eAppendix

Correcting HIV prevalence estimates for survey non-participation using Heckman-type selection models

Till Bärnighausen, Jacob Bor, Speciosa Wandira-Kazibwe, David Canning

Contents

Sample description 3

HIV testing laboratory procedures 4

Plausibility check of estimated HIV prevalence in men 4

Identification of valid exclusion restrictions 5

Limitations of Heckman-type selection models 6

References 8

eTABLE 1A: Descriptive statistics (men) 9

eTABLE 1B: Descriptive statistics (women) 10

eTABLE 2: Consent regressions (women) 11

eTABLE 3: Contact regressions (women) 13

eTABLE 4A: Consent regressions: other independent variables (men) 14

eTABLE 4B: Consent regressions: exclusion restrictions (men) 15

eTABLE 5A: Consent regressions: other independent variables (women) 16

eTABLE 5B: Consent regressions: exclusion restrictions (women) 17

eTABLE 6A: Contact regressions: other independent variables (men) 18

eTABLE 6B: Contact regressions: exclusion restrictions (men) 19

eTABLE 7A: Contact regressions: other independent variables (women) 20

eTABLE 7B: Contact regressions: exclusion restrictions (women) 21

Sample description

The households eligible for household interviews in the 2007 Demographic and Health Survey in Zambia were selected as follows: First, a stratified random sample of 320 clusters of households was drawn from an enumeration list of 16,757 household clusters, which had been previously used in the 2000 Census of Population and Housing of the Republic of Zambia.[1] The 18 strata were defined by province and rural *versus* urban location. Second, a complete listing of all households in the 320 selected clusters was conducted, and a systematic sample of 8000 households was drawn. One cluster was excluded, leading to a sample of 7969 eligible households. In some households the interview could not take place because the household structure was no longer occupied, no eligible household member was present at the time of the fieldworker visit, or none of the present eligible household members consented to be interviewed with the Household Questionnaire (see Figure 1 in the article).[1]

"Any adult member of the household who is capable of providing information needed to fill in the Household Questionnaire" was eligible to serve as respondent for the Household Questionnaire.[2] In the household interview, the respondent was asked to name all household members, and all visitors who had stayed at the household during the previous night, and to provide information on their sex and age. All men aged 15-59 years and all women aged 15-49 years "who were either permanent residents of the households in the 2007 ZDHS [Zambian Demographic and Health Survey] sample or visitors present in the household on the night before the survey were eligible" for the individual interview and HIV testing (7146 men and 7408 women). Figure 1 shows the samples used in HIV prevalence estimation in this study.

Eighteen men and seven women with known HIV status had missing information for at least one of the independent variables included in the contact regressions; an additional 80 men and 458 women with known HIV status had missing information for at least one of the independent variables included in the consent regressions (Figure 1 in the article). These individuals were thus not included in the respective selection and imputation models. Not being able to use information on these individuals in predicting HIV status of HIV survey non-participants does not result in bias or inconsistency in our results provided that the missing variables in the data, other than HIV status, are missing at random.

HIV testing laboratory procedures

After obtaining consent, interviewers used a sterile finger-prick lancet to collect a dried blood spot (DBS) sample on a filter paper card. HIV status was determined by antibody testing with an initial enzyme-linked immunoabsorbent assay (ELISA) test (Vironostika HIV Uni-Form II Plus O, Biomerieux), followed by a confirmatory ELISA (DADE Behring HIV-1/2) if the initial result was HIV-positive.[1]

Plausibility check of estimated HIV prevalence in men

We tested the plausibility of the finding in our selection model that HIV prevalence in men who did not consent to HIV testing was higher than in survey participants by dividing the 34 interviewers (including the ‘interviewer’ we assigned to all interviewers with fewer than 50 individual interviews) into 17 ‘more successful’ ones (i.e. an interviewer with a consent rate above the median of 0.804) and 17 ‘less successful’ ones (i.e. an interviewer with a consent rate below 0.804). The mean consent rate among 'less successful' interviewers was 73.9% compared to 85.6% among 'more successful' interviewers. We then compared the HIV prevalence rate found by 'more successful' versus 'less successful' interviewers. 'More successful' interviewers found significantly higher HIV prevalence (14.6%) than 'less successful' interviewers (10.6%, *P*-value of difference <0.001). We can think of 'more successful' interviewers as obtaining consent to an HIV test from the same types of men as 'less successful' interviewers, plus from an additional 11.7% of men who usually do not consent when the interviewer is 'less successful'. If this is the case, the observed prevalence rate in the population is the weighted average of the prevalence rate in men who usually consent and those men who consent only to 'more successful' interviewers. The prevalence rate x among those 11.7% of men who only consent when their interviewer is 'more successful' solves the equation

$$0.146 = \frac{0.117 * x + 0.739 * 0.106}{0.117 + 0.739}$$

i.e. a prevalence rate x of 40.0%. This calculation suggests that there must be a very high prevalence rate among men who usually refuse to test but agree to test with a 'more successful' interviewer.

Identification of valid exclusion restrictions

In general, the identification of a valid exclusion restriction involves three steps. First, the researcher must consider which of the variables available in a survey could be associated with survey participation. Second, she must discard variables that could have affected the outcome of interest. This second test will eliminate the overwhelming number of variables. For instance, most respondent or household characteristics recorded in the Demographic and Health Surveys could conceivably have affected HIV status. In addition to the exclusion restrictions used in this study, we did not identify any other plausible selection variables in the Zambia 2007 Demographic and Health Survey. However, in other surveys and contexts other variables than those used in this study may be plausible selection variables. Detailed knowledge of the mechanisms of particular HIV surveys will be helpful in the search for such variables. Differences in time or space in survey conditions or operating procedures may affect contact rates (such as the weather or the types of vehicles fieldworkers use when looking for eligible participants) or consent rates (such as the availability of alternative HIV testing services or incentives for survey participation) without determining HIV status independent of the selection effect. Third, the researcher must test whether the plausible selection variable is indeed significantly associated with survey participation in a selection model, controlling for other observed variables. In the case of multiple dummy variables representing different categories of the same concept, such as interviewer identity, the relevant test for this purpose is one of joint significance of all selection variables.

Limitations of Heckman-type selection models

The Heckman-type selection models used in this study assume that the error terms follow a bivariate normal distribution.[3] Future extensions of this work could include relaxation of this distributional assumption in non-parametric selection models. Furthermore, while some of the imputation approaches assuming "missing at random" can simultaneously predict missing values in any number of variables,[4] Heckman-type selection models can only be used to predict values of the outcome variable. While this condition may be a disadvantage when many variables suffer from large proportions of missing values, in the case of HIV surveys the outcome variable of principal interest, HIV status, is commonly the one variable with the largest proportion of missing values, while only small proportions of explanatory variables (such as sex, age, and education) are missing. Heckman-type selection models are thus likely to be appropriate approaches to control for non-participation in many HIV surveys.

References

1. Central Statistical Office (CSO), Ministry of Health (MOH), Tropical Diseases Research Centre (TDRC), et al. *Zambia Demographic and Health Survey 2007*. Calverton, Maryland: CSO and Macro International Inc.; 2009.

2. ORC Macro. *Demographic and health survey interviewer's manual*. Calverton, Maryland: ORC Macro; 2006.

3. Dubin JA, Rivers D. Selection bias in linear regression, logit and probit models. In: Fox J, Long JS, eds. *Modern methods of data analysis*. Newbury Park: Sage Publications; 1990:410-443.

4. Little R, Rubin DB. *Statistical analysis with missing data*. Hoboken, New Jersey: Wiley; 2002.

eTABLE 1A: Descriptive statistics (men)

		Respondents who consented to HIV testing			Respondents who refused HIV testing	Eligible HH members who did not interview
		HIV+	HIV-	Total		
Variables in contact regressions						
Age category (%)	15-19	5	24	21	22	20
	20-24	6	17	16	18	17
	25-29	14	15	15	16	18
	30-34	22	13	14	14	15
	35-39	20	10	12	10	13
	40-44	15	7	8	6	6
	45-49	10	6	6	6	4
	50-54	5	4	5	4	4
	55-59	3	3	3	3	3
Wealth quintile (%)	Poorest	10	19	18	15	7
	2nd	12	15	15	15	13
	3rd	19	20	20	21	16
	4th	32	24	25	24	33
	Wealthiest	26	21	22	25	31
Educational attainment (mean grade)		8	8	8	8	8
Location (%)	Large city	14	8	9	12	11
	Small city	9	6	7	10	13
	Town	32	27	27	26	37
	Countryside	46	59	58	52	40
Variables in consent regressions						
Married (%)	Yes	72	54	56	55	
Age at first sex (%)	Never had sex	4	14	13	17	
	≤ 15 yrs	33	32	32	28	
	> 15 yrs	63	54	55	55	
Number of partners in last 12 months (%)	None	13	25	24	27	
	One	62	60	60	61	
	Multiple	25	15	16	11	
High risk sex in last 12 months (%)		30	28	29	23	
Condom use at last sex (%)		24	16	17	16	
STD in last 12 months (%)		12	4	5	4	
Smokes tobacco (%)		32	24	25	20	
Drinks alcohol (%)		54	39	41	38	
Knows someone who died of AIDS (%)		64	56	57	52	
Would care for relative with AIDS (%)		98	95	96	94	
Ever tested for HIV (%)		34	22	24	22	
HIV Testing						
Respondent consent rate (%)		100	100	100	0	n/a
HIV status (%)		100	0	13	n/a	n/a
Total *N* in subsample		649	4514	5163	1301	670

n/a = not applicable, STD = sexually transmitted diseases. Percentages may not sum to 100% due to rounding. Some of the descriptive statistics in the colums HIV+, HIV-, and Total are based on samples smaller than the total *N* because of missing values. The samples are defined in Figure 1 and in the text.

eTABLE 1B: Descriptive statistics (women)

		Respondents who consented to HIV testing			Respondents who refused HIV testing	Eligible HH members who did not interview
		HIV+	HIV-	Total		
Variables in contact regressions						
Age category (%)	15-19	8	25	22	23	28
	20-24	15	20	20	22	24
	25-29	24	18	19	19	15
	30-34	22	13	15	14	12
	35-39	16	9	10	10	9
	40-44	9	7	8	7	6
	45-49	5	7	7	6	6
Wealth quintile (%)	Poorest	9	17	16	15	16
	2nd	11	19	17	18	14
	3rd	17	20	20	20	19
	4th	33	23	24	24	25
	Wealthiest	31	21	23	23	26
Educational attainment (mean grades)		7	6	7	6	6
Location (%)	Large city	12	7	8	12	10
	Small city	12	6	7	8	10
	Town	37	27	29	25	29
	Countryside	39	59	56	55	51
Variables in consent regressions						
Married (%)	Yes	55	61	60	61	
Age at first sex (%)	Never had sex	3	15	13	14	
	≤ 15 yrs	34	33	33	29	
	> 15 yrs	63	52	54	57	
Number of partners in last 12 months (%)	None	24	26	25	26	
	One	74	73	73	73	
	Multiple	2	1	1	1	
High risk sex in last 12 months (%)		20	13	14	14	
Condom use at last sex (%)		17	9	10	9	
STD in last 12 months (%)		9	4	5	3	
Smokes tobacco (%)		1	1	1	0	
Drinks alcohol (%)		17	9	10	10	
Knows someone who died of AIDS (%)		61	57	58	55	
Would care for relative with AIDS (%)		98	95	96	94	
Ever tested for HIV		53	39	41	40	
HIV Testing						
Respondent consent rate (%)		100	100	100	0	n/a
HIV status (%)		100	0	17	n/a	n/a
Total *N* in subsample		947	4766	5713	1324	359

n/a = not applicable, STD = sexually transmitted diseases. Percentages may not sum to 100% due to rounding. Some of the descriptive statistics in the colums HIV+, HIV-, and Total are based on samples smaller than the total *N* because of missing values. The samples are defined in Figure 1 and in the text.

eTABLE 2: Consent regressions (women)

		Selection model (bivariate probit)				Imputation model (probit)	
		HIV survey participation		HIV status		HIV status	
		dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Age category	20-24	-0.007	-0.041 - 0.027	0.072	0.024 - 0.121	0.066	0.024 - 0.109
	25-29	0.007	-0.028 - 0.042	0.192	0.132 - 0.252	0.181	0.127 - 0.235
	30-34	0.013	-0.023 - 0.049	0.258	0.193 - 0.323	0.247	0.185 - 0.309
	35-39	0.013	-0.028 - 0.054	0.280	0.211 - 0.349	0.270	0.204 - 0.336
	40-44	0.008	-0.035 - 0.052	0.190	0.117 - 0.260	0.180	0.112 - 0.248
	45-49	0.033	-0.008 - 0.075	0.117	0.047 - 0.188	0.114	0.047 - 0.182
Educational attainment (mean)		0.006	0.002 - 0.010	0.003	-0.001 - 0.007	0.003	0.000 - 0.007
Wealth quintile	2nd	-0.003	-0.038 - 0.033	0.010	-0.032 - 0.051	0.009	-0.029 - 0.048
	3rd	-0.005	-0.044 - 0.033	0.034	-0.011 - 0.080	0.032	-0.010 - 0.074
	4th	0.030	-0.015 - 0.075	0.072	0.016 - 0.128	0.069	0.018 - 0.120
	Wealthiest	0.046	-0.004 - 0.096	0.042	-0.019 - 0.103	0.042	-0.014 - 0.099
Location	Small city	0.147	0.085 - 0.209	0.037	-0.072 - 0.147	0.053	-0.041 - 0.148
	Town	0.148	0.061 - 0.236	-0.005	-0.096 - 0.087	0.009	-0.066 - 0.084
	Countryside	0.179	0.060 - 0.298	-0.071	-0.161 - 0.019	-0.052	-0.116 - 0.013
Married	Yes	-0.043	-0.095 - 0.008	-0.113	-0.167 - -0.059	-0.108	-0.157 - -0.060
Age at first sex	≤ 15 yrs	0.049	0.006 - 0.092	0.198	0.129 - 0.267	0.190	0.125 - 0.255
	> 15 yrs	0.065	-0.040 - 0.053	0.160	0.105 - 0.215	0.148	0.101 - 0.195
Number of partners in last 12 months	One	0.036	-0.023 - 0.096	-0.031	-0.090 - 0.023	-0.026	-0.077 - 0.026
	Multiple	0.099	0.024 - 0.174	0.003	-0.104 - 0.110	0.013	-0.089 - 0.116
High risk sex in last 12 months		-0.048	-0.115 - 0.019	-0.035	-0.086 - 0.016	-0.035	-0.080 - 0.011
Condom use at last sex		0.024	-0.008 - 0.057	0.072	0.029 - 0.116	0.070	0.029 - 0.110
STD in last 12 months		0.062	0.024 - 0.101	0.132	0.075 - 0.189	0.130	0.075 - 0.185
Smokes tobacco		0.106	0.041 - 0.172	-0.082	-0.153 - -0.010	-0.068	-0.126 - -0.009
Drinks alcohol		0.006	-0.030 - 0.041	0.070	0.029 - 0.111	0.065	0.028 - 0.103
Knows someone who died of AIDS		0.013	-0.007 - 0.033	-0.018	-0.043 - 0.006	-0.016	-0.037 - 0.006
Would care for relative with AIDS		0.028	-0.019 - 0.075	0.049	0.000 - 0.098	0.048	0.004 - 0.091
Ever tested for HIV		0.010	-0.012 - 0.031	0.029	0.006 - 0.051	0.027	0.006 - 0.047
Number of observations N		6572				5248	
Censored		1324					
Uncensored		5248					
Correlation between HIV survey participation and HIV status		ρ = -0.26, 95% CI = (-0.65 - 0.23)					
Wald test of independent equations		χ^2 (1) = 1.10, probability > χ^2 = 0.294					
Wald test of exclusion restrictions on HIV survey participation		χ^2(40) = 186.85, probability > χ^2 < 0.001					

eTABLE 2: Consent regressions (women): table subtext

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV-positive status. For dummy variables, the marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster. In addition to the variables shown, the imputation model and each of the two equations in the selection model include dummy variables for language, ethnicity, religion and region. The HIV survey participation equation further includes dummy variables for interviewer identity. See eTables 5A and 5B for effect estimates for the variables not shown in this table.

eTABLE 3: Contact regressions (women)

		Selection model (bivariate probit)				Imputation model (probit)	
		HIV survey participation		HIV status		HIV status	
		dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Age category	20-24	-0.009	-0.037 - 0.019	0.104	0.050 - 0.158	0.105	0.061 - 0.149
	25-29	0.022	-0.006 - 0.050	0.213	0.150 - 0.276	0.213	0.165 - 0.261
	30-34	0.032	0.004 - 0.060	0.280	0.210 - 0.350	0.280	0.224 - 0.337
	35-39	0.019	-0.018 - 0.055	0.303	0.232 - 0.375	0.304	0.245 - 0.363
	40-44	0.034	-0.004 - 0.072	0.226	0.156 - 0.296	0.227	0.164 - 0.290
	45-49	0.047	0.010 - 0.084	0.158	0.090 - 0.227	0.159	0.092 - 0.226
Educational attainment (mean)		0.006	0.002 - 0.010	0.004	0.000 - 0.008	0.004	0.000 - 0.008
Wealth quintile	2nd	0.002	-0.037 - 0.040	0.023	-0.018 - 0.065	0.023	-0.018 - 0.065
	3rd	-0.004	-0.046 - 0.039	0.052	0.008 - 0.097	0.052	0.009 - 0.096
	4th	0.019	-0.029 - 0.068	0.082	0.028 - 0.136	0.082	0.030 - 0.135
	Wealthiest	0.010	-0.048 - 0.068	0.048	-0.008 - 0.104	0.048	-0.008 - 0.105
Location	Small city	0.164	0.100 - 0.227	0.067	-0.046 - 0.180	0.067	-0.034 - 0.169
	Town	0.164	0.081 - 0.246	0.008	-0.083 - 0.098	0.008	-0.071 - 0.086
	Countryside	0.194	0.085 - 0.302	-0.068	-0.164 - 0.029	-0.068	-0.135 - -0.001
Number of observations N		7389				5706	
Censored		1683					
Uncensored		5706					
Correlation between HIV survey participation and HIV status		ρ = -0.00, 95% CI = (-0.65 - 0.65)					
Wald test of independent equations		χ^2 (1) = 0.00, probability > χ^2 = 0.998					
Wald test of exclusion restrictions on HIV survey participation		χ^2(54) = 203.36, probability > χ^2 < 0.001					

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV-positive status. For dummy variables, the marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster. In addition to the variables shown, the imputation model and each of the two equations in the selection model include dummy variables for region. The HIV survey participation equation further includes dummy variables for household visit on the first day of fieldwork in the cluster and for interviewer identity. See eTables 7A and 7B for the variables not shown in this table.

eTABLE 4A: Consent regressions: other independent variables (men)

		Selection model (bivariate probit)				Imputation model (probit)	
		HIV survey participation		HIV status		HIV status	
		dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Language	Bemba	0.026	-0.028 - 0.081	0.010	-0.057 - 0.077	0.015	-0.030 - 0.060
	Lozi	0.085	-0.006 - 0.175	0.075	-0.046 - 0.196	0.097	-0.007 - 0.201
	Nyanja	0.091	0.040 - 0.143	-0.005	-0.072 - 0.063	0.020	-0.021 - 0.061
	Tonga	0.088	0.023 - 0.153	-0.013	-0.103 - 0.078	0.023	-0.040 - 0.087
	Other	0.046	-0.040 - 0.133	-0.015	-0.140 - 0.109	-0.008	-0.101 - 0.084
Ethnicity	Lunda (L)	0.038	-0.053 - 0.130	0.032	-0.053 - 0.118	0.033	-0.030 - 0.096
	Lala	-0.056	-0.145 - 0.032	-0.023	-0.098 - 0.051	-0.030	-0.070 - 0.011
	Ushi	-0.082	-0.179 - 0.016	0.062	-0.038 - 0.163	0.021	-0.051 - 0.093
	Lamba	-0.032	-0.115 - 0.050	-0.067	-0.164 - 0.030	-0.051	-0.095 - -0.007
	Tonga	0.018	-0.032 - 0.068	0.011	-0.049 - 0.072	0.012	-0.028 - 0.052
	Luvale	-0.008	-0.086 - 0.070	0.013	-0.072 - 0.099	0.005	-0.054 - 0.063
	Lunda (NW)	-0.033	-0.131 - 0.064	-0.052	-0.132 - 0.029	-0.042	-0.086 - 0.001
	Mbunda	0.014	-0.092 - 0.120	-0.038	-0.118 - 0.041	-0.026	-0.078 - 0.025
	Kaonde	0.057	-0.007 - 0.121	-0.021	-0.100 - 0.058	0.001	-0.051 - 0.054
	Lozi	0.016	-0.045 - 0.078	0.032	-0.037 - 0.102	0.027	-0.024 - 0.078
	Chewa	-0.008	-0.063 - 0.047	-0.006	-0.054 - 0.042	-0.009	-0.040 - 0.022
	Nsenga	-0.014	-0.077 - 0.049	0.039	-0.024 - 0.103	0.023	-0.022 - 0.068
	Ngoni	0.042	-0.010 - 0.094	-0.031	-0.084 - 0.021	-0.016	-0.051 - 0.019
	Mambwe	0.021	-0.069 - 0.111	-0.049	-0.116 - 0.018	-0.028	-0.064 - 0.007
	Namwanga	0.033	-0.022 - 0.087	-0.061	-0.127 - 0.006	-0.039	-0.076 - -0.002
	Tumbuka	0.111	0.072 - 0.150	-0.084	-0.144 - -0.024	-0.035	-0.066 - -0.004
	Other	-0.021	-0.064 - 0.023	-0.004	-0.042 - 0.033	-0.009	-0.032 - 0.015
Religion	Protestant	-0.003	-0.033 - 0.026	0.023	-0.007 - 0.053	0.016	-0.004 - 0.035
	Muslim	-0.077	-0.147 - -0.008	-0.028	-0.097 - 0.041	-0.036	-0.070 - -0.002
Region	Copperbelt	0.022	-0.093 - 0.138	-0.006	-0.074 - 0.062	-0.023	-0.059 - 0.012
	Eastern	-0.032	-0.177 - 0.113	-0.023	-0.091 - 0.045	-0.014	-0.057 - 0.029
	Luapula	-0.001	-0.308 - 0.306	-0.008	-0.082 - 0.066	0.017	-0.036 - 0.069
	Lusaka	0.109	-0.001 - 0.218	-0.030	-0.106 - 0.045	-0.006	-0.052 - 0.041
	Northern	0.061	-0.025 - 0.147	-0.073	-0.129 - -0.018	-0.042	-0.074 - -0.010
	Northwestern	0.082	-0.022 - 0.187	-0.064	-0.176 - 0.048	-0.036	-0.111 - 0.040
	Southern	0.004	-0.148 - 0.156	-0.005	-0.089 - 0.080	-0.021	-0.066 - 0.024
	Western	0.053	-0.090 - 0.196	-0.062	-0.142 - 0.017	-0.045	-0.088 - -0.002

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation or HIV-positive status. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster. L = luapula, NW = northwestern.

eTABLE 4B: Consent regressions: exclusion restrictions (men)

		Selection model (bivariate probit) HIV survey participation					
		dy/dx	95% CI			dy/dx	95% CI
Individual interviewer	2	0.036	-0.042 - 0.113	Individual interviewer	30	-0.050	-0.230 - 0.130
	3	0.061	-0.007 - 0.129	(continued)	31	-0.040	-0.218 - 0.139
	4	-0.040	-0.165 - 0.085		32	-0.051	-0.259 - 0.157
	5	-0.103	-0.289 - 0.084		33	-0.041	-0.245 - 0.163
	6	-0.058	-0.203 - 0.087		Other	-0.006	-0.094 - 0.082
	7	-0.102	-0.257 - 0.054				
	8	0.014	-0.102 - 0.131				
	9	-0.002	-0.136 - 0.132				
	10	0.119	0.046 - 0.193				
	11	0.060	-0.067 - 0.186				
	12	0.121	-0.049 - 0.291				
	13	0.074	-0.149 - 0.297				
	14	0.068	-0.039 - 0.175				
	15	-0.105	-0.281 - 0.070				
	16	0.030	-0.094 - 0.154				
	17	-0.126	-0.280 - 0.028				
	18	-0.055	-0.183 - 0.072				
	19	0.118	0.036 - 0.201				
	20	0.049	-0.067 - 0.164				
	21	-0.060	-0.221 - 0.102				
	22	-0.111	-0.272 - 0.051				
	23	-0.031	-0.225 - 0.163				
	24	0.060	-0.091 - 0.211				
	25	-0.098	-0.257 - 0.061				
	26	-0.089	-0.258 - 0.079				
	27	-0.075	-0.234 - 0.084				
	28	0.043	-0.087 - 0.173				
	29	-0.123	-0.306 - 0.061				

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster.

eTABLE 5A: Consent regressions: other independent variables (women)

		Selection model (bivariate probit)				Imputation model (probit)	
		HIV survey participation		HIV status		HIV status	
		dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Language	Bemba	0.025	-0.031 - 0.081	0.027	-0.038 - 0.092	0.024	-0.035 - 0.083
	Lozi	0.032	-0.152 - 0.217	-0.046	-0.181 - 0.088	-0.039	-0.161 - 0.082
	Nyanja	0.062	0.012 - 0.113	0.073	0.005 - 0.140	0.073	0.012 - 0.134
	Tonga	0.069	0.019 - 0.119	0.110	0.028 - 0.193	0.109	0.032 - 0.187
	Other	0.028	-0.071 - 0.127	-0.043	-0.168 - 0.081	-0.037	-0.145 - 0.071
Ethnicity	Lunda (L)	0.096	-0.017 - 0.210	-0.030	-0.114 - 0.055	-0.025	-0.098 - 0.048
	Lala	-0.056	-0.150 - 0.038	0.025	-0.068 - 0.118	0.014	-0.065 - 0.093
	Ushi	0.044	-0.019 - 0.107	-0.058	-0.132 - 0.015	-0.052	-0.117 - 0.014
	Lamba	-0.005	-0.078 - 0.067	0.062	-0.037 - 0.161	0.055	-0.036 - 0.146
	Tonga	0.027	-0.026 - 0.081	-0.048	-0.093 - -0.004	-0.043	-0.082 - -0.004
	Luvale	-0.062	-0.144 - 0.020	0.042	-0.050 - 0.134	0.033	-0.049 - 0.114
	Lunda (NW)	-0.080	-0.173 - 0.014	-0.004	-0.105 - 0.096	-0.012	-0.102 - 0.078
	Mbunda	-0.016	-0.120 - 0.088	-0.041	-0.111 - 0.030	-0.039	-0.101 - 0.023
	Kaonde	-0.012	-0.084 - 0.059	-0.015	-0.087 - 0.056	-0.017	-0.079 - 0.045
	Lozi	-0.022	-0.090 - 0.047	0.033	-0.032 - 0.098	0.027	-0.030 - 0.084
	Chewa	-0.033	-0.096 - 0.031	0.033	-0.037 - 0.103	0.027	-0.036 - 0.089
	Nsenga	-0.033	-0.095 - 0.028	0.085	0.015 - 0.154	0.075	0.013 - 0.137
	Ngoni	-0.001	-0.073 - 0.071	0.037	-0.037 - 0.111	0.033	-0.034 - 0.100
	Mambwe	0.014	-0.080 - 0.107	0.038	-0.053 - 0.128	0.037	-0.046 - 0.121
	Namwanga	0.060	0.010 - 0.110	-0.047	-0.106 - 0.012	-0.042	-0.092 - 0.008
	Tumbuka	0.041	-0.022 - 0.104	-0.036	-0.094 - 0.022	-0.031	-0.082 - 0.020
	Other	0.010	-0.035 - 0.055	0.016	-0.030 - 0.062	0.015	-0.028 - 0.057
Religion	Protestant	0.024	-0.008 - 0.056	0.016	-0.011 - 0.044	0.017	-0.008 - 0.042
	Muslim	0.030	-0.049 - 0.110	-0.016	-0.109 - 0.076	-0.013	-0.098 - 0.071
Region	Copperbelt	0.066	-0.219 - 0.351	-0.062	-0.111 - -0.014	-0.058	-0.099 - -0.016
	Eastern	0.171	0.078 - 0.265	-0.108	-0.159 - -0.057	-0.095	-0.130 - -0.059
	Luapula	0.087	0.015 - 0.159	-0.066	-0.123 - -0.009	-0.051	-0.100 - -0.003
	Lusaka	0.110	-0.025 - 0.245	-0.085	-0.156 - -0.014	-0.071	-0.125 - -0.017
	Northern	0.142	0.045 - 0.239	-0.103	-0.149 - -0.056	-0.087	-0.125 - -0.050
	Northwestern	0.075	-0.108 - 0.257	-0.037	-0.186 - 0.112	-0.032	-0.160 - 0.096
	Southern	0.086	-0.059 - 0.231	-0.083	-0.130 - -0.037	-0.075	-0.113 - -0.037
	Western	0.052	-0.124 - 0.228	0.042	-0.137 - 0.221	0.045	-0.124 - 0.213

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation or HIV-positive status. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster. L = luapula, NW = northwestern.

eTABLE 5B: Consent regressions: exclusion restrictions (women)

		Selection model (bivariate probit)					
		HIV survey participation					
		dy/dx	95% CI			dy/dx	95% CI
Individual interviewer	2	-0.078	-0.160 - 0.005	Individual interviewer	30	-0.220	-0.491 - 0.051
	3	-0.019	-0.106 - 0.069	(continued)	31	-0.408	-0.687 - -0.130
	4	-0.231	-0.769 - 0.306		32	-0.345	-0.631 - -0.058
	5	-0.145	-0.639 - 0.348		33	-0.273	-0.560 - 0.014
	6	-0.100	-0.563 - 0.363		34	-0.029	-0.258 - 0.200
	7	-0.155	-0.651 - 0.342		35	-0.012	-0.221 - 0.197
	8	-0.335	-0.865 - 0.194		36	-0.036	-0.259 - 0.187
	9	-0.070	-0.449 - 0.308		37	-0.177	-0.441 - 0.086
	10	-0.093	-0.555 - 0.369		38	-0.178	-0.437 - 0.082
	11	-0.172	-0.682 - 0.339		39	-0.106	-0.366 - 0.154
	12	-0.391	-0.749 - -0.033		40	-0.163	-0.421 - 0.096
	13	-0.251	-0.578 - 0.076		41	-0.197	-0.491 - 0.096
	14	-0.241	-0.577 - 0.095		Other	-0.122	-0.295 - 0.050
	15	-0.186	-0.517 - 0.145				
	16	-0.102	-0.208 - 0.005				
	17	-0.094	-0.195 - 0.006				
	18	0.123	0.074 - 0.173				
	19	0.017	-0.182 - 0.216				
	20	0.032	-0.135 - 0.200				
	21	0.024	-0.146 - 0.193				
	22	-0.167	-0.424 - 0.091				
	23	0.071	-0.104 - 0.245				
	24	-0.055	-0.271 - 0.161				
	25	-0.292	-0.566 - -0.017				
	26	-0.040	-0.278 - 0.198				
	27	0.023	-0.226 - 0.271				
	28	0.082	-0.076 - 0.240				
	29	0.003	-0.202 - 0.209				

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster.

eTABLE 6A: Contact regressions: other independent variables (men)

		Selection model (bivariate probit)				Imputation model (probit)	
		HIV survey participation		HIV status		HIV status	
		dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Region	Copperbelt	-0.096	-0.252 - 0.060	-0.039	-0.077 - 0.000	-0.036	-0.066 - -0.006
	Eastern	0.170	0.070 - 0.270	-0.031	-0.090 - 0.028	-0.020	-0.052 - 0.012
	Luapula	0.047	-0.099 - 0.193	0.012	-0.035 - 0.058	0.014	-0.026 - 0.054
	Lusaka	0.174	0.048 - 0.299	-0.005	-0.070 - 0.060	0.008	-0.031 - 0.046
	Northern	0.054	-0.106 - 0.214	-0.068	-0.127 - -0.009	-0.054	-0.080 - -0.028
	Northwestern	0.039	-0.055 - 0.133	-0.080	-0.139 - -0.020	-0.066	-0.090 - -0.041
	Southern	0.127	0.010 - 0.243	-0.003	-0.052 - 0.046	0.001	-0.039 - 0.041
	Western	0.091	-0.051 - 0.233	0.008	-0.046 - 0.062	0.015	-0.026 - 0.056

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation or HIV-positive status. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster.

eTABLE 6B: Contact regressions: exclusion restrictions (men)

		Selection model (bivariate probit)					
		HIV survey participation					
		dy/dx	95% CI			dy/dx	95% CI
First day		0.015	-0.012 - 0.041				
Household interviewer	2	0.000	-0.101 - 0.100	Household interviewer	30	-0.261	-0.471 - -0.050
	3	0.048	-0.044 - 0.139	(continued)	31	0.111	0.008 - 0.215
	4	0.049	-0.118 - 0.215		32	0.152	0.060 - 0.243
	5	0.119	0.007 - 0.232		33	0.104	-0.093 - 0.301
	6	-0.049	-0.244 - 0.146		34	0.136	0.015 - 0.257
	7	-0.064	-0.249 - 0.121		35	-0.020	-0.214 - 0.174
	8	0.095	-0.021 - 0.211		36	0.017	-0.210 - 0.244
	9	0.002	-0.157 - 0.162		37	-0.112	-0.353 - 0.130
	10	0.130	0.020 - 0.239		38	-0.067	-0.336 - 0.201
	11	0.064	-0.088 - 0.215		39	-0.208	-0.472 - 0.055
	12	0.031	-0.134 - 0.197		40	-0.096	-0.321 - 0.130
	13	-0.118	-0.289 - 0.053		41	-0.005	-0.162 - 0.151
	14	-0.158	-0.272 - -0.044		42	-0.187	-0.368 - -0.006
	15	0.022	-0.093 - 0.137		43	-0.110	-0.290 - 0.071
	16	-0.082	-0.243 - 0.079		44	0.066	-0.089 - 0.220
	17	-0.140	-0.298 - 0.018		45	0.009	-0.164 - 0.182
	18	0.060	-0.085 - 0.206		46	0.017	-0.159 - 0.193
	19	-0.017	-0.179 - 0.145		Other	-0.002	-0.099 - 0.096
	20	-0.018	-0.240 - 0.203				
	21	-0.001	-0.167 - 0.165				
	22	0.048	-0.132 - 0.227				
	23	-0.128	-0.325 - 0.070				
	24	0.085	-0.050 - 0.219				
	25	0.008	-0.173 - 0.188				
	26	0.122	-0.014 - 0.258				
	27	0.040	-0.126 - 0.207				
	28	-0.037	-0.196 - 0.123				
	29	-0.069	-0.242 - 0.104				

First day = household visit on the first day of fieldwork in a survey cluster. dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster.

eTABLE 7A: Contact regressions: other independent variables (women)

		Selection model (bivariate probit)				Imputation model (probit)	
		HIV survey participation		HIV status		HIV status	
		dy/dx	95% CI	dy/dx	95% CI	dy/dx	95% CI
Region	Copperbelt	-0.174	-0.347 - -0.002	-0.070	-0.112 - -0.028	-0.070	-0.108 - -0.032
	Eastern	0.162	0.069 - 0.256	-0.061	-0.124 - 0.002	-0.062	-0.099 - -0.025
	Luapula	-0.135	-0.365 - 0.094	-0.070	-0.128 - -0.012	-0.071	-0.108 - -0.034
	Lusaka	-0.031	-0.198 - 0.137	-0.032	-0.112 - 0.048	-0.032	-0.095 - 0.031
	Northern	-0.015	-0.166 - 0.136	-0.094	-0.149 - -0.039	-0.094	-0.126 - -0.062
	Northwestern	-0.175	-0.585 - 0.234	-0.082	-0.132 - -0.032	-0.082	-0.117 - -0.047
	Southern	-0.055	-0.242 - 0.132	-0.049	-0.098 - 0.001	-0.049	-0.088 - -0.010
	Western	0.065	-0.101 - 0.232	-0.007	-0.064 - 0.050	-0.007	-0.054 - 0.040

dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation or HIV-positive status. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster.

eTABLE 7B: Contact regressions: exclusion restrictions (women)

		Selection model (bivariate probit) HIV survey participation					
		dy/dx	95% CI			dy/dx	95% CI
First day		0.047	0.023 - 0.071				
Household interviewer	2	-0.001	-0.098 - 0.096	Household interviewer	30	0.142	0.064 - 0.220
	3	-0.092	-0.229 - 0.044	(continued)	31	0.109	0.003 - 0.215
	4	0.005	-0.115 - 0.126		32	0.159	0.087 - 0.231
	5	0.109	0.001 - 0.217		33	0.178	0.113 - 0.243
	6	0.030	-0.124 - 0.184		34	0.095	-0.044 - 0.234
	7	0.093	-0.053 - 0.239		35	0.184	0.102 - 0.267
	8	0.137	0.051 - 0.223		36	0.182	0.127 - 0.237
	9	0.125	0.021 - 0.230		37	-0.015	-0.192 - 0.162
	10	0.152	0.074 - 0.230		38	0.040	-0.115 - 0.196
	11	0.036	-0.111 - 0.183		39	-0.016	-0.217 - 0.185
	12	0.090	-0.037 - 0.218		40	-0.033	-0.257 - 0.190
	13	-0.152	-0.314 - 0.009		41	-0.026	-0.212 - 0.161
	14	-0.141	-0.299 - 0.017		42	0.142	-0.015 - 0.300
	15	-0.041	-0.173 - 0.091		43	0.154	0.008 - 0.300
	16	-0.049	-0.181 - 0.084		44	0.177	0.070 - 0.284
	17	-0.080	-0.259 - 0.100		45	0.066	-0.220 - 0.351
	18	0.186	0.118 - 0.253		46	0.049	-0.225 - 0.324
	19	0.127	0.010 - 0.244		47	0.083	-0.155 - 0.320
	20	0.127	0.003 - 0.250		48	0.125	0.004 - 0.246
	21	0.168	0.098 - 0.238		49	0.074	-0.073 - 0.221
	22	0.181	0.115 - 0.247		50	0.044	-0.115 - 0.204
	23	0.202	0.169 - 0.236		51	0.049	-0.138 - 0.235
	24	0.168	0.091 - 0.244		52	0.027	-0.164 - 0.218
	25	0.141	0.050 - 0.232		53	-0.010	-0.218 - 0.199
	26	0.162	0.090 - 0.235		Other	0.045	-0.052 - 0.143
	27	0.086	-0.035 - 0.206				
	28	0.174	0.118 - 0.230				
	29	0.162	0.083 - 0.241				

First day = household visit on the first day of fieldwork in a survey cluster. dy/dx = marginal effects evaluated at the sample mean; the effects are expressed as absolute change in the probability of HIV survey participation. The marginal effects represent the probability change in response to the discrete change of the dummy variable value from zero to one. All confidence intervals (CI) are based on standard errors that are adjusted for clustering at the level of the Demographic and Health Survey cluster.