# National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models

Daniel R Hogan,[1,2] Joshua A Salomon,[1,2] David Canning,[2] James K Hammitt,[3,4] Alan M Zaslavsky,[5] Till Bärnighausen[2,6]

▸ Additional data are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/sextrans-2012-050636).

[1]Center for Health Decision Science, Harvard School of Public Health, Boston, Massachusetts, USA
[2]Department of Global Health and Population, Harvard School of Public Health, Boston, Massachusetts, USA
[3]Center for Risk Analysis, Harvard University, Boston, Massachusetts, USA
[4]Toulouse School of Economics (LERNA-INRA), Toulouse, France
[5]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, USA
[6]Africa Centre for Health and Population Studies, University of KwaZulu-Natal, Mtubatuba, South Africa

**Correspondence to**
Dr Daniel R Hogan, Harvard School of Public Health, Department of Global Health and Population, 665 Huntington Ave, Building 1, Room 1104, Boston, MA 02115, USA; dhogan@hsph.harvard.edu

**UNAIDS Report 2012 Guest Editors**
Karen Stanecki
Peter D Ghys
Geoff P Garnett
Catherine Mercer

Accepted 18 August 2012

## ABSTRACT

**Objectives** Population-based HIV testing surveys have become central to deriving estimates of national HIV prevalence in sub-Saharan Africa. However, limited participation in these surveys can lead to selection bias. We control for selection bias in national HIV prevalence estimates using a novel approach, which unlike conventional imputation can account for selection on unobserved factors.

**Methods** For 12 Demographic and Health Surveys conducted from 2001 to 2009 (N=138 300), we predict HIV status among those missing a valid HIV test with Heckman-type selection models, which allow for correlation between infection status and participation in survey HIV testing. We compare these estimates with conventional ones and introduce a simulation procedure that incorporates regression model parameter uncertainty into confidence intervals.

**Results** Selection model point estimates of national HIV prevalence were greater than unadjusted estimates for 10 of 12 surveys for men and 11 of 12 surveys for women, and were also greater than the majority of estimates obtained from conventional imputation, with significantly higher HIV prevalence estimates for men in Cote d'Ivoire 2005, Mali 2006 and Zambia 2007. Accounting for selective non-participation yielded 95% confidence intervals around HIV prevalence estimates that are wider than those obtained with conventional imputation by an average factor of 4.5.

**Conclusions** Our analysis indicates that national HIV prevalence estimates for many countries in sub-Saharan African are more uncertain than previously thought, and may be underestimated in several cases, underscoring the need for increasing participation in HIV surveys. Heckman-type selection models should be included in the set of tools used for routine estimation of HIV prevalence.

## INTRODUCTION

Accurate estimates of HIV prevalence are critical for tracking the epidemic, designing and evaluating prevention and treatment programmes, and estimating resource needs.[1–6] In sub-Saharan Africa, home to about two-thirds of the worldwide 33 million people living with HIV,[1] national population-based surveys[7–9] have become an essential data source for estimating HIV prevalence in many countries.[10–12] A potential threat to the validity of survey-based prevalence estimates is that not all individuals eligible to participate in a survey can be contacted, and some who are contacted do not consent to HIV testing. Incomplete participation in testing can lead to selection bias, and a recent paper found evidence for substantial downward bias in existing national HIV prevalence estimates for Zambian men due to selective survey non-participation.[13] The evaluation of possible bias in HIV prevalence estimates for other countries in sub-Saharan Africa is thus important for HIV research and policy.

Previous authors have suggested that non-participation may lead to bias in HIV prevalence estimates,[10 14 15] but official estimates of HIV prevalence in sub-Saharan Africa rely heavily on population-based surveys, which often have low participation rates.[1] An analysis of the Demographic and Health Surveys (DHS), which are the most common nationally representative surveys for HIV prevalence in sub-Saharan Africa, reveals average rates of non-participation in HIV testing of 23% for adult men and 16% for adult women in the region, with a high of 37% for men in Zimbabwe 2005–2006 and a low of 3% for women in Rwanda 2005,[16] and the most recent national population-based survey in South Africa reported an overall non-participation rate of 32% for HIV testing among adults.[7] Analyses of the DHS have adjusted HIV prevalence estimates for testing non-participation by imputing missing HIV test results with probit regressions, controlling for differences in observed characteristics between testing participants and non-participants, such as gender, urban residence, wealth and indicators of sexual behaviour, as recommended by WHO.[16–18] Based on this conventional imputation approach, non-participants were estimated to have higher HIV prevalence than participants in about half of the DHS examined, but this did not result in substantially different estimates of overall HIV prevalence when compared with the complete-case estimates that ignored missing observations.[16] These results have been interpreted to mean that non-participation in HIV testing surveys is likely to have minimal impact on prevalence estimates.[16 19] However, the conventional imputation approach has two important limitations. First, it assumes that no unobserved variables associated with HIV status influence participation in HIV testing. Second, it ignores regression parameter uncertainty in the imputation model, resulting in confidence intervals (CI) that are too small.

The first limitation of conventional imputation is that non-participants are assumed to be 'missing at random', implying that the expected HIV status of non-participants is the same as that for

participants with the same measured covariates.[20] However, if any unobserved variable is correlated with testing and HIV status, this condition will be violated. In particular, HIV status itself may influence participation.[15 21] Individuals who know that they are HIV-positive (because they have tested in the past) may fear stigma, exclusion or abuse if others learn about their HIV status.[22 23] Individuals who suspect that they are HIV-positive (eg, based on past sexual behaviour) may fear confirmation of their suspicions.[24] The limited available empirical evidence supports the hypothesis that HIV status correlates with participation. A longitudinal study in Malawi showed that among persons aware of a previous HIV test result those who had tested HIV-positive were 4.6 times less likely to consent to a new HIV test than those who had tested HIV-negative.[15] In South Africa, a population-based, longitudinal study found that HIV-positive individuals were substantially less likely to consent to an HIV test than HIV-negative individuals, and that among HIV-positive individuals those who certainly knew their status were least likely to participate in testing.[21]

To address these issues, Bärnighausen *et al*[13] estimated HIV prevalence in the Zambian 2007 DHS with a Heckman-type selection model. This approach can control for correlation between HIV status and HIV testing participation that remains after selection on observed characteristics has been taken into account. The national HIV prevalence estimate in adult Zambian men was 21% after correcting for selection on unobserved factors, compared with 12% in those with valid HIV tests or based on conventional imputation.

This study aims to derive adjusted estimates of national HIV prevalence in other sub-Saharan African countries using Heckman-type selection models to correct for selective non-participation in nationally representative surveys. It also employs a novel method for computing 95% CI around imputation-based HIV point estimates of prevalence that incorporates regression parameter uncertainty, which more accurately reflects the additional uncertainty introduced when imputing HIV status.

## METHODS

### Survey data

We examined data from 24 DHS (table 1).[25] A typical survey involved a two-stage sampling design stratified by region and urban versus rural setting.[25 26] Interviewing teams first completed a 'household' questionnaire with one household member to establish which household members were eligible for an 'individual' interview and for HIV testing. Members of the interviewing team then elicited informed consent for HIV testing from the eligible household members and conducted the tests. A typical survey team included a team leader, a field editor and 3–6 interviewers who were usually matched to the gender of eligible participants (table 1). In some surveys, health professionals travelled with teams to conduct HIV testing, while in others interviewers were trained to obtain consent and blood samples (table 1).

### Models to estimate HIV prevalence

We compared three strategies for handling missing HIV test results when estimating HIV prevalence from DHS data, following the analytic approach in Bärnighausen *et al*[13] and extending it to improve the computation of CI. These models included: (1) an unadjusted *complete-case* analysis in which missing observations are ignored and prevalence is calculated among those with valid HIV tests, (2) a *conventional imputation* approach that imputes missing HIV status conditional on observed covariates using a probit regression and (3) a Heckman-type *selection model* approach, which can correct for selection on unobserved factors when imputing HIV status for missing observations. Eligible individuals were missing valid HIV test results in the DHS for two main reasons: (1) the individual was successfully contacted but refused to consent to an HIV test or (2) the interview team failed to contact or interview the individual. For both conventional imputation and selection modelling approaches, we ran separate regressions to predict missing HIV status in either the 'non-consent' or 'non-contact' groups.

Although uncommon in the biomedical literature, Heckman-type selection models have been widely used for more than 3 decades in economics and other social sciences to estimate regression coefficients in the presence of missing data problems.[27 28] The selection model used in this analysis is a bivariate probit regression comprised of a selection equation that predicts HIV test participation and an outcome equation that predicts HIV status, linked through a correlation parameter, ρ, that reflects covariance between HIV status and testing participation, conditional on observed covariates.[13 27] A negative estimate of ρ implies that HIV-positive individuals were less likely to participate in HIV testing than HIV-negative individuals, all else being equal, and in this case the model will predict higher probabilities of being HIV-positive among non-participants. To improve the identification of the model, selection variables subject to an exclusion restriction are included in the selection equation. The exclusion restriction requires that the selection variables affect HIV testing participation but are not correlated with HIV status. We accounted for the complex survey design when estimating regression covariance matrices and used household sampling weights to obtain national representative prevalence estimates (see online technical appendix and reference 13 for details).

### Selection variables

We used the same selection variables as Bärnighausen *et al* to predict participation in HIV testing within Heckman-type selection models.[13] For individuals who completed an individual interview but refused to consent to an HIV test (*consent regressions*), the identity of the interviewer who conducted the individual questionnaire was chosen as the selection variable based on a long line of work in the survey sciences showing that interviewer characteristics (eg, motivation, extraversion, experience with HIV testing and attitudes about HIV research) can influence consent to testing.[29–31] The DHS surveys in this study varied in terms of which survey team members were responsible for obtaining consent and blood samples for HIV testing, which we have grouped into four categories (table 1).[25] For surveys that included interviewers who did not obtain consent and conduct testing, these interviewers could affect consent through their impact over the course of the lengthy individual interview on respondents' confidence in the survey process, or attitudes towards the survey team or participating in HIV research.

For individuals who were eligible to participate but could not be contacted or refused to be interviewed (*contact regressions*), the identity of the interviewer who conducted the household interview was chosen as one of two selection variables, as these interviewers may differ in their ability to obtain information on when the missing individual would return, in the frequency of their follow-up visits or their ability to obtain consent for the individual interview. We included a second selection variable, indicating whether or not the household was visited on the first day that a team conducted interviews in a

**Table 1** HIV testing strategies and personnel responsibilities in 24 Demographic and Health Surveys (DHS) as described in DHS survey reports, 2001–2009, with HIV testing participation rates for adult men and women.

| HIV testing strategy and personnel | Country | Year | Pr. HH* | No. of teams | No. of interviewers† | No. of testers‡ | % Participating§ | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Men | Women |
| (1) Consent on individual questionnaire; interviewers conducted HIV testing | Cote d'Ivoire | 2005 | 1/1 | 10 | 2F, 2M | – | 76 | 79 |
| | Malawi | 2004 | 1/3 | 22 | 4–5F, 1M | (2–3) | 63 | 70 |
| | Tanzania | 2003–2004 | 1/1 | 11 | 4F, 1M | – | 77 | 84 |
| | Tanzania | 2007–2008 | 1/1 | 14 | 4F, 1M | – | 80 | 90 |
| | Zimbabwe | 2005–2006 | 1/1 | 14 | 3–4F, 2–3M | – | 63 | 76 |
| (2) Consent on household questionnaire; interviewers conducted HIV testing | Lesotho | 2004 | 1/2 | 12 | 3F, 1M | – | 68 | 81 |
| | Liberia | 2007 | 1/1 | 19 | 2F, 2M | – | 81 | 88 |
| | Sierra Leone | 2008 | 1/2 | 24 | 2F, 1M | – | 87 | 90 |
| | Zambia | 2007 | 1/1 | 12 | 3F, 3M | – | 72 | 77 |
| (3) Consent on household questionnaire; subset of interviewers conducted HIV testing | Cameroon | 2004 | 1/2 | 14 | 3F, 1M | (≥2) | 90 | 92 |
| | Ethiopia | 2005 | 1/2 | 30 | 4F, 2M | (2) | 76 | 83 |
| | Mali | 2006 | 1/3 | 25 | 3 | (2) | 85 | 93 |
| | Niger | 2006 | 1/2 | 20 | 3F, 1M | (1) | 84 | 91 |
| | Senegal | 2005 | 1/3 | 15 | 3F, 1M | (2) | 75 | 84 |
| | Swaziland | 2006–2007 | 1/1 | 10 | 3–4F, 1–2 M | (2–3) | 78 | 87 |
| | Rwanda | 2005 | 1/2 | 15 | 3F, 1M | (2) | 96 | 97 |
| (4) Consent on household questionnaire; health worker or technician conducted HIV testing | Burkina Faso | 2003 | 1/3 | 12 | 3F, 1M | 1 | 86 | 92 |
| | Democratic Republic of Congo | 2007 | 1/2 | 234 | 1–3 | 1 | 86 | 90 |
| | Ghana | 2003 | 1/2 | 15 | 4 | 1 | 80 | 89 |
| | Guinea | 2005 | 1/2 | 10 | 4F, 1M | 1 | 88 | 92 |
| | Kenya | 2003 | 1/2 | 17 | 4F, 1M | 1 | 70 | 76 |
| | Kenya¶ | 2008–2009 | 1/2 | 23 | 4F, 2M | 2 | 79 | 86 |
| | Mali | 2001 | 1/3 | 25 | 3F | 1 | 76 | 85 |
| | Zambia | 2001 | 1/3 | 12 | 3–4F, 1M | 2 | 73 | 79 |

*Proportion of sampled households that were eligible for HIV testing and the men's individual questionnaire.
†Number of female and male interviewers per team. Team interviewer gender composition was not described in the reports for the Democratic Republic of Congo 2007, Mali 2006 and Ghana 2003 surveys.
‡Number of individuals who conducted HIV testing per team. Numbers in parenthesis indicate the number of interviewers on a team who also conducted HIV testing. The symbol '–' indicates that all interviewers conducted HIV testing.
§Percent participating in survey HIV testing.
¶Kenya 2008–2009 also had two voluntary counselling and testing counsellors on each team.
F, female; M, male; Pr. HH, Proportion of sampled households.

cluster, since households visited earlier would have more opportunities to be revisited in the event an eligible member was absent on the first visit.

A key assumption of our approach is that the identity of the survey interviewer and the day of the survey that a household is first visited correlate with testing but not with HIV status. We tested the statistical significance of the association between the selection variables and HIV testing in each consent regression and each contact regression, separately by survey and sex, using Wald tests with a two-sided p value of 0.05. It is highly implausible that the identity of the interviewer in a DHS survey could causally determine respondent HIV status at the time of the interview,[29] and we controlled for observed factors that were used to match interviewers to respondents, such as region and urban setting, which could induce non-causal association between interviewer identity and the HIV status of potential survey participants.

### Uncertainty estimation

Previous approaches to imputing HIV status for missing observations in the DHS have focused on sampling uncertainty conditional on the estimated regression equations when calculating standard errors (SE) or 95% CI for estimates of HIV prevalence.[13 16 17] This approach overstates the precision of imputation-based HIV prevalence estimates because it ignores estimation uncertainty about the imputation regression parameters. We incorporated this additional source of uncertainty with a parametric simulation approach for the conventional imputation and selection model-based imputation strategies.[32 33] The sampling distribution for predicted prevalence among those without a valid HIV test was approximated by calculating prevalence from imputed HIV status for each of the 10 000 regression parameter sets drawn from a multivariate normal distribution parameterised by the maximum likelihood estimates for the regression coefficients and their covariance matrix. To obtain CI for national prevalence estimates, the 10 000 draws from the sampling distribution for imputed prevalence among non-participants were combined with 10 000 draws for prevalence among those with a valid HIV test, which were simulated from a binomial distribution defined by the complete-case analysis. We induced correlation between these two sets of prevalence values using a copula method[34] with correlation coefficients obtained from bootstrapped prevalence estimates in a subset of surveys (further details are described in the online technical appendix). We conducted all statistical analyses in Stata V.11 (StataCorp, College Station, Texas, USA) and prepared figures with R V.2.11.1 (R Foundation for Statistical Computing, Vienna, Austria).

## RESULTS

### Final survey sample

Our final analysis included results from 12 of the 24 DHS surveys that we examined (table 1) as the selection model could not be used in several cases. DHS surveys for Mali 2001, Democratic Republic of Congo 2007 and Zambia 2001–2002 were missing unique identifiers linking an individual's questionnaire responses to their HIV test results or were missing an

interviewer identity variable and therefore could not be analysed. Results for Burkina Faso 2003, Cameroon 2004, Guinea 2005, Kenya 2003, Kenya 2008–2009 and Sierra Leone 2008 were excluded because the estimate of the selection model correlation parameter was near its boundary ($|\rho|>0.9$) in at least one regression, indicating that model parameters were not well identified. Models with $|\rho|>0.9$ also typically had highly significant p values. Last, the independent effects of region and interviewer identity could not be estimated for Niger 2006, Tanzania 2003–2004 and Tanzania 2007–2008 DHS.

### Selection variables

Across 48 selection models (including separate regressions for consent and contact, by sex and survey), interviewer identity was significantly associated with HIV testing participation (at $p<0.05$), even after controlling for observed factors that were used to match interviewers to respondents such as region and urban setting, in 46 cases. The two exceptions were the consent regression for men ($p=0.07$) and for women ($p=0.16$) in Swaziland 2006–2007, see online supplementary table 1. Among the 24 contact regressions, the coefficient for the indicator variable denoting whether or not a household was contacted on the first day that an interviewing team visited a cluster was only significantly associated with participation in the Zambia 2007 women survey (see online supplementary table 1).

### Prevalence estimates

National estimates of adult HIV prevalence, by survey and separately for men and women, are depicted in figure 1 for the complete-case, conventional imputation and Heckman-type selection model approaches (see supplementary table 1 for more detailed results). Selection model point estimates of national HIV prevalence were greater than those based on a complete-case analysis for 10 out of 12 surveys for men and 11 out of 12 surveys for women. In comparison with conventional imputation, selection model point estimates were greater for eight of 12 surveys for men and 11 of 12 surveys for women. These differences were statistically significant in three surveys—Cote d'Ivoire 2005, Mali 2006 and Zambia 2007—which had significant negative values for the selection model correlation parameter ($\rho$) in either the *consent* or *contact* regression for men, indicating strong evidence of higher HIV prevalence among men who did not participate in HIV testing. HIV prevalence estimates derived from the selection modelling approach led to changes in the sex ratio of HIV prevalence. As compared with conventional imputation, the selection model estimated a lower female-to-male prevalence ratio in seven surveys out of 12 surveys. However, the female-to-male prevalence ratio decreased in five of the seven surveys that had substantial changes in HIV prevalence point estimates, defined as a greater than one percentage point change for either men or women (Cote d'Ivoire, Mali, Swaziland, Zambia and Zimbabwe).

Allowing for the possibility that factors not measured in the DHS may influence HIV testing participation resulted in much greater uncertainty around prevalence estimates, with 95% CI for HIV prevalence being 4.5 times wider on average for the selection model estimates compared with those from conventional imputation. On the other hand, in most cases, the 95% CI around the selection model estimates were substantially tighter than the most extreme bounds possible (see in figure 1), which are derived by assuming that all non-participants were uniformly either HIV-negative (for the lower bound) or HIV-positive (for the upper bound). Incorporating regression parameter uncertainty led to 95% CI that were 1.2 times larger for the conventional imputation estimates and 4.9 times larger for the selection model estimates, as compared with the CI obtained for those same models when only sampling uncertainty was accounted for and regression parameter uncertainty was ignored.

### Sensitivity analyses

Sensitivity analyses of two key assumptions of the bivariate probit selection model used in this analysis suggested that our findings were relatively robust to deviations from key model assumptions. First, a simulation experiment based on the Zambia 2007 DHS, which assessed the sensitivity of the selection model to violations of its assumption that interviewer effects on participation do not vary with respect to respondent HIV status, indicated that the large adjustment to the HIV prevalence estimate for men could not be explained by a violation of this assumption. Second, estimates of the correlation parameter $\rho$ from a semi-non-parametric selection model (which relaxes the assumption of bivariate normality of the error terms[35]) were modestly correlated with those from the parametric model. A full description of these analyses can be found in the online technical appendix.

## DISCUSSION

Heckman-type selection models offer a means of testing and correcting for sample selection in HIV testing surveys. We investigated the applicability of one variant of this type of selection model, which uses the identity of survey interviewer and the timing of the interview, to DHS datasets from sub-Saharan Africa. We could not apply this approach in half the data sets we examined either because data on the selection variables were missing or the models could not be identified. Our analysis of the 12 DHS for which we could apply the approach indicated that the relationship between HIV status and participation in HIV testing may vary across surveys, but likely leads to underestimates of prevalence in several countries. Additionally, ignoring selection on unobserved factors with conventional imputation approaches substantially overstates the precision of HIV prevalence estimates in many sub-Saharan African countries.

Among the final sample of 12 surveys, the Heckman-type selection model results can be viewed as a sensitivity analysis of conventional HIV prevalence estimates.[36] The selection model estimates agree with, and add credibility to, existing prevalence estimates for countries such as Liberia, Rwanda and Senegal. However, on average the selection model estimates had CI that were 4.5 times larger than those from conventional imputation, indicating that we are unable to precisely estimate the effect that bias due to low participation rates may have on HIV prevalence estimates in many surveys. Thus, for many countries, including those in southern Africa, policy makers should consider using a wider range of potential values when making decisions that depend on national levels of HIV prevalence. Last, selection model estimates resulting in significant, large increases in estimated HIV prevalence among men in Cote d'Ivoire, Mali and Zambia are most concerning and suggest that renewed focus on HIV prevention in men would be particularly justified in these countries.[37]

The narrow CI frequently reported around conventional estimates of national HIV prevalence reflect a false precision resulting from the assumption that testing non-participants are 'missing at random'.[16 17] The selection model approach relaxes this assumption as it does not assume that the correlation parameter $\rho$ equals 0 with certainty; the wider CI around selection

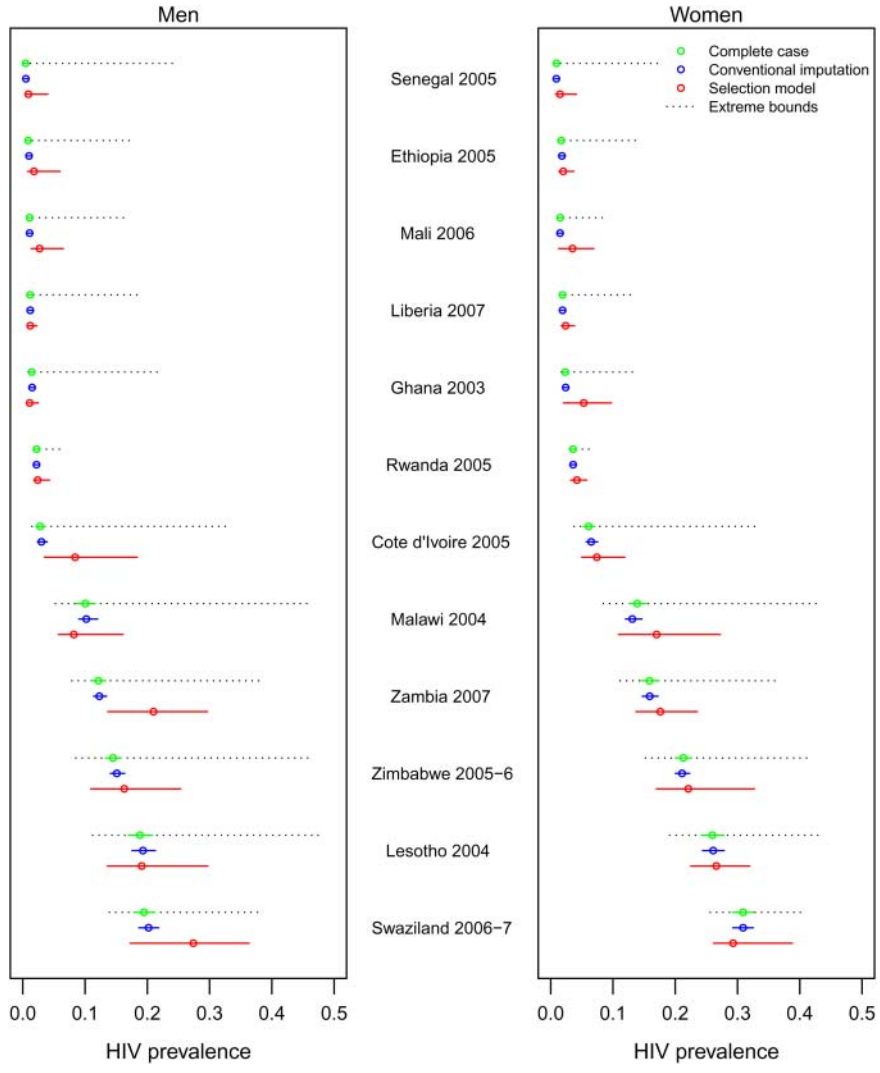



**Figure 1** National adult HIV prevalence estimates with 95% CI derived from three modelling approaches for men and women from 12 Demographic and Health Surveys conducted in sub-Saharan Africa, 2001–2009. Women aged 15–49 years were eligible to be tested for HIV. The age range for men was 15–59 years, with the exceptions of Cote d'Ivoire, Liberia and Swaziland (15–49 years) and Malawi and Zimbabwe (15–54 years). HIV infection was defined as infection with either HIV-1 or HIV-2. Apart from the selection variables described in the text, all other covariates were shared by the two model components of the selection models and the conventional imputation probit regressions. For 'consent' regressions, these variables were: age, educational attainment, household wealth quintile as constructed from an index of household assets, urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behaviour in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with AIDS, willingness to care for a family member with AIDS, and having had a previous HIV test. For 'contact' regressions, these variables were: sex, age, education, wealth quintile, urban setting and region (see details in online technical appendix). 'Extreme bounds' assume that all those missing a valid HIV test are uniformly HIV-positive or HIV-negative.

model-based estimates reflect uncertainty about the strength of the relationship between HIV status and testing. These results offer a quantification of uncertainty around values for population HIV prevalence that is more conservative yet more accurate than conventional approaches[16 17] and typically much narrower than an extreme bounds approach (figure 1). The CI estimated using our approach are also more interpretable from a sampling theory perspective than sensitivity analyses that apply fixed factors to existing estimates.[14 15]

Underestimating the uncertainty around HIV prevalence estimates derived from national population-based surveys has important implications, as it will impact the weight placed on other sources of HIV surveillance data and overstate the precision of measures of HIV burden used in global and national HIV policymaking. For example, in its recent report on the global epidemic, the United Nations Joint Programme of HIV/AIDS (UNAIDS) estimates HIV prevalence in sub-Saharan Africa by rescaling models fit to antenatal clinic (ANC) data so that they are compatible with population-based survey estimates for prevalence.[1 38] If there is less certainty about estimates from population-based surveys than previously thought, weighing ANC data more heavily in such analyses may be appropriate. This would also have implications for estimating HIV incidence, which can be derived from the epidemic models used by UNAIDS[39] or estimated from changes in HIV prevalence between two population-based surveys.[40] Adjustments of HIV prevalence estimates will also affect indicators of antiretroviral treatment coverage,[1] for instance, as measured by the US President's Emergency Plan for AIDS Relief programme[41] and model-based predictions of future HIV trends.[39]

Our study has several limitations. For surveys in which health workers or technicians obtained HIV test consent and blood samples (table 1, category 4), we could only control for the identities of interviewers in the selection model. Although interviewer identity was a significant predictor of HIV testing participation in all except one of these five surveys, the majority of them had selection models with estimates of ρ near the boundary for at least one group, suggesting model identification problems. Future surveys should record the identities of the individuals responsible for conducting HIV testing, in addition to interviewer identity, to allow for broader applicability of selection models. Bayesian methods that enable the estimation of selection model parameters in cases like Tanzania where maximum likelihood techniques fail to converge may also enable wider application of these methods.

Heckman-type selection models can be sensitive to violations of model assumptions,[28] and methodological work is needed to

establish diagnostic tests and robustness checks for applied researchers. The selection model implemented here assumes that the error terms for the selection and outcome equations are distributed bivariate normal, and therefore relies on parametric assumptions for extrapolation. The plausibility of this assumption can be tested with semi- or non-parametric selection models.[42] In an initial sensitivity analysis, we found a modest correlation between estimates for ρ obtained from a semi-non-parametric model and the parametric model used in our main analysis. However, as explained in the online technical appendix, further development of these methods is needed to establish strong tests of assumption validity.

The choice of selection variables can also impact selection model estimates,[28] but we only identified one variable that consistently predicted HIV testing. Our use of interviewer identity as a selection variable has a behavioural justification[29] and has been used in at least three previous studies employing Heckman-type selection models of HIV in Africa.[13 43 44] It is unlikely that interviewers could affect respondent HIV status, and we controlled for the variables used to match interviewers with respondents, namely region and sex. In simulations consistent with the Zambia 2007 data, we found that violations of this assumption would be unlikely to explain the large adjustment to prevalence estimated for adult men in Zambia.

Ideally, the validity and precision of HIV prevalence estimates could be improved through increased HIV testing participation. Increasing contact rates could be achieved through renewed emphasis on revisiting households to test absent members or encouraging individuals who are unwilling to complete the questionnaire to participate in HIV testing. Improving consent rates may be possible if an oral swab is used instead of collecting blood[45 46] and approaches such as financial incentives,[47 48] resampling previous refusers or offering test results and referral to care could be investigated. A deeper understanding of what characteristics predict an individual's propensity to test, and how they relate to HIV status, would be useful, and more research on methods for improving HIV testing participation during large-scale surveys is needed.

In the absence of increased HIV testing participation, we recommend that Heckman-type selection models be included among the toolkit of routine analyses when estimating HIV prevalence, deriving epidemic indicators from HIV prevalence or modelling the determinants of HIV status, as a check on the robustness of conventional methods. To facilitate these efforts, survey reports should describe interview team composition and include unique identifiers for those responsible for contacting households, obtaining consent and conducting HIV tests. Common software packages implement the bivariate probit model, including Stata, SAS and R. We also suggest that analysts incorporate parameter uncertainty when calculating CI around imputation-based estimates. We used a parametric simulation approach to do this;[32] the bootstrap and Bayesian algorithms could be useful alternatives in other settings.[49 50]

In conclusion, Heckman-type selection models provide a useful addition to the set of tools used for the estimation of HIV prevalence from national surveys. In settings where they can be identified, selection models offer a means of assessing potential problems with conventional estimates of HIV prevalence and may suggest substantially revised estimates in some cases. Our analysis indicates that national HIV prevalence estimates for many countries in sub-Saharan Africa are more uncertain than previously thought, and may be underestimated in several cases. This suggests that more emphasis should be put on increasing participation in HIV testing in surveys that aim to establish national prevalence rates.

### Key messages

- National population-based surveys that include HIV testing are a critical source of evidence on HIV prevalence in sub-Saharan Africa.
- Selection models can be used to correct HIV prevalence estimates derived from these surveys for selection bias due to non-participation in HIV testing.
- This study suggests that important uncertainty remains around estimates of HIV prevalence in sub-Saharan Africa and that HIV prevalence may be underestimated in several countries.
- More emphasis should be placed on increasing participation in HIV surveys.

**Acknowledgements** We thank the participants of the UNAIDS Reference Group on Estimates, Modelling and Projections meetings in Seattle, October 2011 and Boston, April 2012 for helpful discussion.

**Contributors** DRH, JAS, DC, TB: conceived the study; DRH: obtained and analysed the data; DRH, JAS, DC, JKH, AMZ, TB: contributed to analytic methods and interpretation of results; DRH: wrote the first draft of the manuscript; JAS, DC, JKH, AMZ, TB: revised the manuscript before submission.

**Funding** DRH was supported by a Harvard University Dissertation Completion Fellowship and a T-32 Training Grant from the National Institute of Allergy and Infectious Diseases (AI 007433). DC received funding support from the William and Flora Hewlett Foundation (2008-2302 and 2011-6455) and the National Institute of Aging (5P30AG024409). TB received funding support through the National Institute of Child Health and Human Development (1R01-HD058482-01) and the National Institute of Mental Health (1R01-MH083539-01). JAS, JKH and AMZ have no financial disclosures.

**Competing interests** None.

**Ethics approval** Ethics committee approval was not required for this work. All data were analysed anonymously.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** Stata code demonstrating how to implement Heckman-type selection models for imputing HIV status is available at our academic website: http://dvn.iq.harvard.edu/dvn/dv/CPDS/faces/study/StudyPage.xhtml?studyId=75001&versionNumber=2.

## REFERENCES

1. **UNAIDS.** *Global report: UNAIDS report on the global AIDS epidemic 2010*. Geneva: UNAIDS, 2010.
2. **Brown T,** Salomon JA, Alkema L, *et al*. Progress and challenges in modelling country-level HIV/AIDS epidemics: the UNAIDS Estimation and Projection Package 2007. *Sex Transm Infect* 2008;**84**(Suppl 1):i5–10.
3. **Schwartlander B,** Stover J, Walker N, *et al*. AIDS. Resource needs for HIV/AIDS. *Science* 2001;**292**:2434–6.
4. **Salomon JA,** Hogan DR, Stover J, *et al*. Integrating HIV prevention and treatment: from slogans to impact. *PLoS Med* 2005;**2**:e16.
5. **Stover J,** Johnson P, Zaba B, *et al*. The Spectrum projection package: improvements in estimating mortality, ART needs, PMTCT impact and uncertainty bounds. *Sex Transm Infect* 2008;**84**(Suppl 1):i24–30.
6. **Hecht R,** Stover J, Bollinger L, *et al*. Financing of HIV/AIDS programme scale-up in low-income and middle-income countries, 2009–31. *Lancet* 2010;**376**:1254–60.
7. **Shisana O,** Rehle T, Simbayi LC, *et al*. *South African national HIV prevalence, incidence, behaviour and communication survey 2008: A turning tide among teenagers?* Cape Town: Human Sciences Research Council, 2009.
8. **Central Statistical Office (CSO)**, Ministry of Health (MOH), Tropical Diseases Research Centre (TDRC), *et al*. *Zambia Demographic and Health Survey 2007*. Calverton, Maryland, USA: CSO and Macro International Inc, 2009.
9. **National AIDS Coordinating Agency (NACA)**, Central Statistics Office (CSO) and Other Development Partners. *The Botswana AIDS impact survey II (BAIS II): Popular report*. Gaborone: National AIDS Coordinating Agency, 2005.

10. **Boerma JT,** Ghys PD, Walker N. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet* 2003;**362**:1929–31.
11. **Ghys PD,** Walker N, McFarland W, *et al*. Improved data, methods and tools for the 2007 HIV and AIDS estimates and projections. *Sex Transm Infect* 2008;**84**(Suppl 1): i1–4.
12. **Gouws E,** Mishra V, Fowler TB. Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data. *Sex Transm Infect* 2008;**84**(Suppl 1):i17–23.
13. **Bärnighausen T,** Bor J, Wandira-Kazibwe S, *et al*. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology* 2011;**22**:27–35.
14. **Garcia-Calleja JM,** Gouws E, Ghys PD. National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates. *Sex Transm Infect* 2006;**82**(Suppl 3):iii64–70.
15. **Reniers G,** Eaton J. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS* 2009;**23**:621–9.
16. **Mishra V,** Barrere B, Hong R, *et al*. Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sex Transm Infect* 2008;**84**(Suppl 1): i63–70.
17. **Mishra V,** Vaessen M, Boerma JT, *et al*. HIV testing in national population-based surveys: experience from the Demographic and Health Surveys. *Bull World Health Organ* 2006;**84**:537–45.
18. **WHO/UNAIDS**. *Guidelines for measuring national HIV prevalence in population-based surveys*. Geneva: WHO/UNAIDS, 2005.
19. **UNAIDS.** *Global Report: Methodology—Understanding the latest estimates*. Geneva: UNAIDS, 2010.
20. **Rubin DB.** Inference and missing data. *Biometrika* 1976;**63**:581–92.
21. **Bärnighausen T,** Tanser F, Malaza A, *et al*. HIV status and participation in HIV surveillance in the era of antiretroviral treatment: a study of linked population-based and clinical data in rural South Africa. *Trop Med Int Health* 2012;**17**:e103–10.
22. **Weiser SD,** Heisler M, Leiter K, *et al*. Routine HIV testing in Botswana: a population-based study on attitudes, practices, and human rights concerns. *PLoS Med* 2006;**3**:e261.
23. **Kalichman SC,** Simbayi LC. HIV testing attitudes, AIDS stigma, and voluntary HIV counselling and testing in a black township in Cape Town, South Africa. *Sex Transm Infect* 2003;**79**:442–7.
24. **Kranzer K,** McGrath N, Saul J, *et al*. Individual, household and community factors associated with HIV test refusal in rural Malawi. *Trop Med Int Health* 2008;**13**:1341–50.
25. **Measure DHS.** Demographic and Health Surveys (DHS) Final Reports. Secondary Demographic and Health Surveys (DHS) Final Reports. 2011. http://www.measuredhs.com
26. **Macro International Inc.** *Sampling manual. DHS-III basic documentation*. Calverton, Maryland: Macro International Inc, 1996.
27. **Dubin J,** Rivers D. Selection bias in linear regression, logit and probit models. *Sociological Methods Res* 1990;**18**:360–90.
28. **Winship C,** Mare R. Models for sample selection bias. *Annu Rev Sociol* 1992;**18**:327–50.
29. **Bärnighausen T,** Bor J, Wandira-Kazibwe S, *et al*. Interviewer identity as exclusion restriction in epidemiology. *Epidemiology* 2011;**22**:446.
30. **Groves R,** Couper M. *Nonresponse in household interview surveys*. New York: Wiley, 1998.
31. **Blohm M,** Hox J, Koch A. The influence of interviewers' contact behavior on the contact and cooperation rate in face-to-face household surveys. *Int J Public Opin Res* 2007;**19**:97–111.
32. **King G,** Tomz M, Wittenberg J. Making the most of statistical analyses: Improving interpretation and presentation. *Am J Political Sci* 2000;**44**:341–55.
33. **Timpone R.** Estimating aggregate policy reform effects: New baselines for registration, participation, and representation. *Political Anal* 2002;**10**:154–77.
34. **Trivedi PK,** Zimmer DM. Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics* 2005;**1**:111.
35. **Stover J,** Brown T, Martson M. Updates to the spectrum/EPP model to estimate HIV trends for adults and children. *Sex Transm Infect* 2012;UNAIDS 2012 supplement.
36. **Geneletti S,** Mason A, Best N. Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only "solution". *Epidemiology* 2011;**22**:36–9.
37. **The United Nations Joint Programme of HIV/AIDS (UNAIDS)**. *Working with men for HIV prevention and care. UNAIDS best practice collection. Key material*. Geneva: UNAIDS, 2001.
38. **Alkema L,** Raftery AE, Brown T. Bayesian melding for estimating uncertainty in national HIV prevalence estimates. *Sex Transm Infect* 2008;**84**(Suppl 1):i11–16.
39. **Brown T,** Bao L, Raftery AE, *et al*. Modelling HIV epidemics in the antiretroviral era: the UNAIDS Estimation and Projection Package 2009. *Sex Transm Infect* 2010;**86** (Suppl 2):ii3–10.
40. **Hallett TB,** Zaba B, Todd J, *et al*. Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. *PLoS Med* 2008;**5**:e80.
41. **The President's Emergency Plan for AIDS Relief.** *Planning and reporting: the next generation indiciators reference guide, version 1.1, August 2009*. Washington DC: United States President's Emergency Plan for AIDS Relief, 2009.
42. **Das M,** Newey WK, Vella F. Nonparametric estimation of sample selection models. *Rev Econ Stud* 2003;**70**:33–58.
43. **Reniers G,** Araya T, Berhane Y, *et al*. Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. *BMC Public Health* 2009;**9**:163.
44. **Janssens W,** van der Gaag J, de Wit T. *Refusal bias in the estimation of HIV prevalence*. Amsterdam: Amsterdam Institute for International Development, 2009.
45. **Pugatch DL,** Levesque BG, Lally MA, *et al*. HIV testing among young adults and older adolescents in the setting of acute substance abuse treatment. *J Acquir Immune Defic Syndr* 2001;**27**:135–42.
46. **Spielberg F,** Critchlow C, Vittinghoff E, *et al*. Home collection for frequent HIV testing: acceptability of oral fluids, dried blood spots and telephone results. HIV Early Detection Study Group. *AIDS* 2000;**14**:1819–28.
47. **Thornton R.** The demand for, and impact of, learning HIV status. *Am Econ Rev* 2008;**98**:1829–63.
48. **Haukoos JS,** Witt MD, Coil CJ, *et al*. The effect of financial incentives on adherence with outpatient human immunodeficiency virus testing referrals from the emergency department. *Acad Emerg Med* 2005;**12**:617–21.
49. **Efron B,** Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1986;**1**:54–77.
50. **Gelman A,** Carlin JB, Stern HS, *et al*. *Bayesian Data Analysis*. 2nd edn. Boca Raton, FL, USA: Chapman & Hall/CRC, 2004.

**Web Table A**

The following tables present detailed HIV prevalence estimates for men and women from 12 DHS surveys. The 'complete case' analysis estimate ('Valid HIV') is reported in the first row for comparison to the imputation results that appear below. Prevalence estimates for those who refused consent and for those who could not be contacted were imputed with either conventional probit regressions ('Imputed') or Heckman-type selection models ('Heckman'). The 'total' estimate combines the three categories for a national estimate of adult HIV prevalence. Estimates of the selection model correlation parameter $\rho$ are reported in the final column. Prevalence estimates are survey weighted but sample sizes (N) are unweighted. P-values testing if the 'selection variables' (i.e., interviewer identities or a variable indicating whether or not a survey team contacted a household on the first day the team visited a cluster) were associated with participation are reported under each table.

CotedIvoire2005

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 3893 | 0.028 (0.021, 0.036) | 0.028 (0.021, 0.036) | |
| Predicted via consent | 588 | 0.034 (0.030, 0.051) | 0.163 (0.041, 0.403) | -0.59 (-0.86, -0.04) |
| Predicted via contact | 663 | 0.035 (0.028, 0.050) | 0.248 (0.034, 0.685) | -0.68 (-0.93, 0.02) |
| Total | 5144 | 0.030 (0.024, 0.039) | 0.084 (0.035, 0.184) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.115$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 4535 | 0.061 (0.052, 0.071) | 0.061 (0.052, 0.071) | |
| Predicted via consent | 620 | 0.072 (0.062, 0.092) | 0.120 (0.019, 0.327) | -0.20 (-0.65, 0.36) |
| Predicted via contact | 611 | 0.079 (0.068, 0.097) | 0.097 (0.018, 0.264) | -0.08 (-0.52, 0.40) |
| Total | 5766 | 0.065 (0.057, 0.075) | 0.074 (0.050, 0.119) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.6$.

Ethiopia2005

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5118 | 0.009 (0.006, 0.012) | 0.009 (0.006, 0.012) | |
| Predicted via consent | 905 | 0.017 (0.014, 0.029) | 0.024 (0.000, 0.273) | -0.10 (-0.97, 0.95) |
| Predicted via contact | 749 | 0.013 (0.010, 0.020) | 0.107 (0.010, 0.453) | -0.66 (-0.93, 0.06) |
| Total | 6772 | 0.010 (0.007, 0.014) | 0.018 (0.008, 0.060) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.193$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5953 | 0.017 (0.013, 0.022) | 0.017 (0.013, 0.022) | |
| Predicted via consent | 843 | 0.029 (0.025, 0.041) | 0.052 (0.003, 0.210) | -0.19 (-0.72, 0.48) |
| Predicted via contact | 345 | 0.027 (0.022, 0.037) | 0.021 (0.000, 0.141) | 0.08 (-0.65, 0.73) |
| Total | 7141 | 0.018 (0.014, 0.024) | 0.020 (0.013, 0.037) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.237$.

Ghana2003

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 4263 | 0.014 (0.011, 0.019) | 0.014 (0.011, 0.019) | |
| Predicted via consent | 753 | 0.016 (0.015, 0.029) | 0.000 (0.000, 0.083) | 0.69 (-0.88, 1.00) |
| Predicted via contact | 321 | 0.014 (0.011, 0.023) | 0.001 (0.000, 0.027) | 0.48 (-0.28, 0.87) |
| Total | 5337 | 0.015 (0.011, 0.020) | 0.011 (0.009, 0.025) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.865$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5285 | 0.023 (0.019, 0.028) | 0.023 (0.019, 0.028) | |
| Predicted via consent | 410 | 0.028 (0.025, 0.039) | 0.446 (0.008, 0.934) | -0.86 (-1.00, 0.32) |
| Predicted via contact | 251 | 0.021 (0.018, 0.029) | 0.041 (0.000, 0.460) | -0.16 (-0.94, 0.89) |
| Total | 5946 | 0.024 (0.020, 0.028) | 0.053 (0.021, 0.097) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.9$.

Lesotho2004

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 2246 | 0.188 (0.170, 0.208) | 0.188 (0.170, 0.208) | |
| Predicted via consent | 553 | 0.205 (0.182, 0.239) | 0.243 (0.064, 0.498) | -0.10 (-0.59, 0.45) |
| Predicted via contact | 504 | 0.200 (0.180, 0.225) | 0.145 (0.001, 0.615) | 0.15 (-0.79, 0.88) |
| Total | 3303 | 0.193 (0.175, 0.213) | 0.191 (0.136, 0.297) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.966$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 3032 | 0.260 (0.242, 0.278) | 0.260 (0.242, 0.278) | |
| Predicted via consent | 514 | 0.269 (0.248, 0.295) | 0.283 (0.096, 0.537) | -0.03 (-0.50, 0.45) |
| Predicted via contact | 212 | 0.258 (0.239, 0.280) | 0.314 (0.080, 0.629) | -0.11 (-0.58, 0.43) |
| Total | 3758 | 0.261 (0.244, 0.279) | 0.266 (0.225, 0.320) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.707$.

Liberia2007

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5241 | 0.012 (0.009, 0.015) | 0.012 (0.009, 0.015) | |
| Predicted via consent | 768 | 0.014 (0.014, 0.024) | 0.018 (0.001, 0.091) | -0.06 (-0.56, 0.48) |
| Predicted via contact | 461 | 0.013 (0.011, 0.019) | 0.011 (0.000, 0.071) | 0.05 (-0.52, 0.59) |
| Total | 6470 | 0.012 (0.010, 0.016) | 0.012 (0.009, 0.023) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.581$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 6533 | 0.019 (0.016, 0.023) | 0.019 (0.016, 0.023) | |
| Predicted via consent | 561 | 0.023 (0.021, 0.031) | 0.074 (0.010, 0.237) | -0.34 (-0.72, 0.20) |
| Predicted via contact | 351 | 0.022 (0.019, 0.028) | 0.047 (0.004, 0.169) | -0.21 (-0.65, 0.34) |
| Total | 7445 | 0.019 (0.016, 0.023) | 0.024 (0.017, 0.038) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.098$.

Malawi2004

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 2404 | 0.100 (0.087, 0.116) | 0.100 (0.087, 0.116) | |
| Predicted via consent | 837 | 0.098 (0.086, 0.128) | 0.071 (0.001, 0.341) | 0.14 (-0.74, 0.84) |
| Predicted via contact | 553 | 0.119 (0.105, 0.143) | 0.024 (0.000, 0.183) | 0.53 (-0.37, 0.92) |
| Total | 3794 | 0.102 (0.090, 0.120) | 0.082 (0.058, 0.161) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.205$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 2864 | 0.139 (0.126, 0.153) | 0.139 (0.126, 0.153) | |
| Predicted via consent | 975 | 0.111 (0.102, 0.137) | 0.254 (0.050, 0.598) | -0.41 (-0.83, 0.31) |
| Predicted via contact | 231 | 0.129 (0.118, 0.148) | 0.152 (0.019, 0.428) | -0.08 (-0.64, 0.55) |
| Total | 4070 | 0.131 (0.120, 0.147) | 0.170 (0.109, 0.272) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.799$.

Mali2006

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 3946 | 0.011 (0.007, 0.016) | 0.011 (0.007, 0.016) | |
| Predicted via consent | 303 | 0.012 (0.011, 0.026) | 0.032 (0.000, 0.342) | -0.26 (-0.93, 0.82) |
| Predicted via contact | 384 | 0.011 (0.009, 0.022) | 0.177 (0.053, 0.442) | -0.76 (-0.92, -0.41) |
| Total | 4633 | 0.011 (0.007, 0.016) | 0.027 (0.014, 0.065) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.946$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 4804 | 0.015 (0.011, 0.021) | 0.015 (0.011, 0.021) | |
| Predicted via consent | 225 | 0.017 (0.014, 0.028) | 0.366 (0.000, 0.990) | -0.86 (-1.00, 0.96) |
| Predicted via contact | 124 | 0.016 (0.013, 0.024) | 0.184 (0.015, 0.539) | -0.67 (-0.93, 0.00) |
| Total | 5153 | 0.015 (0.011, 0.020) | 0.035 (0.013, 0.069) | |

Consent regression p-value for interviewers $= 0.002$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.114$.

Rwanda2005

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 4742 | 0.022 (0.018, 0.027) | 0.022 (0.018, 0.027) | |
| Predicted via consent | 90 | 0.034 (0.028, 0.052) | 0.138 (0.000, 0.811) | -0.46 (-1.00, 0.98) |
| Predicted via contact | 122 | 0.031 (0.026, 0.042) | 0.010 (0.000, 0.403) | 0.24 (-0.97, 0.99) |
| Total | 4954 | 0.022 (0.018, 0.027) | 0.024 (0.018, 0.043) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $= 0.043$.
Contact regression p-value for first day in cluster $= 0.847$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5677 | 0.036 (0.031, 0.041) | 0.036 (0.031, 0.041) | |
| Predicted via consent | 65 | 0.055 (0.047, 0.069) | 0.268 (0.002, 0.784) | -0.58 (-0.98, 0.80) |
| Predicted via contact | 93 | 0.039 (0.034, 0.047) | 0.209 (0.000, 0.840) | -0.54 (-0.99, 0.89) |
| Total | 5835 | 0.036 (0.031, 0.041) | 0.042 (0.032, 0.058) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.546$.

Senegal2005

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 3303 | 0.005 (0.003, 0.008) | 0.005 (0.003, 0.008) | |
| Predicted via consent | 487 | 0.005 (0.004, 0.016) | 0.035 (0.003, 0.193) | -0.52 (-0.86, 0.12) |
| Predicted via contact | 564 | 0.005 (0.003, 0.012) | 0.014 (0.000, 0.157) | -0.28 (-0.84, 0.57) |
| Total | 4354 | 0.005 (0.003, 0.008) | 0.009 (0.003, 0.040) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.127$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 4520 | 0.009 (0.007, 0.013) | 0.009 (0.007, 0.013) | |
| Predicted via consent | 577 | 0.008 (0.007, 0.016) | 0.052 (0.003, 0.252) | -0.52 (-0.87, 0.18) |
| Predicted via contact | 245 | 0.010 (0.007, 0.020) | 0.033 (0.001, 0.218) | -0.31 (-0.80, 0.42) |
| Total | 5342 | 0.009 (0.007, 0.013) | 0.015 (0.008, 0.041) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.333$.

Swaziland2006-7

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 3630 | 0.195 (0.178, 0.212) | 0.195 (0.178, 0.212) | |
| Predicted via consent | 526 | 0.229 (0.215, 0.248) | 0.669 (0.133, 0.964) | -0.73 (-0.98, 0.34) |
| Predicted via contact | 509 | 0.228 (0.212, 0.247) | 0.447 (0.001, 0.989) | -0.42 (-0.98, 0.91) |
| Total | 4665 | 0.202 (0.186, 0.218) | 0.274 (0.173, 0.364) | |

Consent regression p-value for interviewers $= 0.069$.
Contact regression p-value for interviewers $= 0.037$.
Contact regression p-value for first day in cluster $= 0.383$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 4624 | 0.309 (0.292, 0.326) | 0.309 (0.292, 0.326) | |
| Predicted via consent | 383 | 0.294 (0.279, 0.313) | 0.057 (0.000, 0.923) | 0.54 (-1.00, 1.00) |
| Predicted via contact | 283 | 0.327 (0.310, 0.344) | 0.343 (0.001, 0.942) | -0.03 (-0.90, 0.89) |
| Total | 5290 | 0.309 (0.292, 0.325) | 0.293 (0.262, 0.388) | |

Consent regression p-value for interviewers $= 0.163$.
Contact regression p-value for interviewers $= 0.033$.
Contact regression p-value for first day in cluster $= 0.801$.

Zambia2007

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5163 | 0.121 (0.110, 0.133) | 0.121 (0.110, 0.133) | |
| Predicted via consent | 1318 | 0.117 (0.110, 0.131) | 0.520 (0.219, 0.825) | -0.75 (-0.94, -0.23) |
| Predicted via contact | 653 | 0.153 (0.141, 0.167) | 0.248 (0.014, 0.697) | -0.24 (-0.83, 0.60) |
| Total | 7134 | 0.123 (0.114, 0.134) | 0.210 (0.137, 0.297) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.236$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5713 | 0.159 (0.145, 0.173) | 0.159 (0.145, 0.173) | |
| Predicted via consent | 1400 | 0.158 (0.148, 0.173) | 0.245 (0.075, 0.501) | -0.22 (-0.65, 0.30) |
| Predicted via contact | 283 | 0.172 (0.158, 0.188) | 0.172 (0.013, 0.514) | 0.00 (-0.64, 0.64) |
| Total | 7396 | 0.159 (0.147, 0.172) | 0.176 (0.137, 0.235) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $< 0.001$.

Zimbabwe2005-6

| Men | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 5555 | 0.145 (0.133, 0.158) | 0.145 (0.133, 0.158) | |
| Predicted via consent | 1604 | 0.153 (0.142, 0.169) | 0.197 (0.053, 0.420) | -0.12 (-0.57, 0.37) |
| Predicted via contact | 1585 | 0.172 (0.158, 0.188) | 0.193 (0.014, 0.558) | -0.06 (-0.69, 0.63) |
| Total | 8744 | 0.151 (0.141, 0.164) | 0.163 (0.109, 0.253) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.118$.

| Women | | Imputed | Heckman | |
|---|---|---|---|---|
| | N | HIV (95% CI) | HIV (95% CI) | $\rho$ (95% CI) |
| Valid HIV | 7494 | 0.213 (0.200, 0.226) | 0.213 (0.200, 0.226) | |
| Predicted via consent | 1390 | 0.203 (0.194, 0.214) | 0.333 (0.062, 0.677) | -0.27 (-0.76, 0.41) |
| Predicted via contact | 970 | 0.212 (0.202, 0.224) | 0.119 (0.000, 0.840) | 0.24 (-0.93, 0.97) |
| Total | 9854 | 0.211 (0.200, 0.223) | 0.221 (0.170, 0.327) | |

Consent regression p-value for interviewers $< 0.001$.
Contact regression p-value for interviewers $< 0.001$.
Contact regression p-value for first day in cluster $= 0.817$.

**Online Technical Appendix. National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models**

**Contents**



## 1. Heckman-type selection model equations

Dubin and Rivers described the model equations that extend Heckman's original method to the case of a dichotomous outcome, such as HIV status.[1, 2] The equation that predicts participation in HIV testing for individual *i* ($s_i$) is the following probit model [3]:

$$s_i^* = \beta_s x_i + \phi z_i + u_i$$
$$s_i = 1 \; if \; s_i^* > 0, s_i = 0 \; otherwise$$

where *x* are observed characteristics, *z* are selection variables subject to an exclusion restriction, and *u* is random error. HIV status $h_i$ is observed if $s_i = 1$. The equation for the HIV status of individual *i* ($h_i$) is predicted with a second probit model:

$$h_i^* = \beta_h x_i + \varepsilon_i$$
$$h_i = 1 \; if \; h_i^* > 0, h_i = 0 \; otherwise$$

where *x* are observed characteristics and ε is random error. The error terms *u* and ε are assumed to be distributed bivariate normal, and the parameter $\rho = corr(u, \varepsilon)$ measures the magnitude and direction of the correlation between participation and HIV status on the probit scale after controlling for the variables in *x.* A negative value of ρ would indicate that individuals who are more likely to be HIV positive are less likely to participate in testing, conditional on observed variables. Note that the conventional imputation probit model is nested within the bivariate probit selection model, and it can be thought of as a selection model that assumes $\rho$=0 with certainty.

## 2. Regression variables

The DHS system uses standardized questionnaires, and country specific questions are recoded to allow for comparisons across countries and surveys.[4] We used the same set of variables in conventional probit and selection model-based imputation regression models across surveys whenever possible, following previous work.[3] For those who completed an individual questionnaire, these variables included age, educational attainment, household wealth quintile as constructed from an index of household assets, urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behavior in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with AIDS, willingness to care for a family member with AIDS, and having had a previous HIV test.[3] In some cases, we used only one of two variables when they were highly collinear (e.g., when there was nearly complete overlap between ethnicity and language). In a small departure from the Zambian 2007 analysis, we defined the "married" variable with three categories (i.e., never married, currently married, and formerly married), as widowed individuals may be at high risk for HIV infection. For those individuals for whom information was only available from the household questionnaire, we controlled for sex, age, education, wealth quintile, urban setting, and region. In Senegal 2005, which had low prevalence among men, we used wider age categories to ensure that there were HIV positive individuals in each category. Rates of missing observations for covariates were low across surveys, typically within the range of 2-4% of individuals missing at least one covariate observation on the individual questionnaire. We formed a single HIV status variable for surveys that reported HIV-1 and HIV-2 status.

For the selection models, we operationalized interviewer identity by creating a dummy variable for each interviewer. Interviewers who conducted at least 50 interviews were assigned their own dummy variable and those who conducted fewer than 50 interviewers were combined in an 'other interviewer' dummy variable.[3] Estimating the effect that interviewers who conduct very few interviews have on participation in testing is difficult and can lead to lack of identification or to numerical problems in obtaining estimates. In Malawi 2004 we used 30 interviews as the minimum threshold when assigning interviewers unique dummy variables, as many interviewers in these surveys did not complete at least 50 interviews. We explored using a threshold of 30 interviews across surveys but encountered model convergence issues with this approach in some settings.

## 3. Accounting for survey design

We employed household sampling weights to calculate nationally representative estimates of HIV prevalence for all three modeling strategies. The use of household weights is more appropriate than individual weights, which are adjusted for non-participation, as we correct for non-participation in our analysis. We incorporated sampling weights after estimating regression models, as the variables used to construct the sampling weights were included as regression covariates. Thus, for both imputation-based modeling strategies, regressions were fit without sampling weights, HIV status was predicted for those without a valid HIV test, and then a sampling-weighted average was calculated for those predictions. We accounted for survey strata and household clustering when estimating the covariance matrix of regression parameters.

## 4. Parametric simulation of 95% confidence intervals

We employed a parametric simulation approach to generate uncertainty intervals around imputation-based HIV prevalence estimates, which incorporates uncertainty about imputed HIV status and sampling variation.[5, 6] We simulated the sampling distribution of predicted prevalence for the two groups of people who were missing a valid HIV test—those who could not be contacted and those who refused consent—using the same procedure for conventional imputation and selection-model-based imputation strategies. First, we fit the regression model and saved the maximum likelihood estimates of the coefficients and their covariance matrix, which was adjusted to account for the complex survey design. In the case of the selection model, these coefficients included those from the selection and outcome equations and the correlation parameter $\rho$. Next, 10,000 regression parameter sets were drawn from a multivariate normal distribution parameterized by the coefficients and covariance matrix obtained in the first step.[5] For each set of regression parameter draws, we predicted HIV status and calculated sampling-weighted mean prevalence for those missing a valid HIV test. Aggregating these prevalence estimates across simulation draws approximated the sampling distribution of imputed prevalence for those missing a valid HIV test.

Obtaining 95% confidence intervals for national estimates of HIV prevalence required combining the uncertainty around imputed prevalence estimates for nonparticipants as described above with the sampling uncertainty around the prevalence estimate for those with observed HIV status. To incorporate uncertainty for the latter, we first simulated 10,000 prevalence values from a binomial distribution, parameterized with a probability equal to the complete case estimate for prevalence and a population size appropriate for the complex survey design. To approximate the sampling distribution for national HIV prevalence, the simulated values for HIV prevalence among those with a valid HIV test cannot be combined at random with the simulated values for imputed prevalence for those missing a valid HIV test because of correlated sampling uncertainty around these estimates. To address this, we induced correlation between the sets of simulated prevalence values with an empirical distribution copula method.[7] This procedure involves rank-ordering two vectors and then re-ordering them so as to induce a pre-specified amount of correlation in their values. We first used the copula method to combine the two vectors of imputed prevalence values (i.e., estimates for those who could not be contacted and those who refused consent). Then, we combined this vector with the simulated values from the sampling distribution for the complete-case analysis.

For the copula method, we used the average of the correlation coefficients calculated from comparisons of bootstrapped draws around prevalence estimates from analyses of the Cote d'Ivoire, Zambia and Zimbabwe surveys (correlation coefficients were similar across surveys and between men and women; see section below for description of bootstrapping procedure). For the conventional imputation analyses that relied on a probit regression, the correlation between imputed prevalence for those who refused consent and those who could not to be contacted was 0.66, and the correlation between the combined imputed prevalence for those who did not have a valid HIV test and those with a valid HIV test was 0.67. For the selection model analyses, the correlation between imputed prevalence for those who refused consent and those who could not to be contacted was 0.46, and the correlation between the combined imputed prevalence for those who did not have a valid HIV test and those with a valid HIV test was 0.17.

## 5. Participation rates

The proportion of eligible individuals participating in HIV testing in the 12 DHS surveys included in the final analysis ranged from 63 to 96% in men and 70 to 97% in women, with higher participation rates among women (Table 1). Non-consent was the more common cause of non-participation in HIV testing for women, while men had similar rates of non-participation due to non-consent and non-contact. Considering men and women separately, the span in non-participation outcomes between the most and the least successful interviewers, in terms of either non-consent or non-contact, had a median value of 30 or more percentage points in all cases. All surveys had at least one interviewer with a non-participation rate below 9%, with the exceptions of Zambia 2007 (for which the lowest non-contact rate for men was 13%) and Zimbabwe 2005-6 (where the lowest non-contact rate for men was 12%).

## 6. Bootstrapped confidence intervals

The parametric simulation approach to generating 95% confidence intervals for imputation-based prevalence estimates makes strong distributional assumptions. The bootstrap is a more robust approach but was not feasible to implement for many surveys, for example due to collinearity between interviewer identities and the region variable. For comparison to the parametric simulation approach, we obtained bootstrapped confidence intervals for HIV prevalence imputed with the selection modeling approach in the Cote d'Ivoire 2005, Zambia 2007, and Zimbabwe 2005-6 surveys. To construct a bootstrap data set, we resampled clusters of households within each stratum. Across these three surveys, the bootstrapped 95% confidence intervals for HIV prevalence from the selection modeling approach for those refusing consent, for those who could not be contacted, and for the total national estimate were less conservative than those obtained from the parametric simulation approach, as shown below:

| | Cote d'Ivoire 2005 | | Zambia 2007 | | Zimbabwe 2005-6 | |
|---|---|---|---|---|---|---|
| | Simulation | Bootstrap | Simulation | Bootstrap | Simulation | Bootstrap |
| Men | | | | | | |
| No consent | 4.1, 40.3 | 8.6, 24.6 | 21.9, 82.5 | 34.6, 66.2 | 5.3, 42.0 | 10.7, 30.9 |
| No contact | 3.4, 68.5 | 12.3, 43.4 | 1.4, 69.7 | 8.2, 46.8 | 1.4, 55.8 | 6.2, 35.0 |
| National | 3.5, 18.4 | 5.5, 12.3 | 13.7, 29.7 | 17.0, 25.0 | 10.9, 25.3 | 12.9 20.3 |
| Women | | | | | | |
| No consent | 1.9, 32.7 | 6.0, 20.7 | 7.5, 50.1 | 14.2, 35.7 | 6.2, 67.7 | 14.5, 49.7 |
| No contact | 1.8, 26.4 | 5.2, 18.0 | 1.3, 51.4 | 0.1, 28.9 | 0.0, 84.0 | 0.0, 41.6 |
| National | 5.0, 11.9 | 6.1, 9.7 | 13.7, 23.5 | 15.1, 20.2 | 17.0, 32.7 | 18.4, 26.6 |

## 7. Semi-nonparametric selection model

The parametric selection model used in the main analysis assumes that the error terms in the selection model are distributed bivariate normal. If this assumption was violated, it could impact the accuracy of the model's imputation results. There are limited choices among existing software packages for implementing models that relax the bivariate normality assumption. For our application, we used a semi-nonparametric selection model that approximates the unknown densities of the two error terms by Hermite polynomial expansions.[8] This is implemented in Stata in the –snp2s- command.[8] This approach is somewhat limited for our purposes as the

intercepts are not identified and therefore cannot be used for imputation. Thus, we only used it to estimate the selection model correlation parameter, $\rho$, for comparison to the estimate from the parametric model used in the main analysis. The semi-nonparametric model is computationally intensive to fit, so we only replicated the *consent* regressions for the sensitivity analysis. For each regression, we compared models fit under two possible specifications for the orders of the polynomial expansions: 3 for the selection model and 3 for the outcome model vs. 4 for the selection model and 4 for the outcome model. The preferred model was selected based on a likelihood ratio test,[8] except in a few cases where only one of the two expansion specifications converged, in which case the results from the converged model were used. Semi-nonparametric estimates of $\rho$ were modestly correlated with those from the parametric model, with a correlation of 0.27, and tended to be closer to zero. All semi-nonparametric estimates of $\rho$ were covered the 95% CI for $\rho$ estimated with the bivariate probit selection model. The estimate for men in Zambia 2007 was similar but slightly lower, with $\rho$=-0.58 as compared to $\rho$=-0.75 from the parametric model. Given the limitations of this particular semi-nonparametric model, further development of semi- and nonparametric selection models is needed to establish strong tests of the bivariate normality assumption, which is a promising area for future research.

## 8. Simulation experiment of selection model sensitivity

If interviewers differ in their effect on participation in HIV testing, it is worth considering the sensitivity of the selection model to more complex interactions between interviewers and eligible individuals. For example, interviewer impact on participation could vary with the HIV status of respondents, which would violate the assumption of a constant value for $\rho$. Here we consider the case in which more successful interviewers obtain higher consent rates among those with HIV as compared to those without HIV. We used simulation to explore how this form of selection bias would affect estimates obtained from the selection model in comparison to complete case and conventional imputation analyses.

For the simulation, we used a simplified set of parameters informed from the analysis of men who refused consent in the Zambian 2007 DHS. We specified that $\rho = 0$ and generated HIV status for 5,000 individuals as:

$$h_i^* = -1.28 + 0.30 x_{1i} + \varepsilon_i$$

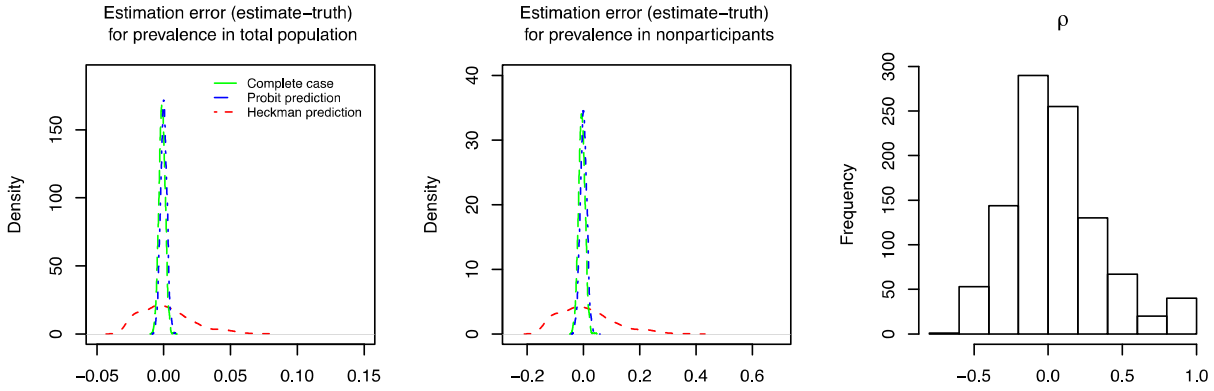
with $\varepsilon$~N(0,1) and $h_i = 1$ if $h_i^* > 0$. The variable $x_1$ denoted urban vs. rural regions, with 40% of the population located in an urban setting. For the base case, we generated participation status for the 5,000 individuals in the data set as follows, with each respondent assigned one of 34 interviewers who had unique effects on participation:

$$s_i^* = 0.7 + 0.24 x_{1i} + \boldsymbol{\phi}\mathbf{z}_i + u_i$$

where $u$ ~ N(0,1), $\mathbf{z}_i$ is a vector that indicates which interviewer was assigned to respondent $i$, $\boldsymbol{\phi}$ is a vector with interviewer-specific participation effects, and $s_i = 1$ if $s_i^* > 0$. For half of the interviewers (group A, the successful interviewers), we assigned each interviewer $j$ a unique participation effect $\phi_j$ ~ Uniform(0.28,0.68), and for the other half of the interviewers (Group

B), we drew $\phi_j$ ~ Uniform(-0.15,0.25). These specifications yielded an average participation rate of 86% in successful interviewer group A and 74% in interviewer group B, matching what was observed in the Zambian data set. They also yielded a distribution of participation rates across interviewers that was comparable to that observed in the Zambian data.

The data for this base case have no selection on unobserved factors and it is useful to compare the performance of the three modeling approaches explored in this paper in this context. As shown here in density plots of prevalence estimation error, comparing true sample means to those estimated with the three different modeling strategies across 1,000 simulated data sets, estimates from the Heckman-type selection model are unbiased but less precise than those obtained from either the complete case or standard probit imputation model:




The complete case analysis, which ignores the effects of $x_1$, leads to a slight underestimate of prevalence, as $x_1$ is associated with higher HIV prevalence and lower participation. For a small number of simulated data sets, the selection model estimated the correlation parameter $\rho$ to be nearly equal to 1 (in many of these cases, the model failed to converge).

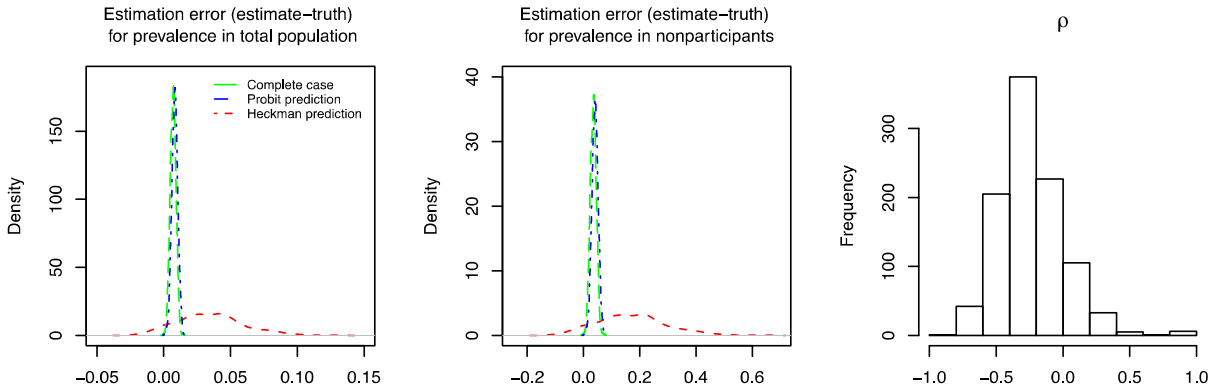
To explore the potential impact of differential interviewer effect by respondent HIV status, we regenerated participation status for respondents who had interviewers from the successful interviewers (group A) as follows:

$$s_i^* = 0.7 + 0.24\, x_{1i} + \lambda h_i \phi_{\mathbf{z}_i} + u_i$$

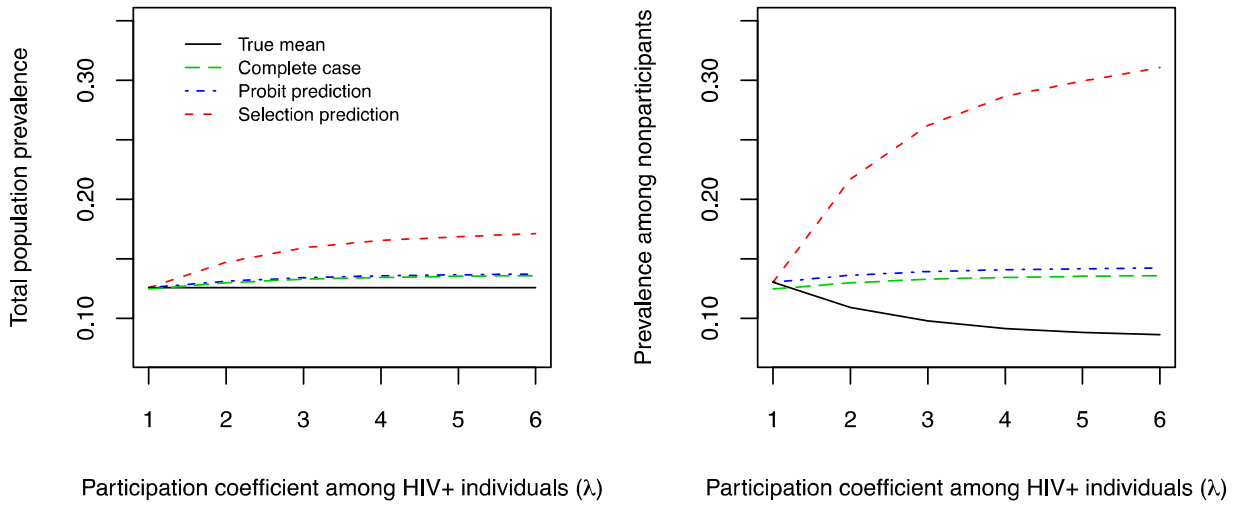
Larger positive values for λ yield higher participation rates for HIV positive individuals among successful interviewers. This mechanism generates selection bias in the data but the bias is of a different form than that which motivates the selection model. To maintain the same overall participation rates in interviewer group A across different values for λ, we reduced the absolute effect that each interviewer in group A had on participation by adjusting the uniform distribution for sampling values of $\phi_j$. These distributions were parameterized as follows:

| λ | Interviewer effect $\phi_j$ |
|---|---|
| 1 | $\phi_j$ ~ (0.28, 0.68) |
| 2 | $\phi_j$ ~ (0.23, 0.63) |
| 3 | $\phi_j$ ~ (0.21, 0.61) |
| 4 | $\phi_j$ ~ (0.19, 0.59) |
| 5 | $\phi_j$ ~ (0.18, 0.58) |
| 6 | $\phi_j$ ~ (0.17, 0.57) |

These parameterizations result in a reduction in the proportion of HIV negative individuals in group A who participate as λ increases, maintaining an over all participation rate of 86% in group A. By way of example, if λ = 3, the data generated under these conditions leads to biased prevalence estimates in all three modeling strategies. The complete case and standard imputation analyses provide similar estimates, which are biased upwards. The selection model predictions are biased upwards to a greater extent than the complete case or conventional imputation model, as the model "corrects" in the wrong direction (i.e., $\rho$ should be positive). The bias arises because there is relatively higher prevalence among consenters in the successful interviewer group, which leads to the model predicting higher prevalence among those who did not consent.




To systematically examine the relationship between λ and the amount of bias in predicted prevalence from different modeling strategies, we plotted mean estimates of prevalence across 1,000 simulations for the different values of λ. In most simulations, a value of λ=6 results in all HIV positive individuals participating within group A. The predicted prevalence estimates obtained from complete case, conventional probit, and selection model strategies are all biased for λ>1:




The magnitude of the difference between estimated and true prevalence increased nonlinearly with λ and suggests that systematic differences in interviewer consent rates by respondent HIV status do have the potential to lead to biased estimates of HIV prevalence with a selection model. However, the magnitude of the change in estimated prevalence in even the most extreme simulations was smaller than that estimated for adult men in the Zambia 2007 survey in the main analysis, suggesting that this violation of the model's assumptions, if it were to occur, would be unlikely to serve as an alternative explanation for our findings.

**9. Software**

Software commands implementing the bivariate probit model used in this study include: -heckprob- in Stata (StataCorp, College Station, TX), PROC QLIM in SAS (SAS Institute Inc., Cary, NC), and the sampleSelection (Henningsen and Toomet) and SemiParBIVProbit packages in R (Marra and Radice) in R (Foundation for Statistical Computing, Vienna, Austria).

## References


1. Heckman J. Sample selection bias as a specification error. Econometrica 1979;**47**(1):153-62.
2. Dubin J, Rivers D. Selection bias in linear regression, logit and probit models. Sociological Methods and Research 1990;**18**(2 & 3):360-90.
3. Bärnighausen T, Bor J, Wandira-Kazibwe S, et al. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology 2011;**22**(1):27-35.
4. Measure DHS. Demographic and Health Surveys (DHS) Final Reports. Secondary Demographic and Health Surveys (DHS) Final Reports. 2011. http://www.measuredhs.com.
5. King G, Tomz M, Wittenberg J. Making the most of statistical analyses: Improving interpretation and presentation. American Journal of Political Science 2000;**44**(2):341-55.
6. Timpone R. Estimating aggregate policy reform effects: New baselines for registration, participation, and representation. Political Analysis 2002;**10**(2):154-77.
7. Trivedi PK, Zimmer DM. Copula modeling: an introduction for practitioners. Foundations and Trends in Econometrics 2005;**1**(1):111.
8. Stover J, Brown T, Martson M. Updates to the Spectrum/EPP model to Estimate HIV Trends for Adults and Children [under review]. Sex Transm Infect 2012;**UNAIDS 2012 supplement**