Handling Missingness in Prevalence Estimates from National Surveys

Oyelola Adegboye, PhD
Menzies School of Health Research, Charles Darwin University, Darwin

Denis Heng-Yan Leung, Tomoki Fujii, Li Siyu
School of Economics, Singapore Management University

This talk

Adegboye, O.; Fiji, T.; Leung., D. Refusal bias in HIV data from the Demographic and Health Surveys: Evaluation, critique and recommendations. *Statistical Methods in Medical Research* **2019**, 1-16.
Adegboye, O.A.; Fujii, T.; Leung, D.H.; Siyu, L. HIV Estimation Using Population-Based Surveys With Non-Response: A Partial Identification Approach. *Statistics in Medicine* **2024**.

When survey participants say ‘Nope’

- Occurs when individuals refuse HIV testing.
- Refusal introduces potential bias.
  - may be related to prior knowledge of HIV status.
- Leads to underestimation or misestimation of HIV prevalence.

Prior approaches

- Garcia-Calleja et al.: Scenario study; did not separate non-contacts/refusals
- Marston et al.: Non-informative non-response; multiple imputation
- Mishra et al.: Logistic regression under non-informative assumption
- Hogan et al.: Selection model requiring a valid instrumental variable
- Reniers & Eaton: Adjusted refusal bias using longitudinal data

Surveys

- Demographic and Health Surveys (DHS): Powering Health Data in the Global South…Malawi DHS
- Others: Antenatal Clinic (ANC)…Malawi ANC
- Malawi Diffusion and Ideational Change Project (MDICP)

Strategies

- Complete-case: Ignores non-responders, biased if missingness is related to HIV status.
- Mean Score Imputation: Assumes MAR; severely underestimates HIV in the presence of refusal bias.
- Auxiliary Data (sentinel surveillance surveys (ANC)

Key definitions and notations

Let $\boldsymbol{D_i}$ be HIV status indicator (1 if positive) and zero otherwise.

$\boldsymbol{\pi \equiv E[D_i]}$, Population HIV prevalence

$\boldsymbol{E[D_i|Z_i]}$, HIV prevalence of certain sub-populations

$\boldsymbol{R_i}$ = Refusal indicator (1 if refuses test, 0 accepts test)

Complete case ($\widehat{\pi_{CC}}$)

Let $\mathbf{I}(\cdot)$ be an indicator function (which takes one if $\cdot$ is true and zero otherwise) and $\boldsymbol{N}$ is the total number of individuals in the MDHS sample.

$\boldsymbol{N_{R_i=0}} = \sum_i \boldsymbol{I}(\boldsymbol{R_i} = \mathbf{0}) = \boldsymbol{N} - \sum_i \boldsymbol{R_i}$ represents the total number of individuals who accept an HIV test.

$$\widehat{\pi_{CC}} = \frac{\sum_{\{\mathrm{i}|\mathrm{R_i}=0\}} D_i \}}{\mathrm{N_{R_i=0}}} \quad (1)$$

$$\widehat{\pi_{CC}} \equiv \boldsymbol{E}[\boldsymbol{D_i}|\mathbf{R_i} = \mathbf{0}], \text{ but not } \boldsymbol{E}[\boldsymbol{D_i}]$$

Mean score imputation ($\hat{\pi}_{\mathrm{MSI}}$)

In the current context, this method requires:

$$\mathrm{P}(D_i = 1|X_i, R_i = 0) = \mathrm{P}(D_i = 1|X_i, R_i = 1) = \mathrm{P}(D_i = 1|X_i) \qquad (2)$$

If an unbiased estimator $\widehat{D_i}$ of $\mathrm{P}(D_i = 1|X_i)$ can be obtained from those with observed HIV status, then we can estimate prevalence by a method equivalent to the mean score imputation (MSI) method, e.g., Pepe et al.,1994 in the missing data literature.

$$\hat{\pi}_{\mathrm{MSI}} = \frac{\sum_i \hat{D}_i}{N} = \frac{\sum_{i|R_i=0} \hat{D}_i + \sum_{i|R_i=1} \hat{D}_i}{N_{R_i=0} + N_{R_i=1}} \qquad (3)$$

Data Required: MDHS only

Inverse Probability Weighting ($\hat{\pi}_{IF}$ & $\hat{\pi}_1$)

Infeasible: $$\hat{\pi}_{IF} = \frac{\sum_{i=1}^{N} \frac{(1-R_i)D_i}{P(R_i=0)}}{\sum_{i=1}^{N} \frac{(1-R_i)}{P(R_i=0)}} \qquad (4)$$

If we replace $P(R_i = 0)$ by an estimator $\hat{P}(R_i = 0) \equiv \hat{P}(R_i = 0|X_i)$

Estimated: $$\hat{\pi}_1 = \frac{\sum_{i=1}^{N} \frac{(1-R_i)D_i}{\hat{P}(R_i = 0|X_i)}}{\sum_{i=1}^{N} \frac{(1-R_i)}{\hat{P}(R_i = 0|X_i)}}, \qquad (5)$$

Horvitz-Thompson (1952) estimator with estimated propensity scores

Data Required: MDHS only

Refusal due to prior knowledge ($\widehat{\pi}_{RE}$)

$$P(R_i = 1|D_i = 1, T_i = 0) = P(R_i = 1|D_i = 0, T_i = 0) = P(R_i = 1|T_i = 0) \quad (6)$$

$$P(D_i = 1|T_i = 1) = P(D_i = 1) \quad (7)$$

where $T_i = 0$ means that a subject does not know his/her HIV status and $T_i = 1$ o.w.

Under these assumptions, it can be shown that: $0 = [\{P(R_i = 0|T_i = 0)P(T_i = 0) + P(T_i = 1)\}(\Delta - 1)]P(D_i = 1)^2 + [-P(D_i = 1|R_i = 0)P(R_i = 0)(\Delta - 1) + P(R_i = 0|T_i = 0)P(T_i = 0) + \{1 - \Delta P(R_i = 1|T\_i = 1)\}P(T_i = 0)]P(D_i = 1) -$

$$P(D_i = 1|R_i = 0)P(R_i = 0) \quad (8)$$

where the RR of refusal Δ is defined as follows: $\Delta \equiv \frac{P(R_i = 1|D_i = 1, T_i = 1)}{P(R_i = 1|D_i = 0, T_i = 1)}$

Estimator $\widehat{\pi}_{RE}$ of $E[D_i]$is the unique root of the quadratic equation on the unit interval

Never Tested estimator ($\hat{\pi}_2$)

Notice that eqs. 6 & 7 imply:
$P(D_i = 1) = P(D_i = 1|T_i = 0) = P(D_i = 1|T_i = 0, R_i = 0).$

This suggests we can estimate the prevalence of HIV by:

$$\hat{\pi}_2 = \sum_{i=1}^{N}(1 - R_i)D_i(1 - T_i)(1 - R_i)(1 - T_i) == \frac{\sum_{\mathrm{i|T_i=0,R_i=0}} \mathrm{D_i}}{N_{\mathrm{T_i=0,R_i=0}}} \qquad (9)$$

Data Required: MDHS only

Bound estimators ($\hat{\pi}_{3\pm}$)

$$P_- = P(D_i = 1, R_i = 0) + WP\left(D_i = 1 \middle| \tilde{T}_i = 1, R_i = 1, M_i = 1\right)P\left(\tilde{T}_i = 1, R_i = 1\right) + P\left(D_i = 1 \middle| \tilde{T}_i = 0, R_i = 0\right)P\left(\tilde{T}_i = 0, R_i = 1\right)$$

$$= P(D_i = 1, R_i = 0) + WP\left(D_i = 1 \middle| \tilde{T}_i = 1, R_i = 1, M_i = 1\right)P\left(\tilde{T}_i = 1, R_i = 1\right) + W'P\left(D_i = 1 \middle| \tilde{T}_i = 0, R_i = 0, M_i = 1\right)P\left(\tilde{T}_i = 0, R_i = 1\right), \qquad (10)$$

$$P_+ = P(D_i = 1, R_i = 0) + WP\left(D_i = 1 \middle| \tilde{T}_i = 1, R_i = 1, M_i = 1\right)P\left(\tilde{T}_i = 1, R_i = 1\right) + P\left(D_i = 1 \middle| \tilde{T}_i = 1, R_i = 1\right)P\left(\tilde{T}_i = 0, R_i = 1\right).$$

$$= P(D_i = 1, R_i = 0) + WP\left(D_i = 1 \middle| \tilde{T}_i = 1, R_i = 1, M_i = 1\right)P(R_i = 1). \qquad (11)$$

Where Z, W = Population adjustment factors, T̃_i = Prior test (result may be unknown), M_i = MDICP population indicator (rural vs. urban)

$\hat{\pi}_{3-}$and $\hat{\pi}_{3+}$ $a$re estimates of P_-andP_+

ANC based ($\hat{\pi}_4$)

Lets C_i to be the index of the district-area in which the ***i***-th individual resides. Then an estimate of the population HIV prevalence is:

$$\widehat{\pi}_4 = \sum_c \widehat{\pi}^c_{ANC} \left(\frac{N_{C_i = c}}{\sum_{c'} N_{C_i = c'}} \right) \quad (12)$$

where $\widehat{\pi}^c_{ANC}$ is the prevalence estimate in district-area ***c*** using the ANC data.

If we let $\widetilde{M_i} = 1$ be an indicator for an individual who has been tested at an ANC site, then $\widehat{\pi}_4$ makes the following assumption:

$$P\left(D_i = 1 \middle| \widetilde{M_i} = 1, C_i = c\right) = P\left(D_i = 1 \middle| \widetilde{M_i} = 0, C_i = c\right) = P(D_i = 1 | C_i = c)$$

(13)

ANC-adjusted

We assume

$$P(R_i = 0) = g(D_i, X_i) \equiv P(R_i = 0|D_i, X_i) \quad (14)$$

for some known function ***g*** that depends on the HIV status D_i and some observable covariates X_i.

Because D_i is unknown for those who refuse an HIV test. Therefore, we make the following assumption:

$$P(R_i = 0|X_i = x, D_i, [\widehat{\pi}\}^c_{ANC}, C_i = c)$$

$$= P(R_i = 0|X_i = x, \widehat{\pi}^c_{ANC}, C_i = c) \quad (15)$$

ANC-adjusted ($\hat{\pi}_{5A}$ & $\hat{\pi}_{5B}$)

Let $\hat{P}(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)$ be an estimator of $P(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)$
then we estimate E$[D_i]$ by

$$\hat{\pi}_5 = \sum_c \sum_x N^{-1}{}_{R_i=0,C_i=c,X_i=x} \sum_{i|R_i=0,C_i=c,X_i=x} \frac{D_i}{\hat{P}(R_i = 0|X_i = x, \hat{\pi}^c_{ANC}, C_i = c)}$$

The first one uses $\hat{\pi}^c_{ANC}$ and a stepwise regression procedure to select from the same list of covariates used in eq.4 to model the propensity score.
The second one uses only $\hat{\pi}^c_{ANC}$ for modelling the propensity score. $\hat{\pi}_{5A}$ and $\hat{\pi}_{5B}$ respectively.

Summary of estimators considered in this study

Method	Name	Key Assumption	Data
$\hat{\pi}_{CC}$	Complete Case	No refusal bias	MDHS
$\hat{\pi}_{MSI}$	Mean Score Imputation	Conditional independence	MDHS
$\hat{\pi}_1$	IPW/Propensity	$P(R_i=0)=P(R_i=0 \mid X_i)$	MDHS
$\hat{\pi}_2$	Never Tested	Prior test independence	MDHS
$\hat{\pi}_{RE}$	Reniers-Eaton	Same as $\hat{\pi}_2 + \Delta$	MDHS+MDICP
$\hat{\pi}_{3\pm}$	Bounds	Monotonicity	MDHS+MDICP+Census
$\hat{\pi}_4$	ANC-based	ANC = population	ANC+Census
$\hat{\pi}_{5A}$	ANC-adjusted IPW (full)	Cond. indep. given ANC	MDHS+ANC
$\hat{\pi}_{5B}$	ANC-adjusted IPW (simple)	Cond. indep. given ANC	MDHS+ANC

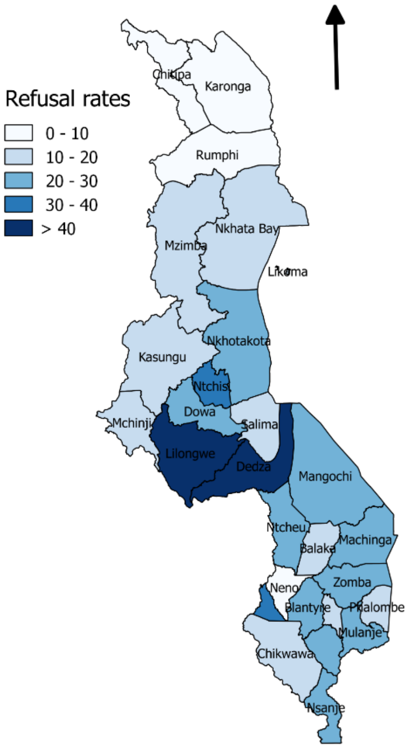
Illustrations

HIV testing refusal patterns

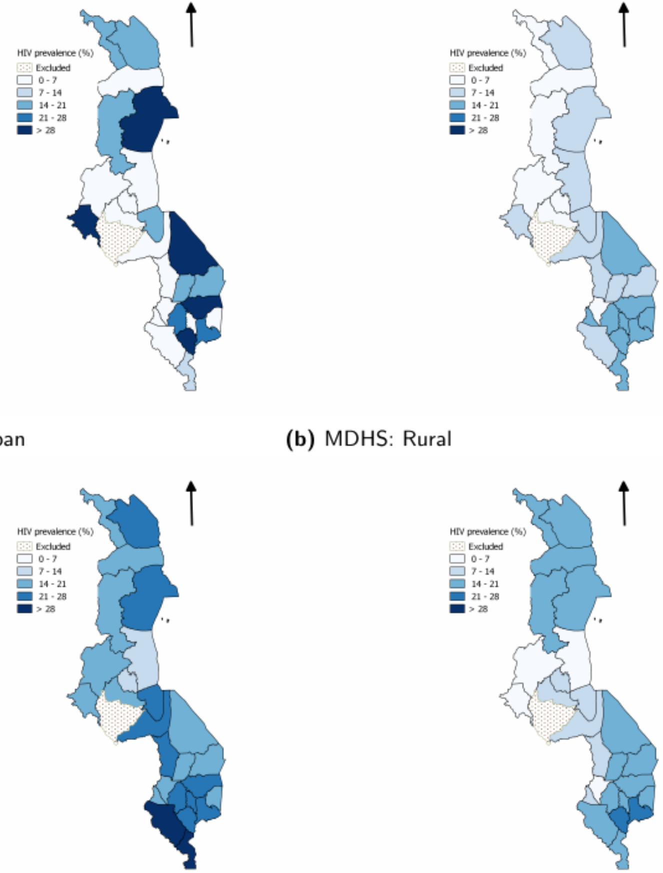
Source	No. Eligible	No. refused	Percent
MDHS	6343	1436	22.6
ANC†	7977	0	0.0
MDICP-3‡	3123	304	9·5
MDICP-4§	2111	115	5·4

†Consent not required
‡Among those contacted in MDICP-3
§Among those tested in MDICP-3 and contacted in MDICP-4

ANC: antenatal clinics; MDHS: Malawi Demographic and Health Survey; MDICP: Malawi Diffusion and Ideational Change Project



Estimated HIV prevalence rates. (a) Complete case estimates using Urban MDHS data. (b) Complete case estimates using Rural MDHS data. (c) District-area estimates using Urban ANC data. (d) District-area estimates using Rural ANC data.



(a) MDHS: Urban

(b) MDHS: Rural

(c) ANC: Urban

(d) ANC: Rural

Adjusted HIV prevalence

Estimator	Men	Women	Overall
$\hat{\pi}_{CC}$	0·1017	0·1521	0·1296
$\hat{\pi}_{MSI}$	0·0998	0·1502	0·1277
$\hat{\pi}_1$	0·0999	0·1490	0·1270
$\hat{\pi}_2$	0·0950	0·1465	0·1238
$\hat{\pi}_{RE}$	0·1006	0·1638	0·1356
$\hat{\pi}_{3-}$	0·0932	0·1421	0·1159
$\hat{\pi}_{3+}$	0·0975	0·1725	0·1323
$\hat{\pi}_4$	—	0·1550†	—
$\hat{\pi}_{5A}^{\ddagger}$	0·1004	0·1490	0·1273
$\hat{\pi}_{5B}^{\dagger\dagger}$	0·0998	0·1510	0·1282

†ANC surveys assume pregnant females rates reflect the national rates

‡Stepwise regression using covariates, X_i and $\hat{\pi}^c_{\text{ANC}}$

††Fixed regression using $\hat{\pi}^c_{\text{ANC}}$ only

District-level HIV prevalence estimates

District	$\hat{\pi}_{CC}$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}^{\dagger}_{5A}$	$\hat{\pi}^{\ddagger}_{5B}$
Blantyre	0·2145	0·2493	0·2043	0·2489	0·2489
Kasungu	0·0535	0·0628	0·0540	0·0625	0·0626
Machinga	0·1210	0·1147	0·1036	0·1147	0·1163
Mangochi	0·2112	0·2409	0·2082	0·2410	0·2409
Mzimba	0·0675	0·0803	0·0634	0·0770	0·0781
Salima	0·0906	0·0763	0·0820	0·0739	0·0755
Thyolo	0·2131	0·2327	0·2116	0·2331	0·2336
Zomba	0·1758	0·174	0·1689	0·1738	0·1738
Mulanje	0·1867	0·1882	0·1805	0·1872	0·1862
Other districts	0·1108	0·1119	0·1098	0·1122	0·1118

†Stepwise regression using X_i and $\hat{\pi}^{c}_{\text{ANC}}$

‡Fixed regression using $\hat{\pi}^{c}_{\text{ANC}}$ only

Concluding remarks and practical implications

- Among people **without prior test results**, refusal rates are similar for HIV-positive and HIV-negative individuals.
- HIV-positive individuals who **know their status** are more likely to refuse retesting.
- Methods using unknown-status individuals (π_{RE}, π_2) and refusal-bound approaches (π_{3-}, π_{3+}) show **no major upward correction**.
- ANC-based estimators (π_4, π_5) also indicate **minimal refusal-related bias**.

Thank you

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(3) 811–826

DOI: 10.1177/0962280219844536

Refusal bias in HIV data from the Demographic and Health Surveys: Evaluation, critique and recommendations

Oyelola A Adegboye,[1] Tomoki Fujii[2] and Denis HY Leung[2]

Abstract

Non-response is a commonly encountered problem in many population-based surveys. Broadly speaking, non-response can be due to refusal or failure to contact the sample units. Although both types of non-response may lead to bias, there is much evidence to indicate that it is much easier to reduce the proportion of non-contacts than to do the same with refusals. In this article, we use data collected from a nationally representative survey under the Demographic and Health Surveys program to study non-response due to refusals to HIV testing in Malawi. We review existing estimation methods and propose novel approaches to the estimation of HIV prevalence that adjust for refusal behaviour. We then explain the data requirement and practical implications of the conventional and proposed approaches. Finally, we provide some general recommendations for handling non-response due to refusals and we highlight the challenges in working with Demographic and Health Surveys and explore different approaches to statistical estimation in the presence of refusals. Our results show that variation in the estimated HIV prevalence across different estimators is due largely to those who already know their HIV test results. In the case of Malawi, variations in the prevalence estimates due to refusals for women are larger than those for men.

Keywords

Bias, Demographic and Health Surveys, missing data, non-response, refusals, Malawi

Get in touch:

oyelola.adegboye@menzies.edu.au
https://x.com/oyeloladegboye
www.linkedin.com/in/oyelolaadegboye/

menzies SCHOOL OF HEALTH RESEARCH

CHARLES DARWIN UNIVERSITY AUSTRALIA