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 (ρ) 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 ρ 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 ρ equals 0 with certainty; the wider CI around selection

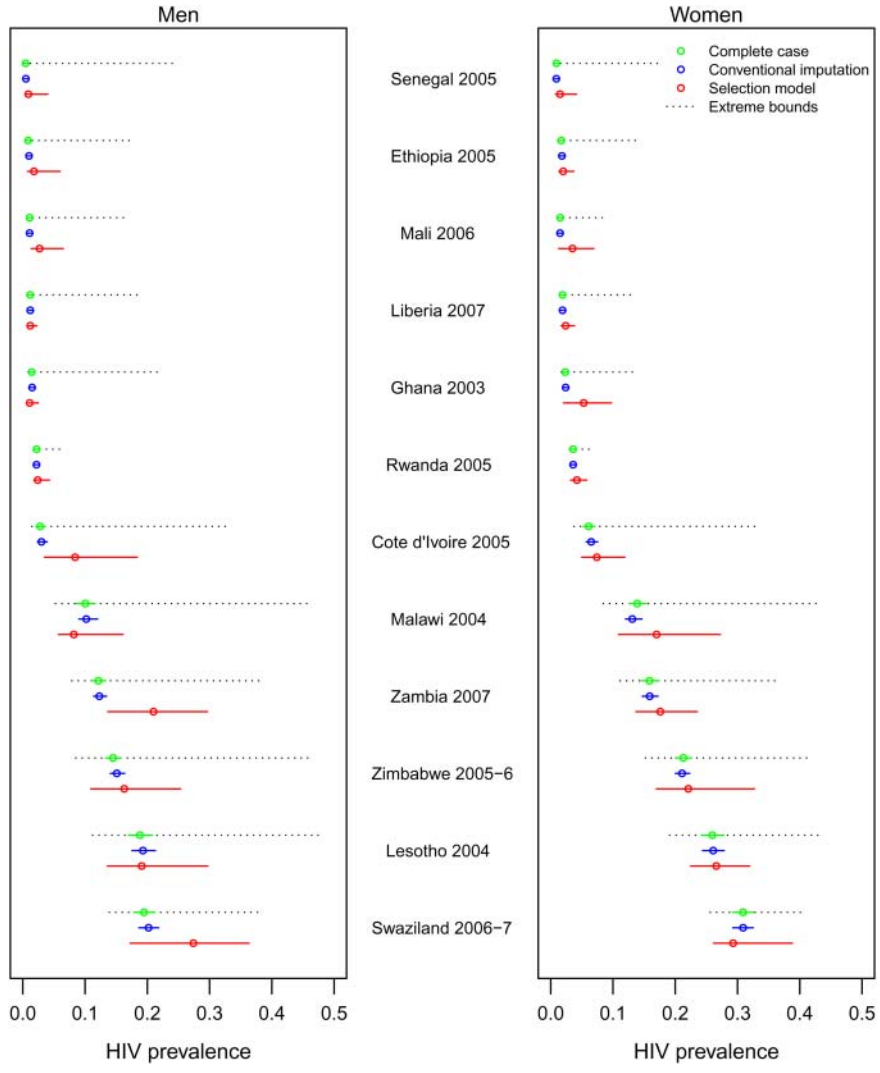

Figure 1 National adult HIV prevalence estimates with 95% CI derived from three modelling approaches for men and women from 12 Demographic and Health Surveys conducted in sub-Saharan Africa, 2001–2009. Women aged 15–49 years were eligible to be tested for HIV. The age range for men was 15–59 years, with 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.