers and Eaton Reniers and Eaton (2009) used MDICP data to estimate Δ and MDHS data to estimate the remaining quantities in eq. (8). Their estimator $\hat{\pi}_{RE}$ of $E[D_i]$ is the unique root of the quadratic equation on the unit interval.

There are a few issues with the assumptions above. First, notice that eqs. (6) and (7) imply:

$$P(D_i = 1) = P(D_i = 1|T_i = 0) = P(D_i = 1|T_i = 0, R_i = 0). \tag{9}$$

This suggests we can estimate the prevalence of HIV by:

$$\hat{\pi}_2 = \frac{\sum_{i=1}^{N}(1 - R_i)D_i(1 - T_i)}{\sum_{i=1}^{N}(1 - R_i)(1 - T_i)}. \tag{10}$$

Therefore, once eqs. (6) and (7) are assumed, we do not need MDICP data to estimate the HIV prevalence. Second, both of these assumptions may be problematic in practice. Eq. (6) is not compelling because individuals may know the risk of HIV infection even without HIV testing. Eq. (7) may also be called into question, because those who have taken HIV tests before may be systematically different from others.

Given these issues, we propose to estimate lower and upper bounds of $P(D_i = 1) \equiv E[D_i]$ under the following assumptions:

$$P(D_i = 1|\tilde{T}_i = 0, R_i = 0) \leq P(D_i = 1|\tilde{T}_i = 0, R_i = 1) \leq P(D_i = 1|\tilde{T}_i = 1, R_i = 1). \tag{11}$$

where $\tilde{T}_i$ differs slightly from the definition of T_i used by Reniers and Eaton Reniers and Eaton (2009) in that $\tilde{T}_i = 0$ means a subject has not taken a prior HIV test and $\tilde{T}_i = 1$ represents a subject has had an HIV test but may or may not know the result of the test. The first inequality in eq. (11) captures the idea that those who refuse to take HIV test are no less likely to be HIV positive than those who participate, given that they have never taken an HIV test before. Note that the first inequality becomes an equality when eq. (6) is satisfied. The second inequality captures the idea that those who have previously taken an HIV test are no less likely to be HIV positive than those who have never taken a test given they refuse to participate in the HIV testing.

In addition to these assumptions, we explicitly account for the fact that MDICP is not representative of the general population of Malawi, because the data are taken only from a few rural districts. We

use $M_i = 1$ to denote individual i belongs to the population that the MDICP sample represents, and zero otherwise. We assume that the relative risk of HIV between MDICP population and non-MDICP population is independent of refusal given that an individual has had a previous HIV test. Mathematically, our assumption implies:

$$\begin{aligned} Z &\equiv \frac{P(D_i = 1|\tilde{T}_i = 1, M_i = 0)}{P(D_i = 1|\tilde{T}_i = 1, M_i = 1)} \\ &= \frac{P(D_i = 1|\tilde{T}_i = 1, R_i = 1, M_i = 0)}{P(D_i = 1|\tilde{T}_i = 1, R_i = 1, M_i = 1)} \\ &= \frac{P(D_i = 1|\tilde{T}_i = 1, R_i = 0, M_i = 0)}{P(D_i = 1|\tilde{T}_i = 1, R_i = 0, M_i = 1)}. \end{aligned} \tag{12}$$

Under this assumption, the numerator and denominator of the last line of eq. (12) can be estimated with the MDHS and MDICP data, respectively. Letting $W \equiv P(M_i = 1) + ZP(M_i = 0)$, we can write:

$$\begin{aligned} P(D_i = 1|\tilde{T}_i = 1, R_i = 1) &= P(D_i = 1|\tilde{T}_i = 1, R_i = 1, M_i = 1)P(M_i = 1) \\ &\quad + P(D_i = 1|\tilde{T}_i = 1, R_i = 1, M_i = 0)P(M_i = 0) \\ &= P(D_i = 1|\tilde{T}_i = 1, R_i = 1, M_i = 1)W, \end{aligned} \tag{13}$$

where we additionally made the assumption that $P(M_i|\tilde{T}_i, R_i) = P(M_i)$. In the MDICP data, we observe the HIV status of those who participate in the first HIV test but refuse the second HIV test. Therefore, we can estimate $P(D_i = 1|\tilde{T}_i = 1, R_i = 1, M_i = 1)$ by the proportion of HIV positives in the first test among those who refuse the second test. To use eq. (13), we also need to estimate W, which in turn requires estimates of $P(M_i = 0), P(M_i = 1)$, and Z. Since the MDICP sample was taken to match closely the rural sample of the 1996 MDHS, we may take $P(M_i = 1)$ to represent the proportion of rural population in Malawi and $P(M_i = 0)$ the urban population, both of which can be estimated using population census data. For Z, we can use the MDHS and MDICP data to estimate the numerator and denominator, respectively.

We also define:

$$Z' = \frac{P(D_i = 1|\tilde{T}_i = 0, R_i = 0, M_i = 0)}{P(D_i = 1|\tilde{T}_i = 0, R_i = 0, M_i = 1)}, \tag{14}$$

and letting $W' \equiv P(M_i = 1) + Z'P(M_i = 0)$, we can write:

$$\begin{aligned} P(D_i = 1|\tilde{T}_i = 0, R_i = 0) &= P(D_i = 1|\tilde{T}_i = 0, R_i = 0, M_i = 1)P(M_i = 1) \\ &\quad + P(D_i = 1|\tilde{T}_i = 0, R_i = 0, M_i = 0)P(M_i = 0) \\ &= P(D_i = 1|\tilde{T}_i = 0, R_i = 0, M_i = 1)W'. \end{aligned} \tag{15}$$

Estimation of eq. (15) follows easily since the numerator and denominator of $Z^{'}$ can be estimated using data from MDHS and MDICP, respectively.

Using eq. (13), we have the following relationship:

$$\begin{aligned} P(D_i=1) &= P(D_i=1,R_i=0)+P(D_i=1|\tilde{T}_i=1,R_i=1)P(\tilde{T}_i=1,R_i=1) \\ &\quad +P(D_i=1|\tilde{T}_i=0,R_i=1)P(\tilde{T}_i=0,R_i=1) \\ &= P(D_i=1,R_i=0) \\ &\quad +WP(D_i=1|\tilde{T}_i=1,R_i=1,M_i=1)P(\tilde{T}_i=1,R_i=1) \\ &\quad +P(D_i=1|\tilde{T}_i=0,R_i=1)P(\tilde{T}_i=0,R_i=1). \end{aligned} \tag{16}$$

Notice that in eq. (16), $P(D_i=1|\tilde{T}_i=0,R_i=1)$ cannot be estimated because test results are not available for those individuals who have had no prior HIV test and decline the current test. Hence, the estimation of eq. (16) is not feasible. However, by eq. (11), (15), and (16), we can form bounds:

$$P_- \leq P(D_i=1) \leq P_+,$$

where

$$\begin{aligned} P_- &= P(D_i=1,R_i=0)+WP(D_i=1|\tilde{T}_i=1,R_i=1,M_i=1)P(\tilde{T}_i=1,R_i=1) \\ &\quad +P(D_i=1|\tilde{T}_i=0,R_i=0)P(\tilde{T}_i=0,R_i=1) \\ &= P(D_i=1,R_i=0)+WP(D_i=1|\tilde{T}_i=1,R_i=1,M_i=1)P(\tilde{T}_i=1,R_i=1) \\ &\quad +W^{'}P(D_i=1|\tilde{T}_i=0,R_i=0,M_i=1)P(\tilde{T}_i=0,R_i=1), \end{aligned} \tag{17}$$

$$\begin{aligned} P_+ &= P(D_i=1,R_i=0)+WP(D_i=1|\tilde{T}_i=1,R_i=1,M_i=1)P(\tilde{T}_i=1,R_i=1) \\ &\quad +P(D_i=1|\tilde{T}_i=1,R_i=1)P(\tilde{T}_i=0,R_i=1). \\ &= P(D_i=1,R_i=0)+WP(D_i=1|\tilde{T}_i=1,R_i=1,M_i=1)P(R_i=1). \end{aligned} \tag{18}$$

We can estimate $P(D_i=1,R_i=0)$, $P(\tilde{T}_i=0,R_i=1)$ and $P(R_i=1)$ with the MDHS data. Other terms can be estimated by eqs. (12), (13), (14), and (15) with the MDICP data. For the computation of the estimators, the following definitions in the MDICP data are used: $\tilde{T}_i=1$ if an individual has a test in MDICP-3, $D_i=1$ if an individual tests positive in MDICP-3 or MDICP-4, $R_i=1$ if an individual tests in MDICP-3 but refuses a test in MDICP-4. Using these estimates, we obtain the estimates $\hat{\pi}_{3-}$ and $\hat{\pi}_{3+}$ of P_- and P_+, respectively.

A third source of data that allows estimation of $E[D_i]$ is the ANC surveys.(The POLICY Project, 2001, National AIDS Commission, 2003a) To produce national prevalence estimates, the district-area prevalence estimates obtained using ANC data are combined with census data. For each district-area c

Table 1: Summary of estimators considered in this study

Estimator	Key equation(s)	Identifying assumption(s)	Data source
$\hat{\pi}_{CC}$	eq. (1)	No refusal bias	MDHS
$\hat{\pi}_{MSI}$	eq. (3)	eq. (2) and No refusal bias conditional on X_i	MDHS
$\hat{\pi}_{IF}$	eq. (5)	Infeasible	
$\hat{\pi}_1$	eq. (4)-eq. (5)	Use $\hat{P}(R_i = 0) \equiv \hat{P}(R_i = 0\|X_i)$ in $\hat{\pi}_{IF}$	MDHS
$\hat{\pi}_2$	eq. (10)	eqs.(6)-(7)	MDHS
$\hat{\pi}_{RE}$	eq. (8)	eqs.(6)-(7); see also Reniers and Eaton (2009)	MDHS, MDICP
$\hat{\pi}_{3+}, \hat{\pi}_{3-}$	eqs. (17)-(18)	eqs. (11)-(12)	MDHS, MDICP, Census
$\hat{\pi}_4$	eq. (19)	eq. (20)	ANC, Census
$\hat{\pi}_{5A}$	eq. (23)	eqs. (21)-(22): Stepwise regression using X_i and $\hat{\pi}^c_{ANC}$	MDHS, ANC
$\hat{\pi}_{5B}$	eq. (23)	eqs. (21)-(22): Fixed regression using $\hat{\pi}^c_{ANC}$ only	ANC

captured in ANC surveys, let w_c be a weight that gives the proportion of individuals living in district-area c from the census (We use 1998 census figures for all district-areas except Likoma and Mzuzu. For Likoma and Mzuzu, separate figures were not given in the 1998 census, so we use figures from the 2008 census). Then an estimator of the population HIV prevalence is:

$$\hat{\pi}_4 = \sum_c \left(\hat{\pi}^c_{ANC} \frac{w_c}{\sum_{c'} w_{c'}} \right), \tag{19}$$

where $\hat{\pi}^c_{ANC}$ is the prevalence estimator in district-area c using ANC data. This method has also been used in cross-national studies comparing ANC-based to population-based survey estimates.(Montana et al., 2008)

If we let $\tilde{M}_i = 1$ be an indicator for an individual who has been tested at an ANC site, then $\hat{\pi}_4$ makes the following assumption:

$$P(D_i = 1|\tilde{M}_i = 1, C_i = c) = P(D_i = 1|\tilde{M}_i = 0, C_i = c) = P(D_i = 1|C_i = c), \tag{20}$$

where C_i is defined as the index of the district-area in which the i-th individual resides, such that $C_i = c$ means that an individual comes from district-area c. In other words, given that individuals are matched by district-area, the prevalence of HIV of the ANC attendees is the same as that in the general population.

When refusal to an HIV test may be due to the (unobservable) HIV status of a sampled unit,(Reniers and Eaton, 2009, Floyd et al., 2013) then the use of known data to estimate $P(R_i = 0)$ will not yield the desired results. This is the classical problem of non-ignorable missingness in the missing data literature.(Little and Rubin, 2002)

We propose a method that mitigates the problem of non-ignorable missingness by using information

routinely recorded in ANC surveys. We assume

$$P(R_i = 0) = g(D_i, X_i) \equiv P(R_i = 0|D_i, X_i) \tag{21}$$

for some known function g that depends on the HIV status D_i and some observable covariates X_i. Of course, eq. (21) cannot be used because D_i is unknown for those who refuse an HIV test. Therefore, we make the following assumption:

$$P(R_i = 0|X_i = x, D_i, \hat{\pi}^c_{ANC}, C_i = c) = P(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c), \tag{22}$$

which says that for an individual in a particular district-area, acceptance of an HIV test is independent of the individual's HIV status, given the covariates and the HIV prevalence in that district-area estimated from the ANC data.

The conditional independence assumption eq. (22) allows us to have a workable solution since $\hat{\pi}^c_{ANC}$ can be obtained using data in every HIV sentinel surveillance report. (National AIDS Commission, 2003a, 2008) Let $\hat{P}(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)$ be an estimator of $P(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)$ which may be based on a logistic regression model. Then, we estimate $E[D_i]$ by

$$\hat{\pi}_5 = \sum_{i=1}^{N} \frac{(1 - R_i)D_i}{\hat{P}(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)} \Bigg/ \sum_{i=1}^{N} \frac{(1 - R_i)}{\hat{P}(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)}. \tag{23}$$

We consider two estimators based on $\hat{\pi}_5$. The first one uses $\hat{\pi}^c_{ANC}$ and a stepwise regression procedure to select from the same list of covariates X_i used in $\hat{P}(R_i = 0|X_i)$ for the estimation of $\hat{\pi}_1$. The second one uses only $\hat{\pi}^c_{ANC}$ for modeling the propensity score. These propensity scores are the used in eq. (23) to give different prevalence estimators, $\hat{\pi}_{5A}$ and $\hat{\pi}_{5B}$, respectively.

A summary of this and other estimators considered in this paper with their key estimation equations, identifying assumptions and data requirement is given in Table 1.

4 Results

4.1 Refusal patterns

We first study the possible bias in the prevalence estimates due to refusals. We begin by summarising the refusal patterns in the data in Table 2. It is clear from the table that the refusal rate of around 23.9% in MDHS is far higher than those in the other two surveys. There are no refusals in the ANC survey as HIV test was carried out based on blood samples left behind for syphilis test and no consent was sought. For MDICP-3, the refusal rate is about 9.5% and for MDICP-4, we obtain a refusal rate of 5.4%, among those who tested in MDICP-3. The refusal rates among men are similar to those in women, in all

Table 2: Refusal patterns in MDHS, ANC and MDICP

Source	No. Eligible	No. refused	Percent
MDHS	6696	1601	23.9
ANC†	7977	0	0.0
MDICP-3‡	3123	304	9.5
MDICP-4§	2111	115	5.4

†Consent not required
‡Among those contacted in MDICP-3
§Among those tested in MDICP-3 and contacted in MDICP-4

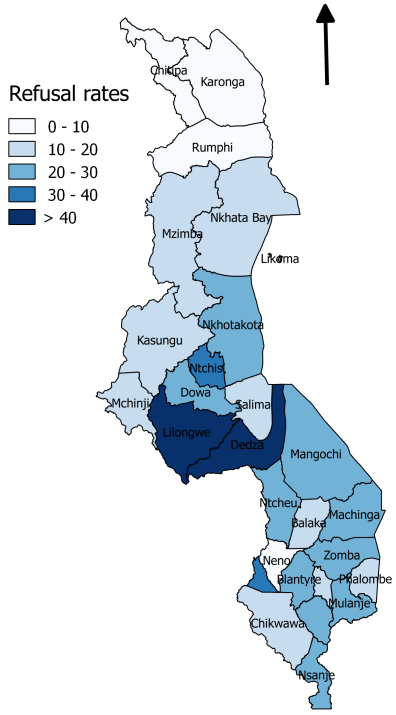


Figure 1: Malawi district level HIV testing refusal patterns in 2004 MDHS .

surveys. For MDHS, the refusal rate for men is $715/2984 \approx 0.240$ and for women is $886/3712 \approx 0.239$; the corresponding figures for MDICP-3 are $141/1490 \approx 0.094$ and $163/1723 \approx 0.094$, respectively, and for MDICP-4, $55/948 \approx 0.058$ and $60/1163 \approx 0.052$, respectively. Similar patterns of refusal rates are reported elsewhere. (Reniers and Eaton, 2009, Obare, 2010) The slight differences between our figures and those reported in Reniers and Eaton Reniers and Eaton (2009) and Obare Obare (2010) can be attributed to the different baseline samples used (For example, Reniers and Eaton Reniers and Eaton (2009) included males aged 15-54 whereas we only used those aged 15-49, in line with the 2004 MDHS report). The district-level HIV refusal map for MDHS shown in Figure 1 indicates higher rates in the central and southern parts of Malawi. There is high variation in the refusal rates across the districts.

4.2 Adjustment of HIV prevalence estimates

We apply various estimators considered in the previous section to MDHS, ANC, and MDICP data. A summary of the results is given in Table 3. For each estimator, we obtain separate HIV prevalence estimates for women and men. The estimates are then combined to derive overall estimates. In deriving these estimates, sampling weights need to be considered. The 2004 MDHS reportNational Statistical Office and ORC Macro (2005) (Tables 12.5, Appendix G.1, and page 452) uses sampling weights for calculating HIV prevalence and adjusted rates. These sampling weights are made up of three types: (1) HIV sampling weights for those who are tested; (2) individual sampling weights for those interviewed but not tested; and (3) household sampling weights for those who are not interviewed and not tested. In Reniers and Eaton, Reniers and Eaton (2009) the sampling weighting scheme of the 2004 MDHS report was applied to the MDHS data but no weights (except by the subgroup proportion of the population) were applied to the MDICP data. Sampling weights do not apply to ANC data since they come from women who visited ANC sites. To facilitate comparison to earlier results, we follow the same strategy as earlier studies in handling sampling weights for the MDHS data and MDICP data. For the ANC data, data are weighted by their proportional representation from census. We return to the discussion of sampling weights and their relationship to refusal bias subsequently.

There are 6696 individuals eligible for HIV testing in our MDHS sample. Out of these individuals, 1601 individuals (886 women and 715 men) expressly refuse to take an HIV test. Among the remaining 5095 individuals, 647 individuals (418 women and 229 men) are found to be HIV positive and 4448 individuals (2408 women and 2040 men) are HIV negative, giving an overall unweighted HIV prevalence of $647/5095 \approx 0.1270$. All subsequent analyses are, however, based on weighted cases, as described earlier. The complete case estimate of HIV prevalence $\hat{\pi}_{CC}$ in women is 0.1347. Similarly, the complete case prevalence estimate for men is approximately 0.1029. The overall estimate combining the women and men estimates is about 0.1194. Compared to $\hat{\pi}_{CC}$, the estimator $\hat{\pi}_{MSI}$ uses additional information from those who do not take an HIV test. For the prediction of HIV status, we use the same set of covariates as those in the MDHS 2004 report, Appendix G, (National Statistical Office and ORC Macro, 2005) that includes both demographic as well as behavioural variables: age, wealth index, education, geographical region, rural/urban residence, age at first sex, work status, marital status, smoking/tobacco use, media exposure, religion, STI or STI symptoms, condom use, higher-risk sex in the last year (sex with a non-marital, non-cohabiting partner), test for AIDS, number of sexual partners in the last 12 months, sexually transmitted disease (STD) in the last year, and willingness to care for a relative with AIDS. Separate logistic regressions are carried out for women and men. The model is then applied to impute HIV status for those who refuse an HIV test. Using this procedure, the prevalence estimates for women and men are 0.1385 and 0.1154, respectively.

The inverse probability estimator $\hat{\pi}_1$ assumes acceptance of HIV testing may be non-random and that

Table 3: HIV prevalence estimates using MDHS, ANC and MDICP data

Estimator	Men	Women	Overall
$\hat{\pi}_{CC}$	0.1029	0.1347	0.1194
$\hat{\pi}_{MSI}$	0.1154	0.1385	0.1274
$\hat{\pi}_1$	0.1118	0.1368	0.1247
$\hat{\pi}_2$	0.0992	0.1319	0.1165
$\hat{\pi}_{RE}$	0.1130	0.1470	0.1306
$\hat{\pi}_{3-}$	0.0935	0.1174	0.1059
$\hat{\pi}_{3+}$	0.1183	0.1556	0.1376
$\hat{\pi}_4$	—	$0.1550^{\dagger}$	—
$\hat{\pi}^{\ddagger}_{5A}$	0.1144	0.1377	0.1265
$\hat{\pi}^{\dagger\dagger}_{5B}$	0.1150	0.1397	0.1278

† Based only on pregnant females in the ANC survey
‡Stepwise regression using covariates, X_i and $\hat{\pi}^c_{ANC}$
††Fixed regression using $\hat{\pi}^c_{ANC}$ only

the probability of acceptance can be captured by some observable covariates. We use the same list of covariates from the MDHS 2004 report for estimating the propensity score for acceptance of HIV testing. Due to some individuals with no information on some of the covariates, the model for men includes only 2304 observations from MDHS, as opposed to the entire sample of 2984 men. Out of the 2304 men, 1759 men accepted an HIV test with a weighted average acceptance rate of 0.835, but the interquartile range of the estimated propensity score is from 0.840 to 0.962. Similarly for women, the model is based on 2623 women instead of the entire sample of 3712 women. Out of these 2623 women, 2019 women accepted an HIV test with a weighted average acceptance rate of 0.747, but the interquartile range of the estimated propensity score is 0.813 to 0.932. So for both men and women, the estimated propensity scores are somewhat different from to their respective means, and $\hat{\pi}_1$ accounts for such differences by adjusting the complete case estimates. Indeed, for women and men, the values of $\hat{\pi}_1$ are 0.1368 and 0.1118, respectively, slightly higher than their complete case counterparts.

Out of the 6696 individuals in our MDHS sample, 5816 report that they do not have a prior HIV test. These individuals form the basis for calculating $\hat{\pi}_2$. Among women who do not have a prior HIV test, 359 have a positive HIV test result while 2138 are HIV negative, giving a weighted HIV prevalence estimate of 0.1319, and the corresponding estimate for men is 0.0992.

A total of 2874 individuals (1539 females and 1335 males) consent to an HIV test and provide complete information for analysis in MDICP-3. Of these individuals, 1996 consent to an HIV test in MDICP-4 and 115 refuse, while the HIV status for the rest is missing for other reasons. Among those individuals who are tested in MDICP-3, 185 (111 females and 74 males) are HIV positive and 2689 (1428 females and 1261 males) are HIV negative.

We repeat the analysis of Reniers and Eaton Reniers and Eaton (2009) using our data. Since we exclude males aged 50-54 years from the MDHS data whereas Reniers and Eaton included them, we do not expect the two sets of estimates to be identical. To compute the estimate using $\hat{\pi}_{RE}$, we need to know

whether an individual has taken the first-round HIV test (MDICP-3), whether the individual knows the test result, the actual test result, and the refusal of the second-round HIV test conducted in MDICP-4. The $\hat{\pi}_{RE}$ estimates for males and females are 0.1130 and 0.1470, respectively, and the combined overall estimate is 0.1306, which is quite similar to the figure of 0.132 in Reniers and Eaton (Table 2). Reniers and Eaton (2009) The same set of data is also used to find $\hat{\pi}_{3-}$ and $\hat{\pi}_{3+}$. The bounds for men are 0.0935 and 0.1183, and for women, they are somewhat wider at 0.1174 and 0.1556, respectively.

To implement the estimator $\hat{\pi}_4$, we first extract the number of ANC attendees and HIV positive cases from the 19 sentinel sites in the 2003 ANC data. (National AIDS Commission, 2003a) The site-specific numbers are then used to represent the HIV prevalence in the rural and urban areas in each of the 28 districts defined in the 2003 ANC Technical Report (Table 2). (National AIDS Commission, 2003b) The resulting rural HIV rates in the 28 districts range from 0.0969 to 0.2315 with a mean of 0.1349 while the urban rates range from 0.0993 to 0.3288 with a mean of 0.2010. Finally, the district-area numbers are weighted by the population size from the 1998 Census data (IPUMS, University Minnesota and Malawi National Statistical Office, 1998 Population and Housing Census) to give an overall HIV prevalence estimate of 0.1550. Since the ANC data is based on pregnant women only, only one HIV prevalence estimate is obtained. Estimates using ANC survey data have been used as indicators for national HIV trends. Kigadye et al. (1993), Fylkesnes et al. (1998), Glynn et al. (2001), Asamoah-Odei et al. (2004)

The estimator $\hat{\pi}_5$ allows refusal to be dependent on the (unobservable) HIV status (for those who refuse testing). To model the propensity score function for (non)-refusal, we impute the unobservable HIV status with HIV prevalence estimates from the ANC data. The ANC prevalence estimates are obtained for different district-areas; for each individual who resides in a particular district-area, his/her HIV status is imputed by $\hat{\pi}^c_{ANC}$.

We consider two estimates based on $\hat{\pi}_5$. The first one, $\hat{\pi}_{5A}$, uses $\hat{\pi}^c_{ANC}$ and a stepwise regression procedure to select from the same list of covariates used in $\hat{\pi}_1$ to model the propensity score. The second one uses only $\hat{\pi}^c_{ANC}$ for modelling the propensity score. These estimated propensity scores are then used in $\hat{\pi}_{5B}$ to give different prevalence estimates.

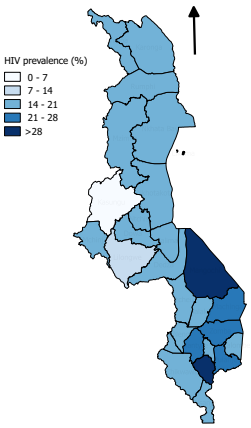
Using $\hat{\pi}^c_{ANC}$ and a selection of other covariates to model the propensity score, the corresponding HIV prevalence estimates, $\hat{\pi}_{5A}$, for women and men are 0.1377 and 0.1144, respectively. When the propensity score is modelled only with $\hat{\pi}^c_{ANC}$, the corresponding HIV prevalence estimates, $\hat{\pi}_{5B}$ for women and men are 0.1397 and 0.1150, respectively.

Table 4 gives the district-level estimates of HIV prevalence estimates using various methods discussed in this paper. There is high variation in HIV prevalence estimates across districts of Malawi, with values ranging from around 5% in Kasungu to as much as 25% in Blantyre. HIV prevalence estimated by $\hat{\pi}_1$, $\hat{\pi}_{5A}$ and $\hat{\pi}_{5B}$ are very similar; in most districts, these estimators give higher values than $\hat{\pi}_{CC}$. On the other hand, $\hat{\pi}_2$ is similar to $\hat{\pi}_{CC}$ in most districts. District-level HIV prevalence rates for urban and rural

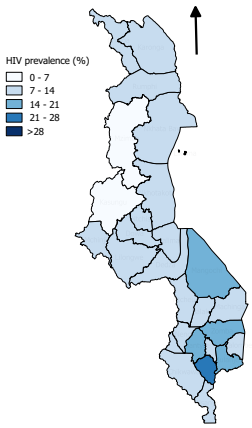
Table 4: District-level HIV prevalence estimates various methods.

District	$\hat{\pi}_{CC}$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}^{\dagger}_{5A}$	$\hat{\pi}^{\ddagger}_{5B}$
Blantyre	0.2234	0.2538	0.2140	0.2561	0.2561
Kasungu	0.0418	0.0478	0.0442	0.0481	0.0482
Machinga	0.1159	0.1108	0.1037	0.1093	0.1108
Mangochi	0.2118	0.2275	0.2024	0.2350	0.2349
Mzimba	0.0523	0.0603	0.0497	0.0585	0.0592
Salima	0.0876	0.0706	0.0844	0.0737	0.0737
Thyolo	0.2150	0.2301	0.2203	0.2343	0.2346
Zomba	0.1780	0.1820	0.1683	0.1817	0.1817
Mulanje	0.1969	0.1986	0.1946	0.2003	0.1993
Lilongwe	0.0375	0.0255	0.0362	0.0349	0.0350
Other districts	0.1093	0.1093	0.1096	0.1106	0.1106

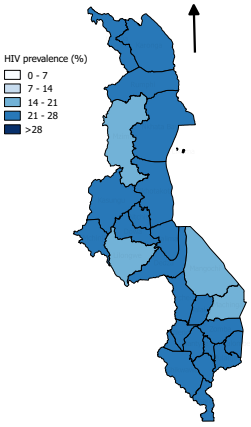
†Stepwise regression using X_i and $\hat{\pi}^c_{ANC}$
‡Fixed regression using $\hat{\pi}^c_{ANC}$ only



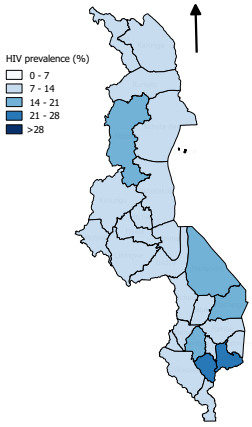
(a) MDHS: Urban



(b) MDHS: Rural



(c) ANC: Urban



(d) ANC: Rural

Figure 2: Estimated HIV prevalence rates. (a) Complete case estimates using Urban MDHS data. (b) Complete case estimates using Rural MDHS data. (c) District-area estimates using Urban ANC data. (d) District-area estimates using Rural ANC data.

areas directly calculated from MDHS and ANC data are presented in Figure 2. In both data sources, HIV prevalence rates are higher in the urban areas than the rural areas.

5 Discussion

This study explored several methods for adjusting refusal bias in HIV prevalence estimates in population-based surveys. It also conducted a thorough investigation of robustness against non-response bias. Compared to the naïve complete case estimator $\hat{\pi}_{CC}$, all point estimators except $\hat{\pi}_2$ give higher adjusted estimates for both men and women (and overall). These results are consistent with those observed in earlier studies. National Statistical Office and ORC Macro (2005), Mishra et al. (2008), Reniers and Eaton (2009), Obare et al. (2009)

Recall that for $\hat{\pi}_2$, the key assumptions are eqs.(6)-(7), which essentially mean that $\hat{\pi}_2$ is a type of complete case estimator applied to those who had never been tested before the 2004 MDHS survey. Hence, it is not surprising that the $\hat{\pi}_2$ estimates are not too different from the naïve $\hat{\pi}_{CC}$ estimates. Both estimators implicitly assume missing completely at random. In the case of $\hat{\pi}_{CC}$, the observed data is considered a random sample of the population. In the case of $\hat{\pi}_2$, the subsample of those with no prior HIV test and who accepted HIV test form a random sample.

Using the remaining methods, the prevalence for men is consistently adjusted upwards (from the complete case estimate) by about one percentage point, irrespective of the method used.

The case for women is somewhat different. The adjustment is method dependent. The results can be broadly classified into three groups, based on the methods used. The first group of methods, which includes $\hat{\pi}_{MSI}$, $\hat{\pi}_1$ and $\hat{\pi}_{5A}$, $\hat{\pi}_{5B}$, uses covariates to model the missing HIV test results (or the propensity that HIV test results are observed). Their results are all quite similar, all give an upward adjustment of HIV prevalence of around 0.5% from the complete case estimate. These methods are related in the sense that they are premised on the HIV status (and hence propensity to accept HIV test) can be modelled using observable demographic and behavioural covariates. Therefore, the methods would not be effective if these covariates have low predictive powers. A multi-country study of bias in HIV estimates from DHS Mishra et al. (2008) found that HIV prevalence is not strongly related to observable covariates.

The methods that combine the MDHS data and MDICP data ($\hat{\pi}_{RE}$, $\hat{\pi}_3$) suggest upward adjustments of about one percentage point. Compared to the complete case estimator, the estimator $\hat{\pi}_{RE}$ adjusts the prevalence of women upwards by 1.3 percent. Reniers and Eaton Reniers and Eaton (2009) found that, compared to those who accept an HIV test, individuals who refuse an HIV test are more than 4.5 times as likely to be HIV positive and hence, the upward adjustment is reasonable based on this fact. On the other hand, $\hat{\pi}_2$, while using the same assumptions as $\hat{\pi}_{RE}$, does not give an upward adjustment of the complete case rates (either men, women or overall). This raises the question of why they are different.

Comparing eq.(8) to eq.(10), we notice that the latter ignores those who refused to be tested (see above for the complete case interpretation of $\hat{\pi}_2$) while the former explicitly estimates the missing HIV status using MDICP data. Hence, $\hat{\pi}_{RE}$ is more similar to a MSI or imputation approach. Pepe et al. (1994), Chen (2000) Naturally, if we assume that eq. (6) and eq. (7) hold and that the MDICP data can be used to replace the missing MDHS data, $\hat{\pi}_{RE}$ uses additional covariate information from observations with missing HIV status, hence more accurate than $\hat{\pi}_2$.

Another method that also uses the MDICP data is $\hat{\pi}_3$. We observe the lower and upper bounds for the HIV prevalence are fairly tight around the complete case estimates. Since these bounds are created with very mild assumptions, the fact that they are very close to the complete case estimates suggests that the refusal bias in the MDHS estimates may be quite small. Between men and women, the bounds for women are much wider. In particular, the upper bound for women is over two percentage points above that of the complete case estimate for women. This result is consistent with the behaviour of $\hat{\pi}_{RE}$, which adjusts the estimate for women upwards.

The third group is the method that uses the ANC data. The ANC survey provides a single prevalence estimate ($\hat{\pi}_4$) for women, and is significantly higher than most of the prevalence rates from other methods. This result is not surprising since ANC surveys only capture data from pregnant women in more urbanised areas who choose to go to an antenatal clinic during their pregnancy and have rates higher than the national average. There are indeed some evidence that applying ANC prevalence directly to give population prevalence estimates leads to biases. (Zaba et al., 2000, Gregson et al., 2002, Gouws et al., 2008) Nevertheless, ANC prevalence does reflect the actual but unknown prevalence within each district-area and is free of refusal (or other kinds of non-response) bias.

6 Conclusion and implication for future research

The motivation for our paper is to provide a coherent and comprehensive conceptual framework for studying survey data with non-response due to refusals. We revisited some existing methods and also introduced new ones. Our paper offers a novel approach to the challenges that refusals create and proposes possible solutions for them. We compared various methods, clarifying their underlying assumptions, implications, and important data requirements. The approach offered in this paper is especially useful for practitioners in charge of planning and analysis. The primary application of our approach is the estimation of HIV prevalence particularly in Africa, where HIV/AIDS remains epidemic or endemic. Our approach is also applicable to other issues and areas with similar challenges.

Longitudinal surveys are still uncommon in many parts of the developing world, since they are difficult to implement and the quality of data from such surveys is often poor because of the difficulty with tracking mobile populations. While longitudinal studies are still relatively rare, the availability of nationally

representative longitudinal studies is on the rise in developing countries. One of our contributions lies in proposing ways to meaningfully bring together the following three very different three types of data: MDHS, ANC, and MDICP. We show how these data can be combined when none of them can allow us to reliably estimate HIV prevalence in Malawi on their own.

A common approach for adjusting (refusal) bias in surveys is by weighting. Methods such as $\hat{\pi}_1$ in this paper, whether using sampling weights, or weights based on fitting a propensity function, use this approach. This approach works only if refusal is independent of the outcome, given the covariates that are used to model the propensity function. In the missing data literature, this condition is called missing at random. However, it can never be confirmed whether the missing at random assumption actually holds. We considered alternative methods to solve this problem, by exploiting information from auxiliary surveys. Using the assumptions of Reniers and Eaton, Reniers and Eaton (2009) we identified a new method ($\hat{\pi}_2$) using only MDHS data. The method uses data from those who have never been tested and do not know their HIV status, and hence, their decision to accept a HIV test is arguably less susceptible to bias.

Further, we introduced a "bound" approach using data from MDICP, by which we estimated the plausible lower and upper bounds ($\hat{\pi}_{3-}, \hat{\pi}_{3+}$) of the prevalence based on a set of weak and reasonable assumptions. This approach is potentially useful because it is often difficult to validate or falsify an underlying assumption. Furthermore, it shows that a carefully designed and implemented localised study may also be helpful for understanding the magnitude of non-response bias.

We also proposed two different methods using the ANC data. The first method ($\hat{\pi}_4$) uses summary statistics from antenatal care units and combines them with census data to obtain prevalence estimates. An advantage of this approach is that no micro-data is needed and therefore the method can be implemented easily. The second method ($\hat{\pi}_5$) combines the MDHS data with the ANC data to produce prevalence estimates. The novel feature of this method is the use of weights based on ANC data that adjust for non-ignorable missingness. Since ANC surveys are relatively free from refusal bias and are carried out at more frequent intervals than DHS, these two methods offer the possibility of obtaining prevalence estimates on a more contemporaneous basis.

In the presence of non-responses, all analytic methods require some assumptions and it is hard to determine what method is best. However, when there are available alternative methods, a way to go about addressing the refusal bias problem is to use all methods and compare their results. In the current study, the prevalence estimates range from 0.0935 to 0.1183 for men, from 0.1174 to 0.1556 for women, and 0.1059 to 0.1376 overall (See Table 3, last column). The relatively narrow range for men tells us that the refusal bias, if it exists at all, is practically not a major issue. The refusal bias for women may be larger but it is still small in absolute value and would be no larger than 3%. As these results indicate, the range reflects (the lack of) limits to which we can place our confidence in our results.

Our findings of acceptable level of refusal bias in the Malawi prevalence estimates can be contrasted from that reported in Obare, Obare (2010) where substantial potential bias is attributed to refusal/absence using the MDICP data. In that report, the percentage of HIV positive is 4.4 among those who accept an HIV test in both MDICP-3 and MDICP-4, compared to 15.5 and 13.0, respectively, for those who refuse or are absent for the test in MDICP-4. However, using our own analysis, we found this difference is due largely to those who already know their HIV test results from MDICP-3. Among those who do not know the results of the first-round HIV test, the proportion of people who refuse is similar between HIV-positive and HIV-negative individuals. Similarly, among those who know the results of the first-round HIV test, the proportion of people who refuse is substantially higher for HIV-positives than HIV-negatives. We may argue that a person who knows his/her HIV positive status is more likely to decline a second test because HIV positive status cannot be changed and the person may feel another test is meaningless. In our paper, the estimates $\hat{\pi}_{RE}$ and $\hat{\pi}_2$ are calculated using those who do not know their HIV status, whereas the bounds $\hat{\pi}_{3-}$ and $\hat{\pi}_{3+}$ explicitly allow for differences in refusal rates between those who know and those do not know their HIV status under a set of weak assumptions. The ANC surveys can be assumed to be free from refusal bias, and $\hat{\pi}_4$ uses this assumption to come up with refusal bias-free prevalence estimates; for $\hat{\pi}_5$, the ANC data is used indirectly to create weights that adjust for refusals. None of the methods considered in this paper show a large upward adjustment from the weighted estimate $\hat{\pi}_1$ and the unadjusted estimate $\hat{\pi}_{CC}$.

References

Global HIV/AIDS Response: Epidemic Update and Health Sector Progress Towards Universal Access: Progress Report 2011. WHO, Geneva, 2011.

J. M. Alho. Adjusting for nonresponse bias using logistic regression. *Biometrika*, 77:617–624, 1990.

E. Asamoah-Odei, J. M. Garcia Calleja, and J. T. Boerma. HIV prevalence and trends in sub-Saharan Africa: no decline and large subregional differences. *Lancet*, 364:35–40, 2004.

J. Billet, M. Philippens, R. Fitzgerald, and I. Stoop. Estimation of nonresponse bias in the European Social Survey: Using information from reluctant respondents. *Journal of Official Statistics*, 23:135–162, 2007.

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.

J. Burton, H. Laurie, and P. Lynn. The long-term effectiveness of refusal conversion procedure on longitudinal surveys. *Journal of the Royal Statistical Society, A*, 169:459–478, 2006.

Y-H Chen. A robust imputation method for surrogate outcome data. *Biometrika*, 87(3):711–716, 09 2000.

G. B. Durrant and F. Steele. Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society, A*, 172:361–381, 2009.

S. Floyd, A. Molesworth, A. Dube, A. C. Crampin, R. Houben, M. Chihana, A. Price, N. Kayuni, J. Saul, N. French, and J. R. Glynn. Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. *AIDS*, 27:233–242, 2013.

K. Fylkesnes, Z. Ndhlovu, K. Kasumba, R. M. Musonda, and M. Sichone. Studying dynamics of the HIV epidemic: population-based data compared with sentinel surveillance in Zambia. *AIDS*, 12:1227–1242, 1998.

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.

Buvé A. Caraël M. Glynn, J. R., R. M. Musonda, M. Kahindo, I. Macauley, F. Tembo, L. Zekeng, and Study Group on Heterogeneity of HIV Epidemics in African Cities. Factors influencing the difference in HIV prevalence between antenatal clinic and general population in sub-Saharan Africa. *AIDS*, 15: 1717–1725, 2001.

E. Gouws, V. Mishra, and T. B. Fowler. Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data. *Sexually Transmitted Infections*, 84 (suppl 1):i17–i23, 2008.

S. Gregson, N. Terceira, M. Kakowa, P. R. Mason, R. M. Anderson, S.K. Chandiwana, and M. Caraël. Study of bias in antenatal clinic HIV-1 surveillance data in a high contraceptive prevalence population in sub-Saharan Africa. *AIDS*, 16:643–652, 2002.

R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little. *Survey Nonresponse*. Chichester: Wiley, 2002.

D. Hawkes and I. Plewis. Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society, A*, 169:479–491, 2006.

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

D. R. Hogan, J. A. Salomon, D. Canning, J. K. Hammitt, A. M. Zaslavsky, and T. Bärnighausen. National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models. *Sexually Transmitted Infections*, 88 (Suppl 2):i17–i23, 2012.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

R. M. Kigadye, A. Klokke, A. Nicoll, K. M. Nyamuryekung'e, M. Borgdorff, L. Barongo, U. Laukamm-Josten, F. Lisekie, H. Grosskurth, and F. Kigadye. Sentinel surveillance for HIV-1 among pregnant women in a developing country: 3 years experience and comparison with a population serosurvey. *AIDS*, 7:849–855, 1993.

F. Kreuter, G. Müller, and M. Trappmann. Nonresponse and measurement error in employment research: Making use of administrative data. *The Public Opinion Quarterly*, 74:880–906, 2010.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. New York: Wiley, 2002.

P. Lynn. Non-response biases in surveys of schoolchildren: the case of the English Programme for International Student Assessment (PISA) samples. *Journal of the Royal Statistical Society, A*, 175: 915–938, 2012.

P. Lynn and P. Clarke. Separating refusal bias and non-contact bias: evidence from UK national surveys. *The Statistician*, 51:319–333, 2002.

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.

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.

L. S. Montana, V. Mishra, and R. Hong. Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa. *Sexually Transmitted Infections*, 84 (Suppl 1):i78–i84, 2008.

National AIDS Commission. *HIV Sentinel Surveillance Report*. Ministry of Health and Population, Malawi, 2003a.

National AIDS Commission. *Estimating National HIV Prevalence in Malawi from Sentinel Surveillance Data: Technical Report*. Ministry of Health and Population, Malawi, 2003b.

National AIDS Commission. *HIV and Syphilis Sero-Survey and National HIV Prevalence and AIDS Estimates Report for 2007*. Ministry of Health, Malawi, 2008.

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

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

F. Obare, P. Fleming, P. Anglewicz, R. Thornton, F. Martinson, A. Kapatuka, M. Poulin, S. Watkins, and H.-P. Kohler. Acceptance of repeat population-based voluntary counseling and testing for HIV in rural Malawi. *Sexually Transmitted Infections*, 85:139–44, 2009.

K. Olson. Do non-response follow-ups improve or reduce data quality?: a review of the existing literature. *Journal of the Royal Statistical Society, A*, 176:129–145, 2013.

M. S. Pepe, M. Reilly, and T. R. Fleming. Auxiliary outcome data and the mean-score method. *Journal of Statistical Planning and Inference*, 42:137–160, 1994.

G. Reniers and J. Eaton. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS*, 23:1–9, 2009.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

I. A. Stoop. Survey nonrespondents. *Field Methods*, 16:23–54, 2004.

The POLICY Project. *Estimating National HIV Prevalence in Malawi from Sentinel Surveillance Data*. The National AIDS Control Programme, Lilongwe, Malawi, 2001.

I. Thomsen and A. M. K. Holmøy. Combining data from surveys and administrative record systems: The Norwegian experience. *International Statistical Review*, 66:201–221, 1998.

G. J. van den Berg, M. Lindeboom, and P. J. Dolton. Survey non-response and the duration of unemployment. *Journal of the Royal Statistical Society, A*, 169:585–604, 2006.

R. M. Yucel and A. M. Zaslavsky. Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association*, 100:1123–1132, 2005.

B. W. Zaba, L. M. Carpenter, J. T. Boerma, S. Gregson, J. Nakiyingi, and M. Urassa. Adjusting antenatal clinic data for improved estimates of HIV prevalence among women in sub-Saharan Africa. *AIDS*, 14:2741–2750, 2000.

E. Zanutto and A. Zaslavsky. Using administrative records to impute for nonresponse. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, editors, *Survey Non-response*. Wiley, Chichester, 2002.

Supplementary material

Adjustment of HIV prevalence estimates (excluding Lilongwe)

In the main article, we carried out analyses using MDHS data from all districts in Malawi. In 2004 MDHS, Lilongwe district has an unusually high refusal rate (54%) and low observed prevalence. (National Statistical Office and ORC Macro, 2005) Here, we give results of a parallel set of analyses after removing Lilongwe from the MDHS data. After removing Lilongwe, the number of individuals in the MDHS data becomes 6287. The refusal rate is around 21.9% in MDHS, which remains considerably higher than the other two sources. For MDHS, the refusal rate for men is $610/2784 \approx 0.22$ and for women is $768/3503 \approx 0.22$ in the MDHS data after excluding Lilongwe.

We apply various estimators considered in the main article to MDHS, ANC, and MDICP data. A summary of the results is given in Table 5. For each estimator, we obtain separate HIV prevalence estimates for women and men. The estimates are then combined to derive overall estimates. We use the sampling weighting scheme described in the main article.

Table 5: HIV prevalence estimates using MDHS, ANC and MDICP data

Estimator	Men	Women	Overall
$\hat{\pi}_{CC}$	0.1120	0.1522	0.1332
$\hat{\pi}_{MSI}$	0.1294	0.1558	0.1433
$\hat{\pi}_1$	0.1296	0.1559	0.1434
$\hat{\pi}_2$	0.1070	0.1491	0.1294
$\hat{\pi}_{RE}$	0.1210	0.1603	0.1417
$\hat{\pi}_{3-}$	0.1026	0.1341	0.1192
$\hat{\pi}_{3+}$	0.1282	0.1714	0.1510
$\hat{\pi}_4$	—	$0.1550^{\dagger}$	—
$\hat{\pi}^{\ddagger}_{5A}$	0.1308	0.1570	0.1387
$\hat{\pi}^{\dagger\dagger}_{5B}$	0.1310	0.1573	0.1449

† Based only on pregnant females in the ANC survey
‡ Stepwise regression using covariates, X_i and $\hat{\pi}^c_{ANC}$
†† Fixed regression using $\hat{\pi}^c_{ANC}$ only

Among the 4909 individuals who took HIV test, 638 (416 women and 222 men) are found to be HIV positive while 4271 (2319 women and 1952 men) are HIV negative. The (weighted) complete case estimate $\hat{\pi}_{CC}$ of HIV prevalence in women is 0.1522, and that for men is 0.1120. The overall complete case prevalence estimate is 0.1332. Other estimates are also derived in the same way as the main article, except that Lilongwe is excluded from the MDHS sample. Note that the estimate using $\hat{\pi}_4$ is identical to that in the main text as it does not depend on MDHS data.

Comparing the results here to those in the main article, where we have included Lilongwe in the MDHS data, two observations emerge. First, for both men and women, the HIV prevalence estimates becomes higher once Lilongwe is excluded. This pattern is observed for all methods considered except for $\hat{\pi}_4$, which remains unchanged as it only uses the ANC data. Second, the exclusion of Lilongwe leads

to a higher increase in the estimated prevalence across all methods except for $\hat{\pi}_4$. As pointed out earlier, the observed prevalence for Lilongwe is unusually low and hence, including data from Lilongwe would place a downward bias on HIV prevalence. Furthermore, even though the refusal rates for Lilongwe men and women are similar ($105/200 \approx 53\%$ and $118/209 \approx 56\%$, respectively), among those who accept an HIV test, the observed HIV rates for men and women are quite different, $7/95 \approx 7.4\%$ and $2/91 \approx 2.2\%$, respectively. Not only the observed HIV prevalence rates are low, but more importantly, the rate for women is much *lower* than that for men. These results run counter to the well established thesis that HIV prevalence for women is higher in men. Hence, by removing these counter-intuitive results from the analysis, the exclusion of Lilongwe affects women's rates more than men's rates.