Demography India (2015)
Vol.44, Issue: 1&2, pp: 17-30

ISSN: 0970-454X

Research Article

Correcting Estimates of HIV Prevalence Due to Survey Non-Participation in India Using Heckman Selection Model

Shrikant Singh[1,*], Swati Srivastava[2] & Ashish Kumar Upadhyay[2]

Abstract

Using the data from the third round of National Family Health Survey and Heckman Selection Model this paper aims to determine the estimates of HIV prevalence in India due to survey non-participation. Interviewer ID was taken as the selection variable, which affects the survey participation but did not affect HIV status independently. Study also compared the estimates of Heckman selection model to conventional imputation model. It has been found that prevalence of HIV is greater among men (0.77; 95% CI= (0.71-0.83)) and women (0.42; 95% CI= (0.39-0.45)), who did not participate in the survey as compare to those who participated in HIV test (0.35 for men & 0.22 for women). Thus, the national estimate for men and women derived from selection model was higher than the unadjusted imputation method. Results of this study demonstrate that the selection variable was significantly associated with the HIV status of the men and women. Further, this study shows the significant association between the survey participation and the HIV status of those who has been interviewed but did not consent to the HIV test, which clarifies that the sample selection led to substantial underestimation of the national HIV prevalence in men and women. Therefore, a valid and efficient way to provide the estimate of HIV prevalence is to incorporate the Heckman selection model instead of the conventional method to provide an estimate of the national prevalence.

Introduction

HIV/AIDS epidemic is one of the critical public health challenges for several developing countries, resulting in gaining significant priority while designing goals and targets in the course of transition from Millennium Development Goals (MDGs) to Sustainable Development Goals (SDGs). Data from various cross-sectional surveys portray the existence of multiple epidemic in most of the countries suffering with concentrated epidemic with diverse population and social networks. Identifying infected and affected population and targeting interventions to match the needs of such population becomes difficult, especially when such groups are socially marginalized and discriminated. Stigmatized populations are frequently hidden and often hard to reach with any service need assessment or other cross-sectional surveys including Demographic and Health Survey (DHS). It is within this context, precise and accurate estimates of HIV prevalence are challenge for the health planners to provide health services, to track the recourse of the epidemic and to assess the effectiveness and efficiency of various targeted interventions designed to change the recourse of the epidemic. All the countries have conducted some HIV surveillance to capture the trend and pattern of the epidemic. Latest release of World Health Organization (WHO) reveals that, globally 36.7 million people were living with HIV at the end of 2015. Worldwide an estimated 0.8 percent of adult aged 15-49 years were living with HIV, although the burden of the epidemic continues to vary considerably between countries and regions. In the Sub-Saharan Africa nearly 1 in every 25 adults (4.4%) were living with HIV and accounting for nearly 70 percent of the people living with HIV worldwide (WHO, 2015). Latest report of UNAIDS shows that, in India about 2.1 million people (including children) were living with HIV. However 0.86 million population newly infected with HIV in 2015(UNAIDS, 2015).

[1]Professor, Department of Mathematical Demography and Statistics, International Institute for Population Sciences, Mumbai
[2]Ph. D. Fellow, International Institute for Population Sciences, Mumbai
*Corresponding author: Shrikant Singh, Email: sksingh31962@gmail.com

In India, HIV/AIDS is embedded as a concentrated epidemic, i.e. infection of HIV/AIDS is intense in particular groups like, female sex workers, male sex workers, men having sex with men, injecting drug users and bridge population (migrants and truckers), which called as the high-risk groups. In India earlier to 1990, socio-economic and other information about 'high risk population' were collected from sentinel surveillance sites, however for 'general population' pregnant women were considered as the proxy, because the women in age 15-49 were found to be sexually active and these communities were readily available through routine ante natal care (ANC) visits (Barnighausen, Bor, Wandira-Kazibwe, & Canning, 2011; Hogan et al., 2012). But the estimates of HIV prevalence obtained from the pregnant women were biased and not good enough as it confronted some problems of overestimation (Boerma, Ghys, & Walker, 2003; Gregson et al., 2002) and underestimation (Gouws, Mishra, & Fowler, 2008). For example-1) the numbers of pregnant women who used to go for ANC services might be low compared to an actual number of pregnant women. 2) The likelihood of HIV infection among pregnant women significantly differed from non-pregnant women. 3) In general, women belong to rural areas, lower educational level and young age group (15-24 years) were considered too frequently get pregnant than their other counterparts. Also, these sentinel surveillances could not provide information about HIV status among men. Therefore, an enormous part of the population was missing. 4) Another reason of underestimation was that, if the women had the symptoms of HIV then she will be the physiologically week to be pregnant compared to other women in the same age groups. 5) Furthermore, the geographic area of sentinel sites was also imperative, because antenatal coverage was not same around the country (Gouws et al., 2008). Therefore, the information about all eligible respondents was not available in a population-based survey which creates the incomplete framework of data, which termed as missing data. Missing data were classified into three categories, namely 1) missing completely at random (MCAR) 2) missing at random (MAR) and 3) missing not at random (MNAR). Classification of missing data depends on the relation between measured variables and the probability of missing data (Kalton, 1996; Roderick J. A. Little, 1987; Rubin, 1987).

In general, the non-participation of the respondents was classified as missing at random. If HIV status was indeed missing at random, then techniques of imputations and multiple imputations were used to deal the selection bias, which associated with the survey participation (Brick & Kalton, 1996; Chinomona & Mwambi, 2015; Rubin, 1987).However, sometimes non-participation of respondents may be "missing not at random" provided; there were some unobserved variables associated with HIV status. For example, if respondents are already aware of their HIV status (whether they had done the test in past or they know that they have some symptom of HIV) they were less likely to consent to HIV test due to fear of confirmation of their status (Kalichman & Simbayi, 2003; Kranzer et al., 2008; Weiser et al., 2006). This non-participation of respondent for HIV testing may result in underestimation of national HIV prevalence. To get a more precise estimate of the HIV epidemic, first time third round of National Family Health Survey (NFHS-3) collected information about the prevalence of HIV/AIDS as part of a national level household survey. Despite the fact that the contribution of ANC surveillance was appreciable, NFHS-3 was considered as the most accurate data on HIV status of general population due to its large and reliable sample size and information collected in community setting (Marsh, Mahy, Salomon, & Hogan, 2014). Many previous studies have used Heckman selection model to the estimates of national HIV prevalence (Barnighausen et al., 2011; Bignami-Van Assche, Salomon, & Murray, 2005; Hogan et al., 2012; Janssens, van der Gaag, Rinke de Wit, & Tanovic, 2014; Lachaud, 2007; Reniers, Araya, Berhane, Davey, & Sanders, 2009).The majority of the studies were conducted in sub-Saharan African countries. However, no study used Heckman selection model to the estimates of HIV prevalence in India. Consequently, to the best of our knowledge, this is the first study, which uses Heckman correction type model and incorporates to revise the prevalence of national HIV estimates by using the third round of NFHS, 2005-06.

Data and methodology

Survey Data

This study used the data of the Indian Demographic Health Survey (IDHS), which also known as National Family Health Survey (NFHS). NFHS-3 (2005-2006) is the first nationwide survey, which

intended to provide information about the HIV estimates among men (15-54 years) and women (15-49 years) in their reproductive age. For the data collection, interviewer team first completed the household questionnaire with one household member and ensured the eligibility of other household members for HIV test by asking age of individuals residing in household. Then selected eligible individuals were given a consent form, in which they consented whether they were ready to give blood sample for HIV test or not. A total of 62,182 women (age-group 15-49) and 64,175 men (age-group 15-54) were eligible for HIV testing. Out of 62182 eligible women 52855 women (85%) were give the blood sample however out of 64175 men about 54549 men (78%) has given the blood sample for HIV test. In NFHS-3 survey, 6 percent of women and 14 percent of men did not complete the individual interview; therefore, they were not eligible for the blood test. However, 6 percent of women and 5 percent of men completed the individual interview but did not provide the blood sample for HIV test (IIPS & Macro., 2007). Detailed information about sample selection procedure and sampling design were available in the NFHS-3 report (IIPS & Macro., 2007).

Analytical approach

This study used three approaches to handle missing HIV test following the analytical approach of Barnighausen et al. (Barnighausen et al., 2011). Three approaches are as follow:1) First, an unadjusted complete case analysis in which missing observations are ignored, and prevalence is calculated among those with valid HIV tests. 2) second, conventional imputation method by using probit regression model, in which missing observation depends on observed variable and 3) third is Heckman selection model in which HIV status of the respondent will depend on observed as well as unobserved variables.

Selection variable

Similar to Barnighausen et al. approach 'Interviewer identity' has been taken as the selection variable to predict the participation of respondents in HIV testing. The identification of a valid selection variable involves in three steps. First, plausible selection variables available in a survey could be associated with survey participation. Second, it must discard variables that could have affected the outcome of interest. Third, plausible selection variable is indeed significantly associated with survey participation in a selection model, controlling for other observed variables.

In this study two terms have often been used- 1) contact regression and 2) consent regression.

Contact regression-The identity of the interviewer, who fails to reach the eligible household, the identity of that interviewer, was considered as selection variable for the regression model.

Consent regression- Interviewers who conducted the individual interview, in which respondent refused to provide blood sample; the identity of that interviewer has been considered as the selection variable for the regression model.

The study has a fundamental assumption that the selection variable (interviewer ID) was associated with survey participation but not to HIV status. The statistical significance association between selection variable and survey participation has been justified by using the Wald test separately for consent regression and contact regression for both sexes.

An overview of Heckman selection model

Heckman selection model is the two-stage statistical method, which is used to determine the selection bias in non-randomly selected sample. The selection model is a bivariate probit regression model which comprises two equations; 1) selection equation and 2) outcome equation. Both the equations are linked with a correlation parameter rho (ρ), which denotes the association between error terms of selection and outcome equation. Correlation parameter may or may not be associated. If parameter rho (ρ) is zero, then there is no association between selection and outcome equation. Which implies that outcome will depend only on observed characteristics, therefore, conventional imputation method, based on an observed variable will be good enough. However, if parameter rho (ρ) is nonzero, which means, selection equation must change the conditional distribution of outcome equation.

Selection of respondents

The information on HIV status for eligible respondent during survey was collected in following way-

1. The interviewer tried to find eligible respondents, whether found (yes, no).
2. If eligible respondents found, they had given the interview (yes, no).
3. If interviewed, they provide a blood sample (yes, no).
4. If blood sample was given, what is the result of the test (yes, no).

Based on the framework as mentioned above, we divided the whole sample into three subgroups.

1. Individuals were contacted, interviewed, and consented to HIV test (valid HIV/complete case)
2. Individuals were contacted, interviewed, and refused to give a blood sample for HIV test (consent group).
3. Individuals were contacted and denied for household interview/ interviewer became fail to communicate with the individual (contact group).

As per the biomarker information was available only for those respondents who were captured by the interviewer, agreed to give a blood sample for HIV test. Therefore, the national estimate of HIV prevalence in India was based only on the first subgroup in NFHS- 3, which created selection bias. This selection bias was incorporated into an analytical approach namely, Barnighausen et al. approach, which helped to estimate prevalence for HIV status in missing population and the overall national estimate.

Selection models and estimation

The aim of present study was to correct HIV prevalence among men and women due to survey nonparticipation, which was correlated with the unobserved characteristic of HIV status. However, the study has information about HIV status among only those respondents who have consented for HIV test. Therefore, the respondents who were selected non-randomly from the population produced the selection bias in the sample. Therefore, a model was introduced, where survey nonparticipation was associated with HIV status (Barnighausen et al., 2011). Heckman correction model was the appropriate model, as it worked in two steps to the selection bias due to survey non-participation. In the first stage, a probit model was formulated for HIV status. Probit regression had specified this relationship

$$h_i^* = x_i Y + u_i \qquad \begin{cases} h_i = 1 & \text{if } h_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

Where h_i^* was the unobserved latent variables which determine the likelihood of being HIV positive, x_i was the observed covariates and u_i was the error term. Here the unobserved latent variables will depend on respective covariates and the random error term.

Now, in the second stage, probit model for survey participation, namely selection model has been formulated

$$s_i^* = x_i \beta + z_i Y + v_i \qquad \begin{cases} s_i = 1 & \text{if } s_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Where s_i^* was the unobserved latent variables which determined the selection of eligible respondents for the interview, xi was the observed covariates, z_i was the exclusion restriction and v_i was the error term. Similarly, here also, unobserved latent variables will depend on several unobserved covariates, exclusion restrictions, and a random error term. The intuition behind bivariate probit model was that; there must be a correlation between unobserved error terms that affects HIV status and HIV survey participation

$$\rho \text{ (rho)} = \text{corr}(u_i, v_i) \qquad (3)$$

Two models for survey nonparticipation were run separately for men and women. First one was consent regression, in which eligible respondents refused to consent to HIV test and the second one was contact regression in which the interviewer did not approach eligible respondents. Therefore, there were four models; consent regression for men, contact regression for men, consent regression for women, contact regression for women.

Let M is the total number of eligible male respondent. Suppose M_1 is the number of respondents who were interviewed and out of M_1, '*x*' individual gave the blood sample. Then the observed HIV status for individual '*x*' is E_1. Again, among the M_2 respondent who were interviewed but refused to give blood samples, predicted probability of being HIV+ for '*y*' individual is E_2. Similarly, among M_3 respondent who were not contacted by interviewer, predicted probability of being infected with HIV for 'z' individual is E_3. Then the estimates of HIV prevalence for population M is

$E_M = (E_1+E_2+E_3)*1/M$

$$= \{\underbrace{\sum_{M_1=0}^{x}(H)x}_{E_1} + \underbrace{\sum_{M_2=0}^{x}\Pr(HIV+/Notconcent)y}_{E_2} + \underbrace{\sum_{M_3=0}^{x}\Pr(HIV+/Notconcent)z}_{E_3}/M \quad (4)$$

Similarly, let F is the total number of eligible female respondent. Let F_1 is the number of female respondent who were interviewed, and out of F_1, 'a' individuals gave the blood sample. Then the observed HIV status for individual 'a' is $\grave{E}_1$. Again, among the F_2 respondent who were interviewed but refused to give blood samples, predicted probability of being HIV+ for 'b' individual is $\grave{E}_2$. Similarly, among F_3 respondent who were not contacted by interviewer, predicted probability of being infected with HIV for 'c' individual is $\grave{E}_3$. Then the estimates of HIV prevalence for population F is

$E_F = (\grave{E}_1+\grave{E}_2+\grave{E}_3)*1/F$

$$= \{\underbrace{\sum_{F_1=0}^{a}(H)a}_{\grave{E}_1} + \underbrace{\sum_{F_2=0}^{a}\Pr(HIV+/Notconcent)b}_{\grave{E}_2} + \underbrace{\sum_{F_3=0}^{a}\Pr(HIV+/Notconcent)c}_{\grave{E}_3}/F \quad (5)$$

All the analysis has been done in STATA, version 13, software.

Results

Table 1 and Table 2 show the distribution of coverage of sampled men and women by their background characteristic in India. Results indicate that while 13987 (22%) men and 9273 (15%) women were eligible, but the interviewer did not elicit blood samples of those for HIV test (either they did not consent to the blood test, or the interviewer did not contact them). Tables3 and 5 present the results of consent regression (where the respondent did not consent to HIV test) for men and women respectively; however, Tables 4 and 6 portray the results of contact regression (where the interviewer did not approach respondent) for men and women respectively.

In selection model following statistics are the important to interpret the result. First, is the Wald test for exclusion restrictions on HIV survey participation, which determined whether there is indeed an association between selection variable and HIV consent. In all four tables, null hypothesis "not have significant effects on consent" is rejected. Therefore, selection variable is the statistically significant factor in determining the survey participation ($P<0.001$, for all model), which confirms that the exclusion restriction is indeed necessary to determine the association with HIV testing.

Table 1: Percent distribution of interviewed men (15-54 years) by HIV testing status, according to background characteristics, India, 2005-06

Background characteristics	Respondents who consented to HIV testing			Respondents who refused HIV testing	Eligible HH members who did not interview
	HIV -	HIV+	Total	Total	Total
Variables taken in contact regression					
Age (in years)					
15-24	33.1	9.7	33.1	34.5	37.3
25-34	27.0	40.0	27.0	27.8	27.1
35-44	23.7	31.5	23.8	22.6	21.1
45-54	16.2	18.9	16.2	15.1	14.5
Education					
No education	19.5	26.7	19.5	21.4	21.3
Primary education	16.0	21.6	16.0	11.9	13.5
Secondary education	17.6	17.5	17.6	14.9	17.1
Higher education	47.0	34.2	46.9	51.9	48.1
Place of residence					
Urban	35.4	40.6	35.4	60.9	48.3
Rural	64.6	59.4	64.6	39.1	51.7
Wealth quintiles					
Poorest	16.1	17.3	16.1	12.5	15.9
Poorer	18.5	15.3	18.5	11.7	17.5
Middle	20.8	20.8	20.8	15.4	14.6
Richer	21.8	31.1	21.8	23.6	21.2
Richest	22.8	15.5	22.8	36.8	30.8
Variables taken in consent regression					
Exposure to mass media					
No	6.4	5.3	6.4	6.7	
Yes	93.6	94.7	93.6	93.3	
Alcohol consumption					
No	67.2	62.1	67.2	74.1	
Yes	32.8	37.9	32.8	25.9	
Marital Status					
Never married	33.7	15.8	33.6	38.9	
Ever married	66.3	84.2	66.4	61.1	
Number of lifetime					
sexual partners					
One	80.6	70.6	80.5	86.6	
More than one	19.4	29.4	19.5	13.5	
Condom use in the last sex					
No	91.2	92.8	91.2	88.9	
Yes	8.8	7.2	8.8	11.1	
Any symptom of STI					
No	99.7	99.0	99.7	99.8	
Yes	0.3	1.0	0.3	0.2	
Total	**51355**	**286**	**51641**	**4057**	**6660**

Table 2: Percent distribution of interviewed women (15-49 years) by HIV testing status, according to background characteristics, India, 2005-06

Background characteristics	Respondents who consented to HIV testing			Respondents who refused HIV testing	Eligible HH members who did not interview
	HIV -	HIV+	Total	Total	Total
Variables taken in contact regression					
Age (in years)					
15-24	36.6	20.1	36.6	38.6	42.6
25-34	31.4	40.6	31.4	30.1	29.3
35 -44	24.0	32.9	24.0	22.6	22.1
45-49	8.0	6.4	8.0	8.8	6.0
Education					
No education	40.2	51.0	40.2	42.7	41.1
Primary education	14.7	21.7	14.7	10.1	12.7
Secondary education	14.9	13.5	14.9	12.3	10.2
Higher education	30.3	13.8	30.2	34.9	36.1
Place of residence					
Urban	32.7	43.4	32.8	52.8	47.0
Rural	67.3	56.6	67.2	47.2	53.0
Wealth quintiles					
Poorest	17.0	13.8	17.0	14.0	19.7
Poorer	19.1	17.5	19.1	16.5	16.5
Middle	20.7	23.3	20.7	16.4	16.2
Richer	20.9	33.2	21.0	17.6	20.7
Richest	22.3	12.3	22.3	34.7	26.9
Variables taken in consent regression					
Exposure to mass media					
No	22.2	20.4	22.2	24.0	
Yes	77.8	79.6	77.8	76.1	
Alcohol consumption					
No	97.9	97.7	97.9	98.7	
Yes	2.1	2.4	2.1	1.3	
Marital status					
Never married	20.4	3.3	20.3	23.8	
Ever married	79.6	96.8	79.7	76.2	
Number of lifetime					
sexual partners					
One	98.2	92.2	98.2	98.4	
More than one	1.8	7.8	1.8	1.6	
Condom use in the last sex					
No	93.9	98.1	93.9	90.9	
Yes	6.1	2.0	6.1	9.1	
Any symptom of STI					
No	98.7	98.6	98.7	99.3	
Yes	1.3	1.4	1.3	0.7	
Total	**53822**	**194**	**54016**	**4704**	**3700**

Table 3: Consent regression for men (15-54 years) for the non-participation due to refusal to give blood sample, India, 2005-06

Background Variables	Heckman selection model (Bivariate probit)				Imputation model (Probit)	
	HIV survey participation		HIV status		HIV survey participation	
	Coefficient	95%C.I.	Coefficient	95%C.I.	Coefficient	95%C.I.
Age (in years)						
15-24	-	-	-	-	-	-
25-34	0.287	(0.017,0.557)	0.017	(-0.039,0.072)	0.300	(0.024,0.575)
35-44	0.335	(0.059,0.611)	0.020	(-0.036,0.076)	0.350	(0.069,0.631)
45-54	0.098	(-0.191,0.387)	0.033	(-0.025,0.091)	0.114	(-0.182,0.409)
Education						
No education	-	-	-	-	-	-
Primary education	0.090	(-0.082,0.262)	0.039	(-0.008,0.087)	0.099	(-0.078,0.277)
Secondary education	0.055	(-0.132,0.241)	0.005	(-0.045,0.056)	0.056	(-0.135,0.248)
Higher education	0.043	(-0.134,0.220)	0.042	(-0.006,0.091)	0.059	(-0.122,0.241)
Wealth status						
Poorest	-	-	-	-	-	-
Poor	-0.076	(-0.302,0.151)	0.069	(0.009,0.129)	-0.062	(-0.295,0.171)
Middle	-0.052	(-0.277,0.173)	0.176	(0.114,0.238)	-0.032	(-0.261,0.199)
Rich	0.031	(-0.203,0.265)	0.094	(0.027,0.161)	0.037	(-0.202,0.277)
Richest	-0.163	(-0.422,0.096)	-0.051	(-0.125,0.023)	-0.195	(-0.459,0.068)
Place of residence						
Urban	-	-	-	-	-	-
Rural	-0.127	(-0.257,0.003)	-0.006	(-0.061,0.049)	-0.132	(-0.263,-0.001)
Number of sexual partner in last 12 months						
One	-	-	-	-	-	-
More than one	0.174	(0.053,0.296)	0.081	(0.043,0.118)	0.192	(0.070,0.315)
Alcohol consumption						
No						
Yes	0.104	(-0.001,0.209)	0.095	(0.064,0.126)	0.116	(0.010,0.223)
Exposure to mass media						
No	-	-	-	-	-	-
Yes	0.025	(-0.264,0.314)	0.232	(0.168,0.296)	0.084	(-0.211,0.379)
Symptom of STI						
No	-	-	-	-	-	-
Yes	0.278	(-0.250,0.805)	-0.136	(-0.330,0.058)	0.265	(-0.274,0.804)
Condom use in last sex						
No	-	-	-	-	-	-
Yes	0.086	(-0.072,0.243)	0.015	(-0.034,0.065)	0.084	(-0.077,0.245)
Marital status						
Never married	-	-	-	-	-	-
Ever married	0.076	(-0.260,0.411)	0.078	(-0.002,0.159)	0.082	(-0.259,0.423)
Correlation between HIV survey participation and HIV status ρ = -0.296 , 95% C.I. = (-.0450, -0.126)						
Wald test of independent equations (rho)=0 : chi2=11.22 , prob > chi2=0.0008						
Wald test of exclusion restriction on HIV survey participation $\chi 2(17)=46.86$ and prob> $\chi 2=0.0001$						

Table 4: Contact regression for men (15-54 years) for the non-participation due to failure of interviewer to contact eligible respondents, India, 2005-06

Background Variables	Heckman selection Model (Bivariate probit)				Imputation (Probit)	
	HIV survey participation		HIV status		HIV survey participation	
	Coefficient	95%C.I.	Coefficient	95%C.I.	Coefficient	95%C.I.
Age (in years)						
15-24	-	-	-	-	-	-
25-34	0.524	(0.388,0.661)	-0.034	(-0.060,-0.009)	0.402	(0.307,0.498)
35-44	0.573	(0.435,0.711)	-0.019	(-0.046,0.007)	0.407	(0.304,0.509)
45-54	0.319	(0.157,0.482)	0.001	(-0.027,0.030)	0.256	(0.136,0.376)
Education						
No education	-	-	-	-	-	-
Primary education	0.048	(-0.103,0.200)	0.125	(0.087,0.163)	0.050	(-0.055,0.154)
Secondary education	0.017	(-0.147,0.181)	0.123	(0.084,0.162)	0.034	(-0.073,0.142)
Higher education	-0.027	(-0.184,0.130)	0.173	(0.132,0.213)	-0.013	(-0.118,0.092)
Wealth status						
Poorest	-	-	-	-	-	-
Poor	-0.002	(-0.193,0.189)	0.066	(0.017,0.115)	-0.016	(-0.176,0.144)
Middle	-0.027	(-0.219,0.164)	0.139	(0.087,0.190)	-0.019	(-0.174,0.135)
Rich	0.033	(-0.165,0.231)	0.052	(-0.004,0.107)	0.029	(-0.125,0.182)
Richest	-0.120	(-0.339,0.099)	-0.126	(-0.189,-0.064)	-0.174	(-0.354,0.006)
Place of residence						
Urban	-	-	-	-	-	-
Rural	-0.115	(-0.225,-0.005)	0.019	(-0.032,0.069)	-0.105	(-0.203,-0.008)
Correlation between HIV survey participation and HIV status ρ = -0.164 , 95% C.I. = (-0.372, 0.058)						
Wald test of independent equations (rho)=0 : chi2=2.10 , prob > chi2=0.1477						
Wald test of exclusion restriction on HIV survey participation $\chi2(11)$=89.93 and prob> $\chi2$=0.000						

Table 5: Consent regression for women (15-49 years) for the non-participation due to refusal to give blood sample, India, 2005-06

Background Variables	Heckman selection Model (Bivariate probit)				Imputation (Probit)	
	HIV survey participation		HIV status		HIV survey participation	
	coefficient	95%C.I.	coefficient	95%C.I.	coefficient	95%C.I.
Age (in years)						
15-24	-	-	-	-	-	-
25-34	0.045	(-0.113,0.203)	0.057	(0.030,0.084)	-0.062	(-0.110, -0.013)
35-44	-0.132	(-0.314,0.049)	0.056	(0.025,0.086)	-0.083	(-0.135, -0.031)
45-49	-0.165	(-0.452,0.121)	-0.012	(-0.053,0.030)	-0.039	(-0.112,0.034)
Education						
No education	-	-	-	-	-	-
Primary education	0.031	(-0.148,0.211)	0.002	(-0.033,0.037)	-0.167	(-0.227,-0.107)
Secondary education	-0.053	(-0.243,0.137)	0.014	(-0.025,0.053)	-0.206	(-0.268,-0.143)
Higher education	-0.027	(-0.198,0.145)	0.049	(0.005,0.093)	-0.112	(-0.167,-0.057)
Wealth status						
Poorest	-	-	-	-	-	-
Poor	0.015	(-0.224,0.255)	0.105	(0.059,0.150)	-0.052	(-0.125, 0.021)
Middle	-0.082	(-0.322,0.158)	0.164	(0.112,0.215)	-0.121	(-0.194,-0.047)
Rich	0.041	(-0.196,0.277)	0.070	(0.011,0.128)	-0.075	(-0.152,0.002)

Richest	-0.162	(-0.462,0.138)	-0.102	(-0.170,-0.034)	0.122	(0.038,0.207)
Place of residence						
Urban	-	-	-	-	-	-
Rural	-0.063	(-0.223,0.098)	-0.184	(-0.245,-0.124)	-0.368	(-0.413,-0.322)
Number of sexual partner in last 12 months						
One	-	-	-	-	-	-
More than one	0.787	(0.556,1.018)	-0.070	(-0.146,0.006)	0.168	(0.032,0.303)
Exposure to mass media						
No	-	-	-	-	-	-
Yes	0.081	(-0.121,0.282)	0.220	(0.182,0.258)	-0.183	(-0.238,-0.128)
Symptom of STI						
No	-	-	-	-	-	-
Yes	-3.563	(-3.769,3.358)	-0.173	(-0.122,-0.076)	-0.365	(-0.571,-0.159)
Condom use in last sex						
No	-	-	-	-	-	-
Yes	0.086	(-0.150,0.322)	-0.077	(-0.122,-0.033)	0.063	(-0.004,0.131)
Marital status						
Never married	-	-	-	-	-	-
Ever married	3.668	(3.422,3.915)	0.162	(-0.078,0.403)	0.193	(-0.388,0.773)
Alcohol consumption						
No	-	-	-	-	-	-
Yes	0.291	(-0.013,0.596)	-0.185	(-0.252,-0.118)	-0.082	(-0.216,0.052)
Correlation between HIV survey participation and HIV status ρ = -0.146 , 95% C.I. = (-0.255, -0.033)						
Wald test of independent equation $\chi2(1)=6.46$ and prob> $\chi2=0.0110$						
Wald test of exclusion restrictions on survey participation and $\chi2=3040$ and prob> $\chi2=0.000$						

Table 6: Contact regression for women (15-49 years) for the non-participation due to failure of interviewer to contact eligible respondents, India, 2005-06

	Heckman selection model (Bivariate probit)				Imputation (Probit)	
Background Variables	**HIV survey participation**		**HIV status**		**HIV survey participation**	
	Coefficient	**95%C.I.**	**Coefficient**	**95%C.I.**	**Coefficient**	**95%C.I.**
Age (in years)						
15-24	-	-	-	-	-	-
25-34	0.256	(0.129,0.382)	0.056	(0.035,0.076)	-0.068	(-0.952,-.038)
35-44	0.154	(0.011,0.297)	0.070	(0.048,0.093)	-0.141	(-0.169,-0.107)
45-49	0.107	(-0.077,0.291)	0.034	(0.002,0.066)	-0.279	(-0.185,-0.103)
Education						
No education	-	-	-	-	-	-
Primary education	-0.040	(-0.182,0.102)	0.036	(0.005,0.067)	-0.341	(-0.377,-0.294)
Secondary education	-0.063	(-0.211,0.086)	0.085	(0.051,0.119)	-0.458	(-0.492,-0.407)
Higher education	-0.201	(-0.342,-0.059)	0.134	(0.092,0.175)	-0.504	(-0.534,-0.451)
Wealth status						
Poorest	-	-	-	-	-	-
Poor	-0.028	(-0.232,0.177)	0.169	(0.125,0.212)	-0.042	(-0.102,0.013)
Middle	0.017	(-0.176,0.210)	0.264	(0.214,0.314)	-0.062	(-0.126,-0.008)
Rich	0.095	(-0.094,0.284)	0.169	(0.114,0.225)	0.022	(-0.046,0.078)
Richest	-0.085	(-0.322,0.152)	-0.045	(-0.109,0.109)	0.335	(0.262,0.397)
Place of residence						
Urban	-	-	-	-	-	-
Rural	-0.086	(-0.217,0.044)	-0.173	(-0.235,-0.112)	-0.267	(-0.313,-0.219)

Correlation between HIV survey participation and HIV status ρ = -0.053 , 95% C.I. = (-0.179, 0.073)
Wald test of independent equation $\chi2(1)=0.68$ and prob> $\chi2=0.4102$
Wald test of exclusion restrictions on survey participation and $\chi2=40.0$ and prob> $\chi2=0.000$

The second parameter is rho (ρ), which measured the correlation between the unobserved error terms of HIV survey participation and HIV status. In Table 3 (consent regression for men), parameter rho (ρ) was negative and significant (ρ=-0.296, 95% C.I= -0.450, -0.126), similarly, in Table 5 (consent regression for women) parameter rho (ρ) was also negative and significant (ρ=-0.146, 95% C.I= -0.255, -0.033) for women.

A Wald test of independent equation validated the association between survey participation and HIV status, which was statistically significant for men ($\chi2$= 11.22, prob> $\chi2$=0.0008) as well as for women ($\chi2$= 6.46, prob> $\chi2$=0.011). Thus, the null hypothesis "no association between HIV survey participation and HIV status" was rejected for both men and women who 'did not consent to the blood test.' Therefore, HIV prevalence was higher among those men and women who were not participated in the survey compared to those who took part in the survey and consented to HIV test. Further, in Table 4 (contact regression for men) and in Table 6 (contact regression for women), the parameter rho (ρ) was negative but not significant for both men (ρ=-0.164, 95% C.I= -0.372, 0.058) and women (ρ=-0.053, 95% C.I= -0.179, 0.073). Wald test of the independent equation did not found any significant association with for men ($\chi2$= 2.10, prob> $\chi2$=0.148) and women ($\chi2$= 6.46, prob> $\chi2$=0.011) therefore the null hypothesis "no relationship between HIV survey participation and HIV status" is accepted for both men and women who were not contacted by the interviewer. This indicated that conventional imputation was good enough to account only for the selection of observed variables when estimating HIV prevalence for men and women.

Table 7: Estimated HIV prevalence for men (15-54 years) and women (15-49 years) derived from Imputation and Heckman selection model in India, 2005-06

Men HIV Prevalence				
	Imputation model (Probit regression)		**Heckman selection model (Bi variate probit regression)**	
	Prevalence	**95% C.I.**	**Prevalence**	**95% C.I.**
Complete case	0.35	(0.28,0.42)	0.35	(0.28,0.42)
Predicted via consent	0.54	(0.53,0.54)	1.82	(1.79,1.86)
Predicted via contact	0.39	(0.38,0.39)	0.95	(0.93,0.96)
National estimates	**0.48**	**(0.42,0.54)**	**0.77**	**(0.71,0.83)**
Women HIV Prevalence				
	Imputation model (Probit regression)		**Heckman selection model (Bi variate probit regression)**	
	Prevalence	**95% C.I.**	**Prevalence**	**95% C.I.**
Complete case	0.22	(0.17,0.26)	0.22	(0.17,0.26)
Those who refused to give blood test	0.43	(0.42,0.44)	0.57	(0.56,0.58)
Those who were not contacted	0.46	(0.45, 0.46)	0.35	(0.35,0.36)
National estimates	**0.35**	**(0.29,0.40)**	**0.42**	**(0.39,0.45)**

Table 7 shows the National estimates of HIV prevalence by imputation model and selection model. Result shows that in Imputation model, when non-participation is adjusted for men, then the national HIV prevalence (prevalence=0.48, 95% C.I. = (0.42, 0.54)) is quite higher than the estimate obtained from the complete case (prevalence=0.35, 95% C.I. = (0.28, 0.42)). Likewise, national estimate of HIV prevalence for women (prevalence=0.35, 95% C.I. = (0.29, 0.40)) is higher than the

estimate obtained from the complete case (prevalence=0.22, 95% C.I. = (0.17, 0.26)), where the unobserved variables are not considered in model.

Further, national estimates obtained from Heckman Selection Model is higher among men (prevalence=0.77, 95% C.I. = (0.71-0.83)) and women (prevalence=0.42, 95% C.I. = (0.39-0.45)) both as compare to estimates obtained from the conventional imputation method for men (prevalence=0.48, 95% C.I=0.42, 0.54)) and women (Prevalence=0.35, 95% C.I= (0.29, 0.40)).
Further, when the national HIV estimate has been adjusted for men and women, by using Heckman selection model, then the adjusted prevalence significantly higher for men (prevalence=0.77, 95% C.I. =0.71-0.83) and women (prevalence=0.42, 95% C.I. =0.39-0.45) both compared to unadjusted complete case for men (Prevalence=0.35, 95% C.I. = (0.28, 0.42)) and women (prevalence=0.22, 95% C.I. = (0.17, 0.26)). Notably in consent regression model, HIV prevalence among men (prevalence=1.82, 95% C.I. =1.79-1.86) was higher than HIV prevalence among the women (prevalence=0.57, 95% C.I. =0.5-0.58). Similarly, in contact regression model, HIV prevalence was higher for men (prevalence=0.95, 95% C.I. =0.93-0.96) than the women (prevalence=0.35, 95% C.I. =0.35-0.36).

Discussion

Heckman selection model is a very powerful approach to investigate the bias in sample selection for HIV test in the population-based survey (Hogan et al., 2012). Results of this study portray that the selection variable was significantly associated with HIV status of the men and women. Further, this study shows the statistically significant association between survey participation and HIV status for those who were interviewed but do not consent to HIV test. It clarifies that the sample selection was lead to substantial underestimation of national HIV prevalence in men and women (Barnighausen et al., 2011). Conversely, the study did not find any association between survey participation and HIV status, who were not contacted by the interviewer. To the best of our knowledge, this is the first study which corrected the national HIV prevalence of men and women using Heckman selection model in India. Findings of this study are consistent with the earlier studies (Barnighausen et al., 2011; Hogan et al., 2012; Montana, Mishra, & Hong, 2008).

Some of the studies revealed that the estimates of HIV prevalence, corrected by selection model have significantly differed from the conventional method used in DHS survey (Barnighausen et al., 2011; Hogan et al., 2012). In this study, the estimates of HIV prevalence has been examined through imputation model as well as selection model and both estimates have been compared. It is evident that estimates obtained from selection model are robust than the conventional method, which assumed that data is “missing at random” (Hogan et al., 2012). In general, Heckman selection model has been widely used in the field of social science and applied econometrics.

However, this is the first ever study using Indian data, which used selection model in the area of epidemiology to investigate the selection bias due to survey non-participation. The central feature of Heckman selection model is the selection of exclusion restriction variable. In this study Interviewer, ID has been taken as selection variable, which is plausible and also have the significant association with HIV survey participation. The choice of this selection variable is because extensive use of this in earlier studies (Barnighausen et al., 2011; Janssens et al., 2014; Reniers et al., 2009).

Further, there is no correlation between survey participation and HIV status for men and women when the interviewer did not contact them. In this case, conventional imputation method is good enough to determine the HIV status of respondents. In that case, HIV prevalence among the non-participated group was less compared to participated group conditioning on observed characteristic (Barnighausen et al., 2011)**.**For example, if a person knows his/her behavioral status in the past, he/she is less likely to participate in the survey.

Conclusions

Findings of this study conclude that the national HIV prevalence for men and women confronted underestimation by the conventional method reported in DHS. Therefore, emphasis should be given to increase the participation of respondents in the survey to establish the national prevalence. Participation rate can be improved by contacting more and more respondents by revisiting to households and eliciting more and more consent for the blood test. Also, the consent rate may be increased by providing some kind of incentives like financial assistance to the respondents (Gouws et

al., 2008Gouws et al., 2008; Marsh et al., 2014). A valid and efficient way to providing the estimate of HIV prevalence is to incorporate Heckman selection model instead of the conventional method to provide an estimate of national prevalence in the large-scale demographic survey.

Strengths and limitations

The result draws its strength by use of Heckman selection model; as this is the first ever study in India which used the selection model to the prevalence of HIV among men and women separately. Heckman selection model is a powerful analytical approach to investigate the selection bias; however, this model has few drawbacks. First, this model necessitates, at least, one exclusion restriction variable. This fundamental requirement created the limitation for this approach. For example, in the demographic health survey, interviewer's ID is available but other than this survey it is not necessary to find the record of such a plausible exclusion restriction. Therefore, Heckman selection model can't work there. Second, in the Demographic health survey, during the field, along with interviewer, some health workers and professional were also conducted the interview and collected blood samples. But, the data could not control the identities of those individuals (health professionals, health workers, etc.) other than appointed interviewers. Therefore, to enhance the applicability of selection models, DHS should record the identities of every responsible individual, who were conducting HIV testing.

Ethical Approval

NFHS-3 obtained informed consent from the individual respondents for the interview, as well as for blood sampling. The data collection procedures were approved by the Institutional Review Board of CDC, Atlanta as well as the IRB of the International Institute for Population Sciences, Mumbai. This paper is based on the secondary data set with no identifiable information on the survey participants and hence no question of human subject violation.

Data Availability

Thanks are due to Data Centre, International Institute for Population Sciences for providing the data for this study.

Acknowledgements

Authors are thankful to the anonymous reviewers of this paper for their valuable comments, which have improved the quality of the paper.

References

Barnighausen, T.; Bor, J., Wandira-Kazibwe, S.; & Canning, D. (2011) 'Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models', *Epidemiology, 22*(1), pp. 27-35, doi: 10.1097/EDE.0b013e3181ffa201

Bignami-Van Assche, S, Salomon, JA, & Murray, CJL. (2005) 'Evidence from national population-based estimates of bias in HIV prevalenc',. Paper presented at the *Population Association of America Annual Meeting*, Philadelphia.

Boerma, J. T., Ghys, P. D., & Walker, N. (2003) 'Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard', *The Lancet, 362*(9399), pp. 1929-1931, doi: 10.1016/s0140-6736(03)14967-7

Brick , JM , & Kalton, G. (1996) 'Handling missing data in survey research', *Stat Methods Med Res, 5*, pp.215–238.

Chinomona, Amos, & Mwambi, Henry. (2015) 'Multiple imputation for non-response when estimating HIV prevalence using survey data', *BMC Public Health, 15*, pp. 1059, doi: 10.1186/s12889-015-2390-1

Gouws, E., Mishra, V. & Fowler, T. B. (2008) '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, 84(1)*, pp. i17-i23, doi: 10.1136/sti.2008.030452

Gregson, S., Terceira, N., Kakowa, M., Mason, P. R., Anderson, R. M., Chandiwana, S. K., & Carael, M. (2002) 'Study of bias in antenatal clinic HIV-1 surveillance data in a high contraceptive prevalence population in sub-Saharan Africa', *AIDS, 16*(4), pp. 643-652.

Hogan, D. R., Salomon, J. A., Canning, D., Hammitt, J. K., Zaslavsky, A. M., & Barnighausen, T. (2012), National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models', *Sex Transm Infect, 88(2)*, pp. 17-23, doi: 10.1136/sextrans-2012-050636

IIPS, & Macro., ORC. (2007) 'National Family Halth Survey 3 (NFHS3) 2005-06'. *Mumbai: International Institute for Population Sciences*.

Janssens, W., van der Gaag, J., Rinke de Wit, T. F., & Tanovic, Z. (2014) 'Refusal bias in the estimation of HIV prevalence', *Demography, 51*(3), pp. 1131-1157, doi: 10.1007/s13524-014-0290-0

Kalichman, S. C., & Simbayi, L. C. (2003) 'HIV testing attitudes, AIDS stigma, and voluntary HIV counselling and testing in a black township in Cape Town, South Africa', *Sex Transm Infect, 79*(6), pp. 442-447.

Kalton, JM Brick & G. (1996) 'Handling Missing Data in Survey Research', (5), pp. 215-238.

Kranzer, K., McGrath, N., Saul, J., Crampin, A. C., Jahn, A., Malema, S., . . . Glynn, J. R. (2008) 'Individual, household and community factors associated with HIV test refusal in rural Malawi', *Trop Med Int Health, 13*(11), pp. 1341-1350. doi: 10.1111/j.1365-3156.2008.02148.x

Lachaud, J. P. (2007) 'HIV prevalence and poverty in Africa: micro- and macro-econometric evidences applied to Burkina Faso', *J Health Econ, 26*(3), pp. 483-504. doi: 10.1016/j.jhealeco.2006.10.007.

Marsh, Kimberly, Mahy, Mary, Salomon, Joshua A., & Hogan, Daniel R. (2014) 'Assessing and adjusting for differences between HIV prevalence estimates derived from national population-based surveys and antenatal care surveillance, with applications for Spectrum 2013', *AIDS (London, England), 28*(4), pp. S497-S505. doi: 10.1097/QAD.0000000000000453

Montana, L. S., Mishra, V., & Hong, R. (2008) 'Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa', *Sex Transm Infect, 84 Suppl 1*, pp. i78-i84. doi: 10.1136/sti.2008.030106

Reniers, G., Araya, T., Berhane, Y., Davey, G., & Sanders, E. J. (2009) 'Implications of the HIV testing protocol for refusal bias in seroprevalence surveys', *BMC Public Health, 9*, pp. 163, doi: 10.1186/1471-2458-9-163

Roderick J. A. Little, Donald B. Rubin (1987) 'Statistical analysis with missing data ',16, pp. 150-155, New York, USA: Wiley series in probability and statistics;: American Educational Research Association and American Statistical Association.

Rubin, Donald B. (1987) 'Multiple Imputation for Nonresponse in Surveys', New York, USA: John Wiley and Sons, Ltd; 1987..

UNAIDS. (2015) 'AIDSinfo', *C. Factsheets (Ed.)*.

Weiser, S. D., Heisler, M., Leiter, K., Percy-de Korte, F., Tlou, S., DeMonner, S., . . . Iacopino, V. (2006) 'Routine HIV testing in Botswana: a population-based study on attitudes, practices, and human rights concerns', *PLoS Med, 3*(7), e261. doi: 10.1371/journal.pmed.0030261

WHO. (2015) 'Global Health Observatory (GHO) data', *HIV/AIDS*.