

Supplementary Material

Descriptive Statistics for HIV dataset

Mean, standard deviation, minimum and maximum values for the numeric variables separated for complete and incomplete cases are presented in Supplementary Table S1. The 95% confidence intervals (CI) for the difference of means of complete versus incomplete are also presented. Although most variables presented significantly different means, with CI not including zero, for complete and incomplete cases, most of the differences are relatively small. For example, the mean village population for villages incomplete is only 40 more than for the villages with complete information (2,820 versus 2,860). Considering the mean values, population in villages with incomplete cases are younger (mean 15.1+14 versus 17.3+14 years old), have higher income (38.8 versus 24.7 dollars/month), live closer to a health clinic (24.8 versus 29.4 minutes), present lower HIV prevalence (4% versus 6%) and more years of education (8.1 versus 6.6 years). Regarding stigma variables, almost all of them are higher in average for the incomplete cases, except disclosure concern and public attitudes.

The density distributions for log of income (better to visualize than the non-log measure), stigma from family and health care workers and years of education are shown in Supplementary Figure S1. These four were the numeric variables that presented the most clear distinction in their density distribution plots when comparing complete and incomplete. The plots, along with the descriptive statistics from the table suggest that the incomplete portion of the sample represents a more affluent (in average) subset of villages with more years of education, higher income and that this population anticipated a higher stigma from health care workers and family due to HIV status.

Categorical variables are presented in Supplementary Table S2, along with frequencies and percentages by completeness of cases. P-values for the chi-square test for difference in frequency between complete and incomplete were all significant, smaller than 0.001, and are not shown in the table. Some of the differences in frequency between the incomplete and complete portion of the sample are highlighted here. Incomplete has a smaller percentage of known HIV+ cases (1.6 versus 4.5) and higher new HIV+ (2.1 versus 1.5). Incomplete also has higher percentage of males (45% versus 36%), higher percentage of the population in the two highest wealth quintiles (50% versus 20% combined) and higher proportion of people away for work for one month or more (14.6% versus 11.6%). Also, incompletes have a higher percentage of men without a partner (17.7% versus 10.1%), and lower percentage working as farmer (24.9% versus 43.9%) and much higher in the salaried trade (24.4% versus 8.2%).

The distributions of the incomplete variables, by HIV status are shown in Supplementary Figure S2. For depressive symptoms score, the distribution is similar for negatives and new positives, and for known positives, slightly higher values and more spread. For alcohol consumption risk level, new positives have wider spread and higher median values than known positives, which present wider spread and higher values than negatives. For the two categorical variables, the figures are also separated by complete and incomplete cases; incomplete cases being the ones that have information for the categorical variable but not for alcohol and depressive symptoms variables. There is a slight difference in proportions of pregnancy status for complete and incomplete cases across the three classes of HIV status. More difference is noticed in the proportion of pregnancy in complete versus incomplete cases for new positives. For occupation variable, there is a difference in distribution in each HIV class in occupation levels when comparing complete and incomplete cases. For new positives, there is a larger proportion in the fisherman and peasant farmer categories in complete cases. There is a higher proportion of salaried tradesperson in incomplete cases, while the proportion of people not employed outside of the house seems to be similar across complete and incomplete cases for all HIV classes.

Table S1. Baseline mean, standard deviation, minimum, maximum, and confidence interval for means difference for numeric variables, by completeness of cases.

	Complete	Incomplete	Total	Mean difference
	n = 9,099	n = 14,838	n = 23,937	
Variable	mean (sd) min; max	mean (sd) min; max	mean (sd) min; max	95% CI mean difference
Village population (x 1000)	2.86 (0.92) 0.55; 3.97	2.82 (1.15) 0.44; 4.42	2.84 (1.07) 0.44; 4.42	(0.01; 0.07)
Age centered at 14 (real age is value plus 14)	17.34 (11.11) 0; 45	15.05 (10.07) 0; 45	15.92 (10.5) 0; 45	(2.02; 2.56)
Monthly income in dollars	24.68 (41.71) 0; 833	38.77 (125.14) 0; 11111	33.42 (102) 0, 11111	(-16.75; -11.43)
Time to health clinic (times 15 minutes)	1.96 (1.99) 0.07; 28	1.65 (1.45) 0; 30	1.77 (1.68) 0; 30.33	(0.28; 0.36)
Anticipated stigma disclosure concern	2.57 (1.06) 0; 4	2.57 (1.05) 0; 4	2.57 (1.05) 0; 4	(-0.03; 0.03)
Anticipated HIV stigma from family	1.08 (0.89) 0; 4	1.26 (0.99) 0; 4	1.19 (0.96) 0; 4	(-0.22; -0.17)
Anticipated HIV stigma from health care workers	0.81 (0.64) 0; 4	0.91 (0.74) 0; 4	0.87 (0.70) 0; 4	(-0.12; -0.08)
Depressive symptoms	5.83 (5.65) 0; 30	-	5.83 (5.65) 0; 30	-
Alcohol consumption risk level score	1.09 (2.2) 0; 12	-	1.09 (2.2) 0; 12	-
Village HIV prevalence (%)	0.06 (0.02) 0.03; 0.1	0.04 (0.01) 0.02; 0.05	0.05 (0.02) 0.02; 0.1	(0.02; 0.02)
Enacted stigma village level	0.98 (0.12) 0.8; 1.4	1.05 (0.16) 0.8; 1.4	1.03 (0.15) 0.8; 1.43	(-0.07; -0.07)
Anticipated stigma due to HIV disclosure concern village level	2.57 (0.26) 2.2; 3.0	2.57 (0.49) 1.9; 3.3	2.57 (0.42) 1.9; 3.3	(-0.01; 0.01)
Anticipated stigma public attitudes village level	1.65 (0.27) 1.4; 2.3	1.62 (0.21) 1.4; 2.0	1.64 (0.23) 1.4; 2.3	(0.02; 0.04)
Anticipated HIV stigma from family village level	1.07 (0.13) 0.9; 1.5	1.26 (0.12) 1.1; 1.7	1.19 (0.15) 0.9; 1.7	(-0.19; -0.19)
Anticipated HIV stigma from health care workers village level	0.81 (0.09) 0.6; 0.9	0.92 (0.16) 0.7; 1.4	0.88 (0.15) 0.6; 1.4	(-0.11; -0.11)
Education (number of years)	6.55 (3.49) 0; 15	8.08 (3.35) 0; 15	7.5 (3.48) 0; 15	(-1.61; -1.43)

Notes. sd = standard deviation, min = minimum, max = maximum.

CI = Approximate confidence interval for mean difference using normal approximation and pooled standard deviation S_p .

$$CI: \bar{x}_1 - \bar{x}_2 \pm 1.96 * S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right), S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Table S2. Baseline frequency distribution for categorical variables by completeness of cases.

Variable	Categories	Complete (n = 9,099) n (row %)	Incomplete (n = 14,838) n (row %)	Total (n = 23,937) n (row %)
HIV status	Negative	8546 (93.9)	14284 (96.3)	22830 (95.4)
	New Positive	140 (1.5)	314 (2.1)	454 (1.9)
	Known Positive	413 (4.5)	240 (1.6)	653 (2.7)
Gender	Male	3342 (36.7)	6707 (45.2)	10049 (42.0)
	Female	5757 (63.3)	8131 (54.8)	13888 (58.0)
Months since study started	1	3599 (39.6)	14838 (100)	18437 (77.0)
	2	5500 (60.4)	-	5500 (23.0)
Religion	Muslim	2184 (24.0)	3870 (26.1)	6054 (25.3)

	Christian & non-Muslim	6915 (76.0)	10968 (73.9)	17883 (74.7)
Wealth index	Lowest quintile	3356 (36.9)	1517 (10.2)	4873 (20.4)
	2nd lowest quintile	2311 (25.4)	2971 (20.0)	5282 (22.1)
	3rd lowest quintile	1590 (17.5)	2867 (19.3)	4457 (18.6)
	4 th lowest quintile	984 (10.8)	2850 (19.2)	3834 (16.0)
	Highest quintile	858 (9.4)	4633 (31.2)	5491 (22.9)
Away for work	Not away	8039 (88.4)	12677 (85.4)	20716 (86.5)
1 month or more	One or more times	1060 (11.6)	2161 (14.6)	3221 (13.5)
Transportation	Free: walking/bike	4882 (53.7)	7370 (49.7)	12252 (51.2)
	Low cost: taxi	958 (10.5)	2870 (19.3)	3828 (16.0)
	High: boda/car	3259 (35.8)	4598 (31.0)	7857 (32.8)
Marital status	Never married	1618 (17.8)	4498 (30.3)	6116 (25.6)
	Married	1556 (17.1)	2204 (14.9)	3760 (15.7)
	Widowed or divorced	5925 (65.1)	8136 (54.8)	14061 (58.7)
Another household member is HIV +	No	7260 (79.8)	12067 (81.3)	19327 (80.7)
	Yes	390 (4.3)	442 (3.0)	832 (3.5)
	Do not know	1449 (15.9)	2329 (15.7)	3778 (15.8)
		n = 9,099	n = 14,762	n = 23,861(1)
Pregnancy of self or partner	No	7272 (79.9)	10918 (74.0)	18190 (76.2)
	Yes	906 (10.0)	1236 (8.4)	2142 (9.0)
	No partner	921 (10.1)	2608 (17.7)	3529 (14.8)
		n = 9,099	n = 12,226	n = 21,325(2)
Occupation	Peasant farmer	3998 (43.9)	3046 (24.9)	7044 (33.0)
	Casual worker	839 (9.2)	1182 (9.7)	2021 (9.5)
	Salaried trade	750 (8.2)	2988 (24.4)	3738 (17.5)
	Fish/Csw/Rest/Bar/Attendant	430 (4.7)	83 (0.7)	513 (2.4)
	Business selling	1287 (14.1)	1799 (14.7)	3086 (14.5)
	Not employed out home	1795 (19.7)	3128 (25.6)	4923 (23.1)

Note. n = sample size. (1) 76 missing for pregnancy. (2) 2,612 missing for occupation.

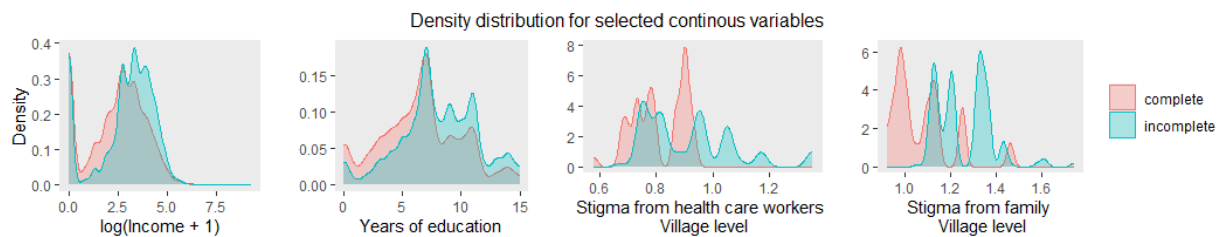


Figure S1. Density distribution.

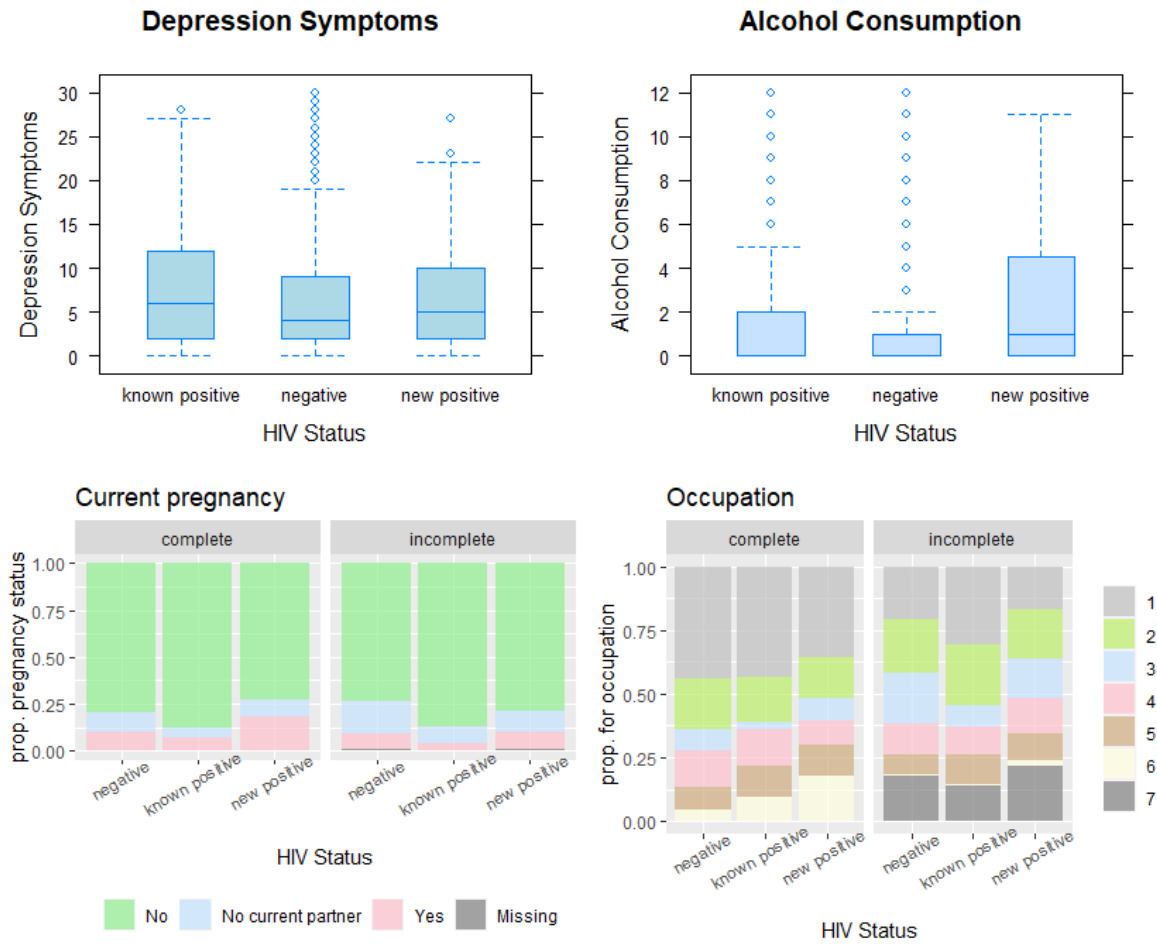


Figure S2. Boxplots for depression score and alcohol consumption (complete data). Bar plots with frequency distribution for pregnancy and occupation. Sample sizes for all plots in Table S2. Legend for occupation: 1 - 'Peasant farmer', 2 - 'Not employed outside home', 3 - 'Salaried tradesperson', 4 - 'Business selling', 5 - 'Casual worker', 6 - 'Fish/commercial sex worker/restaurant/bar/attendant', 7 'Missing'.

Analysis of Variance for trimmed means for overall prediction accuracy and sensitivity.

Table S3. Repeated measures ANOVA for trimmed means for overall prediction accuracy and sensitivity.

Overall Prediction Accuracy, sample size = 1200

Test statistic: $F = 1196.351$

Degrees of freedom 1: 2.45

Degrees of freedom 2: 734.71

p-value: 0

Overall Prediction Accuracy, sample size = 350

Test statistic: $F = 1103.48$

Degrees of freedom 1: 2.16

Degrees of freedom 2: 647.15

p-value: 0

Sensitivity New HIV Positive, sample size = 1200

Test statistic: $F = 1888.55$

Degrees of freedom 1: 1.9
 Degrees of freedom 2: 569.4
 p-value: 0

Sensitivity New HIV Positive, sample size = 350

Test statistic: F = 281.3492
 Degrees of freedom 1: 2.08
 Degrees of freedom 2: 622.27
 p-value: 0

Sensitivity Negative HIV, sample size = 1200

Test statistic: F = 1101.361
 Degrees of freedom 1: 2.47
 Degrees of freedom 2: 741.14
 p-value: 0

Sensitivity Negative HIV, sample size = 350

Test statistic: F = 1055.991
 Degrees of freedom 1: 2.17
 Degrees of freedom 2: 648.46
 p-value: 0

Sensitivity Known HIV Positive, sample size = 1200

Test statistic: F = 430.2524
 Degrees of freedom 1: 1.98
 Degrees of freedom 2: 593.47
 p-value: 0

Sensitivity Known HIV Positive, sample size = 350

Test statistic: F = 230.1897
 Degrees of freedom 1: 2.21
 Degrees of freedom 2: 660.04
 p-value: 0

Pairwise comparison between imputations methods for mean prediction accuracy and sensitivity with the full dataset.

Table S4. Mean difference, confidence interval (CI) and p-value for pairwise comparison between methods.

Overall Prediction Accuracy, sample size = 1200

method	method	mean diff.	CI	p-value
CCA	amelia	-0.095	[-0.101; -0.088]	0
CCA	mice	-0.116	[-0.122; -0.111]	0
CCA	missForest	-0.103	[-0.11; -0.096]	0
CCA	hmisc	-0.117	[-0.123; -0.112]	0
amelia	mice	-0.023	[-0.029; -0.017]	0
amelia	missForest	-0.01	[-0.013; -0.007]	0
amelia	hmisc	-0.023	[-0.029; -0.018]	0
mice	missForest	0.014	[0.008; 0.02]	0
mice	hmisc	-0.001	[-0.003; 0.001]	0.084
missForest	hmisc	-0.015	[-0.021; -0.009]	0

Overall Prediction Accuracy, sample size = 350

method	method	mean diff.	CI	p-value
CCA	amelia	-0.194	[-0.207; -0.181]	0
CCA	mice	-0.176	[-0.185; -0.166]	0
CCA	missForest	-0.201	[-0.215; -0.188]	0
CCA	hmisc	-0.176	[-0.186; -0.166]	0
amelia	mice	0.02	[0.009; 0.031]	0
amelia	missForest	-0.007	[-0.012; -0.003]	0
amelia	hmisc	0.019	[0.008; 0.03]	0
mice	missForest	-0.028	[-0.039; -0.017]	0
mice	hmisc	0	[-0.003; 0.003]	0.907
missForest	hmisc	0.027	[0.017; 0.038]	0

Sensitivity New HIV Positive, sample size = 1200

method	method	mean diff.	CI	p-value
CCA	amelia	-0.457	[-0.482; -0.432]	0
CCA	mice	-0.538	[-0.556; -0.52]	0
CCA	missForest	-0.458	[-0.484; -0.432]	0
CCA	hmisc	-0.547	[-0.565; -0.529]	0
amelia	mice	-0.066	[-0.087; -0.045]	0
amelia	missForest	0.003	[-0.005; 0.012]	0.239
amelia	hmisc	-0.076	[-0.096; -0.056]	0
mice	missForest	0.071	[0.048; 0.094]	0
mice	hmisc	-0.008	[-0.013; -0.003]	0
missForest	hmisc	-0.081	[-0.104; -0.059]	0

Sensitivity New HIV Positive, sample size = 350

method	method	mean diff.	CI	p-value
CCA	amelia	-0.207	[-0.238; -0.176]	0
CCA	mice	-0.235	[-0.26; -0.209]	0
CCA	missForest	-0.194	[-0.227; -0.161]	0
CCA	hmisc	-0.242	[-0.267; -0.216]	0
amelia	mice	-0.022	[-0.042; -0.001]	0.004
amelia	missForest	0.006	[-0.005; 0.017]	0.125
amelia	hmisc	-0.029	[-0.049; -0.01]	0
mice	missForest	0.032	[0.008; 0.057]	0
mice	hmisc	-0.007	[-0.014; -0.001]	0.002
missForest	hmisc	-0.042	[-0.065; -0.019]	0

Sensitivity Negative HIV, sample size = 1200

method	method	mean diff.	CI	p-value
CCA	amelia	-0.092	[-0.099; -0.085]	0
CCA	mice	-0.116	[-0.121; -0.11]	0
CCA	missForest	-0.101	[-0.108; -0.094]	0
CCA	hmisc	-0.117	[-0.122; -0.111]	0
amelia	mice	-0.025	[-0.031; -0.019]	0
amelia	missForest	-0.011	[-0.014; -0.007]	0
amelia	hmisc	-0.025	[-0.031; -0.019]	0
mice	missForest	0.016	[0.01; 0.022]	0
mice	hmisc	-0.001	[-0.003; 0.001]	0.181

missForest	hmisc	-0.016	[-0.023; -0.01]	0
------------	-------	--------	-----------------	---

Sensitivity Negative HIV, sample size = 350

method	method	mean diff.	CI	p-value
CCA	amelia	-0.201	[-0.214; -0.187]	0
CCA	mice	-0.183	[-0.193; -0.172]	0
CCA	missForest	-0.208	[-0.222; -0.194]	0
CCA	hmisc	-0.183	[-0.194; -0.173]	0
amelia	mice	0.02	[0.008; 0.031]	0
amelia	missForest	-0.008	[-0.012; -0.003]	0
amelia	hmisc	0.019	[0.008; 0.031]	0
mice	missForest	-0.028	[-0.04; -0.017]	0
mice	hmisc	0	[-0.003; 0.003]	0.901
missForest	hmisc	0.028	[0.016; 0.039]	0

Sensitivity Known HIV Positive, sample size = 1200

method	method	mean diff.	CI	p-value
CCA	amelia	0.138	[0.111; 0.166]	0
CCA	mice	0.26	[0.242; 0.278]	0
CCA	missForest	0.138	[0.111; 0.164]	0
CCA	hmisc	0.259	[0.24; 0.278]	0
amelia	mice	0.123	[0.104; 0.142]	0
amelia	missForest	-0.002	[-0.012; 0.007]	0.485
amelia	hmisc	0.121	[0.102; 0.141]	0
mice	missForest	-0.128	[-0.15; -0.106]	0
mice	hmisc	-0.002	[-0.009; 0.005]	0.391
missForest	hmisc	0.125	[0.103; 0.147]	0

Sensitivity Known HIV Positive, sample size = 350

method	method	mean diff.	CI	p-value
CCA	amelia	0.136	[0.108; 0.163]	0
CCA	mice	0.218	[0.199; 0.237]	0
CCA	missForest	0.13	[0.105; 0.156]	0
CCA	hmisc	0.212	[0.193; 0.231]	0
amelia	mice	0.087	[0.066; 0.108]	0
amelia	missForest	-0.002	[-0.011; 0.008]	0.616
amelia	hmisc	0.08	[0.058; 0.102]	0
mice	missForest	-0.09	[-0.114; -0.066]	0
mice	hmisc	-0.007	[-0.013; -0.002]	0
missForest	hmisc	0.083	[0.06; 0.105]	0