Evaluating Nonresponse Weighting Adjustment for the Population-Based HIV Impact Assessments Surveys: On Incorporating Survey Outcomes

Tien-Huan Lin[1], Ismael Flores Cervantes[1],
Suzue Saito[2,3], Rommel Bain[4†‡]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850
[2]ICAP at Columbia University, New York, NY 10032
[3]Department of Epidemiology, Mailman School of Public Health of Columbia University, New York, NY 10032
[4]U.S. Centers for Disease Control and Prevention, Atlanta GA 30333

Abstract
Population-based HIV Impact Assessment (PHIA) surveys are being conducted in 14 sub-Saharan African countries to measure HIV prevalence and other key impact indicators by ICAP at Columbia University in collaboration with ministries of health and the U.S. Centers of Disease Control and Prevention (CDC) and other partners. The nonresponse weighting adjustment of the PHIA surveys employs the weighting class method in combination with a tree analysis to identify predictors significant to response propensity. Variable selection for this type of nonresponse adjustment identifies auxiliary variables correlated with response propensity alone and produces one set of weights applicable for all analyses of the survey data. An alternative approach identifies auxiliary variables correlated to both the response probability and selected key outcome variables. This approach may identify a different set of variables for the nonresponse adjustments and may produce more efficient estimates for the key outcome variables. This paper utilizes data from several PHIA studies to examine these weighting adjustments, their effects on selected key estimates, and associated variances.

Key Words: nonresponse adjustment, survey outcome, response propensity, principal component analysis, cluster analysis, gradient boosting

1. Introduction

It is common practice in survey research to attempt to mitigate bias due to unit nonresponse by making weighting adjustments to the base weights that account for the sampled units' unequal selection probabilities. There are various methods of developing these adjustments, all of which depend on the availability of auxiliary variables available for both respondents and nonrespondents. The usual approach is to develop the nonresponse adjustments based on models that predict response propensity (Brick & Kalton, 1996). This form of nonresponse adjustment is a general-purpose strategy that is agnostic to the outcomes of the survey. However, a number of researchers have made the argument that the nonresponse adjustments should take into account both the probability of response and

[†] Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the U.S. Centers for Disease Control and Prevention/the Agency for Toxic Substance and Disease Registry.

the survey outcomes in order to reduce bias while controlling for variance (Little & Vartivarian, 2005). The effectiveness of this approach depends on the availability of auxiliary variables that explain response propensity and are predictive of the survey outcomes at the same time.

In this paper, we take a further step in this direction by including model predictions of the actual survey outcomes in developing the nonresponse adjustments, instead of relying on auxiliary variables that may be related to the survey outcomes. Vartivarian and Little (2002) and more recently Morral, Gore, and Schell (2014) and Fay and Riddles (2017) have applied this approach. However, little work has been done in implementing this approach with stratified multi-stage sample designs. This paper explores empirically two methods of applying this approach using data collected in four African surveys that are part of the Population-Based HIV Impact Assessment (PHIA) project. The PHIA surveys have several phases of data collection, with nonresponse occurring at each phase. Large numbers of auxiliary variables are available at later phases from data collected in prior phases. These variables can be used to develop prediction models for both response propensity and for survey outcomes, that then be applied in compensating for nonresponse at a later phase. The results obtained from the proposed approach are compared with those produced using the standard weighting approach as used in the PHIA surveys based only on models for response propensity.

The paper is organized as follows. In Section 2, we describe the PHIA study and its sample design. Section 3 outlines the weighting adjustment methods employed for these surveys including those for nonresponse. Section 4 describes the two methods for incorporating predictions of survey outcomes in the nonresponse adjustment at the final phase of data collection, the collection of blood samples. Section 5 describes the statistical properties of the weights and compares the estimates produced by the three methods and design effects. We conclude in Section 6 with a discussion of our results.

2. The Population-Based HIV Impact Assessment Surveys

The expansion of anti-retroviral treatment (ART) to more than 12.1 million people in sub-Saharan Africa is one of the most successful global public health programs ever undertaken (UNAIDS, 2016). It is by far the largest initiative for a single disease, with the United States alone investing over 70 billion dollars since 2002 (Avert, 2016). After a decade of the anti-retroviral therapy scale-up, the PHIA project, implemented by ICAP at Columbia University in collaboration with the Ministries of Health, the U.S. Centers for Disease Control and Prevention (CDC), and other partners, is assessing the status of the HIV epidemic in 14 sub-Saharan Africa countries by means of nationally representative surveys that measure estimates such as HIV prevalence, HIV incidence, and viral load suppression. This paper utilizes the data from four countries in sub-Saharan Africa with surveys concluded in 2016 and 2017.

The sample design for the four countries was a two-stage sample in which the primary sampling units (PSUs) were enumeration areas (EAs) as defined by the last population census, and with households selected at the second stage. The sampled individuals included all eligible adults in all selected households and all children in a random subsample of the households. There were three stages in the data collection process:

- The first stage consisted of a household questionnaire that collects a household roster, including the age and sex of each household member, as well as responses to a range of items about the household.
- The second stage comprised personal interviews with each eligible adult and interviews with a parent of each eligible child. These interviews covered an extensive set of topics such as sexual activity, male circumcision, female reproduction, and HIV/AIDS-related knowledge and attitudes.
- The third stage consisted of blood sample collection by trained phlebotomists from respondents who agree to HIV testing at the end of the personal interviews.

Nonresponse occurred at each of the three data collection stages. In this paper, we focus on the blood test nonresponse adjustment for sampled adults age 15-49 at the last stage of data collection.

Table 1 displays the conditional blood test response rates (conditional on interview response) for the four countries. Male and female response rates are shown separately because men and women received different questions in their interviews (e.g., questions about male circumcision and female reproduction).

Table 1: Conditional Blood Test Response Rates for Adults Age 15 to 49 by Sex for Four PHIA Countries

	Country			
Sex	*A*	*B*	*C*	*D*
Male	86.2%	86.9%	88.3%	90.0%
Female	87.8%	86.7%	90.3%	91.9%

3. The Weighting Procedure Used in PHIA to Compensate for Blood Test Nonresponse

In the PHIA weighting process, blood test weights were developed for each country by first adjusting the person-level design weights for interview nonresponse, and then adjusting nonresponse-adjusted interview weights for nonresponse to the blood draw. Finally, the blood sample adjusted weights were poststratified to national age and sex projections. See Lin, Weil, Flores Cervantes, and Saito (2017) for additional details.

In PHIA surveys, the nonresponse adjustments at each stage were computed as the inverses of the weighted response rates in the weighting classes. The weighting classes were created through a two-stage procedure primarily to reduce time and labor (Lin, Weil, Flores Cervantes, & Saito, 2017). The first stage was "feature filtering" or variable selection through the Least Absolute Shrinkage and Selection Operation (LASSO) regression, which is a penalized or regularized regression from the field of machine learning (Tibshirani, 1996). The LASSO regression was implemented via the SAS procedure `PROC HPGENSELECT`. The second stage employed the Chi-squared Automatic Interaction Detector (CHAID) tree classification algorithm (Magidson, 2005) for the final variable selection and for creating the weighting classes, implemented by a stand-alone software SI-CHAID. Both packages took into account the unequal selection probabilities and prior phase nonresponse adjustment by utilizing survey weights in the algorithms. The number of weighting cells created ranged from 23 to 43 for adult males and 32 to 44 for adult females.

4. Alternative Weighting Procedures to Produce More Efficient Estimates for Key Outcome Variables

Our research produced two alternative blood test nonresponse adjusted weights that incorporated survey outcomes into the blood test nonresponse adjustment. The analyses presented in this paper exclude the effect of the poststratification factor. We now describe the alternative methods. Five key survey outcomes of the PHIA study were used in the two methods:

- HIV prevalence estimate;
- Viral load suppression rate among all adults;
- Percent of people living with HIV who were aware of their HIV status;
- Percent of people living with HIV who knew their HIV status and received sustained antiretroviral therapy; and
- Percent of people living with HIV who knew their HIV status, received sustained antiretroviral therapy, and had viral load suppression.

The five key survey outcomes were nested and the latter three were associated to UNAID's 90-90-90 treatment target (http://www.unaids.org/en/resources/909090). The methods described in this section are implemented separately by sex, for reasons discussed in section 3.

4.1 Joint Classification by Response Propensity and Predictive Mean Stratification

The joint classification by response propensity and predictive mean stratification method for adjusting for nonresponse is described in Vartivarian and Little (2002). Their illustration of this method applies to a single survey outcome; however, they suggest options for adjusting this method to work for more outcomes (e.g., using a principal component analysis to reduce the number of variables). The first step, response propensity stratification, uses logistic regression to model and predict response propensities for all sampled cases. The predicted response propensities are then grouped or "stratified" to form a set of propensity strata. The second step, predictive mean stratification, models the survey outcome for respondents using regression analysis. The fitted model is used to predict the survey outcome for both respondents and nonrespondents. Similar to the response propensity stratification method, the predicted survey outcomes are grouped or stratified into a set of strata based on the predictions. The final step forms nonresponse adjustment cells as the cross-classification of the two sets of strata in order to take advantage of both response and outcome models.

For our analysis, we modified the Vartivarian and Little approach in three ways. The first modification, mostly introduced as a time-saving change, was to use the response propensity from the same weighting cells we used for the PHIA blood test nonresponse adjustment instead of modeling a new response propensity from a logistic regression model. The second modification was to use a principal component analysis as suggested by Vartivarian and Little, to reduce the number of survey outcome variables to a smaller set of uncorrelated principal components (PCs) (Pearson, 1901) (see, for example Rao, 1964 and Morrison, 1976). We implemented this analysis using the SAS procedure `PROC PRINCOMP`. For each analysis group, the number of outcome variables was reduced from five to two, retaining on average 90 percent of the total variance.

We used the SAS procedure `PROC GLM` to implement the predictive mean stratification. The predictors in the GLM regression model were filtered by the SAS procedure `PROC GLMSELECT` with forward selection with an initial model that include all auxiliary variables from the household and person interviews. The interview weights were used in both procedures to account for unequal selection probability and interview nonresponse.

The third modification we introduced was to replace the cross-classification of predicted mean and propensity strata by a cluster analysis known as *k*-means model, where the cluster centers are the means of the observations assigned to the cluster. The cluster analysis was implemented using the SAS procedure `PROC FASTCLUS` and the number of clusters or cells created was dependent on the number of blood test nonresponse adjustment cells of the regular PHIA weighting process. The number of cells ranged from 23 to 30 for adult males and 26 to 39 for adult females. These cells were used as weighting classes for computing the blood test nonresponse adjustments for the joint classification weights.

To visualize the weighting classes, we ran a canonical analysis using the SAS procedure `PROC CANDISC` to produce plots of the clusters by their three largest canonical components. Figure 1 shows the plots for the first two canonical components separately by male and female for the four countries. The clusters are indicated by different colors. In all cases, the plots show clear clustering of the predicted values.

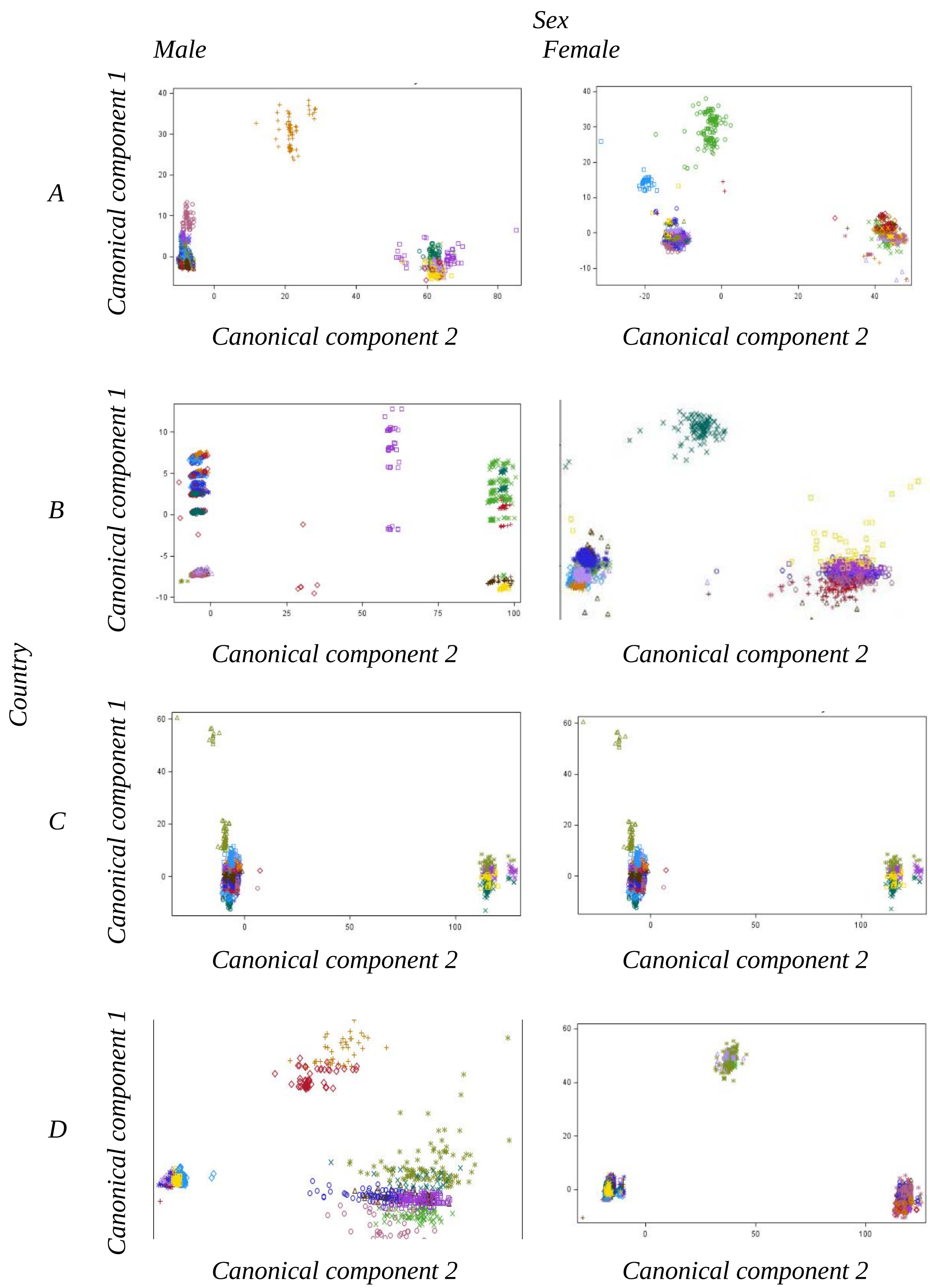

Figure 1: Two-dimension projection of cluster analysis results by sex for four countries.

4.2 Two-Step Approach with Gradient Boosting

The second alternative weighting method is an application of the work of Morral, Gore, and Schell (2014) and Fay and Riddles (2017), labeled as the two-step approach by Fay and Riddles.

In the first step, separate models were fitted for each of the five key survey outcomes using the respondent's household and person interview variables to predict the key outcomes for both respondents and nonrespondents. Both Morral, Gore, and Schell (2014) and Fay and Riddles (2017) used a machine learning algorithm known as gradient boosting (GB) method that fits a prediction model consisting of an ensemble of weak prediction models (based on classification trees). The predictions are based on a "committee" formed from the weak predictions (Hastie, Tibshirani, & Friedman, 2009).

Morral, Gore, and Schell (2014) applied the algorithm with the `xgb` package in R and Fay and Riddles (2017) used both `xgb` and the R package `xgboost` (Chen et al., 2018). We developed our models using `xgboost` with cross-validation to avoid over fitting. The models for the five outcomes from the GB algorithm were used to predict the outcomes for respondents and nonrespondents. In the second step, a GB model for response propensity was fitted using the five predicted survey outcomes for respondents and nonrespondents. The predicted response propensities were grouped by percentiles in order to form weighting classes for the two-step weights. The interview nonresponse adjusted weights were then adjusted for blood test nonresponse by the inverse of the weighted response rates within these weighting classes.

5. Comparison

In this section, we compare the estimates and various statistics for the blood test nonresponse adjusted weights created using the PHIA, joint-classification, and two-step methods. First, we investigate the differences in estimates and variances computed using the weights from each method. We then compare the design effects of the estimators.

5.1 Assessing differences in Estimates and Variances

Table 2 shows the unadjusted blood test estimates (weighted by interview nonresponse adjusted weight, prior to adjusting for blood test nonresponse) and the blood test nonresponse adjusted estimates by the three weighting methods (PHIA, joint classification, and two-step) by sex and country for selected survey outcomes.

Overall, the weighted estimates were lower from the unadjusted estimates, suggesting that all three weighting methods corrected for bias. However, the differences were small. This was because the blood-test response rates for these countries were high and hence the nonresponse adjustment did not have a large impact. For example, for HIV prevalence rate, the PHIA method reduced the estimate by an average of 0.5 percentage points for males and 0.97 percentage points for females across countries. For the joint-classification method, the reduction was moderately larger with an average of 0.5 percentage points for males and 0.98 for females. The largest reduction appeared in males from the two-step method with an average of 0.6 percentage points. On the contrary, females from the two-step method only have a reduction of 0.95 percentage points.

The differences between the PHIA estimates and the alternative weights were much smaller. For the HIV prevalence rate, the average difference between the joint-classification method and PHIA method was 0.15 percentage points for males and 0.13 for females across countries. The average difference between the two-step method and PHIA method was larger for males (0.28 percentage points) but smaller for females 0.1

percentage points). The statistical test[1] for the differences between the PHIA estimates and the other methods showed no differences in most of the cases. For the other survey outcomes, the differences in percentage points in the estimates among the unadjusted, PHIA, joint-classification, and the two-step were higher, but the same pattern as described above holds. These small differences indicate that all three methods correct for bias in a similar fashion.

Table 2: Estimates of Key Survey Outcomes by Country, Gender, and Weighting Method, Unadjusted and Adjusted for Blood Test Nonresponse

HIV Prevalence Rate of adults 15 to 49 years old

Sex	*Method*	*A*	*B*	*C*	*D*
Male	Unadjusted	19.4	8.1	8.8	11.4
	PHIA	18.8	8.1	8.5	10.9
	Joint-class	18.7	7.7*	8.4	10.9
	Two-step	18.4*	7.6*	8.4	10.8*
Female	Unadjusted	30.7	13.2	15.0	17.1
	PHIA	29.2	12.6	14.4	16.4
	Joint-class	29.2	12.3*	14.3	16.3
	Two-step	29.1	12.5	14.3*	16.3

Percentage of HIV positive adults 15-49 years old who are aware of their HIV status

Sex	*Method*	*A*	*B*	*C*	*D*
Male	Unadjusted	68.9	66.3	59.9	68.0
	PHIA	68.0	65.8	58.8	66.4
	Joint-class	67.5	63.9*	58.3	66.4
	Two-step	67.0*	64.0*	58.0*	66.0*
Female	Unadjusted	82.1	76.4	68.9	76.9
	PHIA	81.0	75.7	67.3	75.7
	Joint-class	80.8	74.9*	67.2	75.6
	Two-step	80.9	72.5*	67.2	75.2*

[1] The statistical test takes into account the high correlation between the estimates. That is, the estimates are based on the same data and weighting components except for the blood test adjustment.

Table 2: Estimates of Key Survey Outcomes by Country, Gender, and Weighting Method, Unadjusted and Adjusted for Blood Test Nonresponse (continued)

Percentage of HIV positive adults 15-49 years old who received Sustained Antiretroviral Therapy

Sex	*Method*	*A*	*B*	*C*	*D*
Male	Unadjusted	60.6	55.9	50.2	56.8
	PHIA	59.9	55.5	49.4	55.4
	Joint-class	59.5	53.6*	49.0	55.4
	Two-step	58.9*	54.0*	48.6*	55.1
Female	Unadjusted	73.8	69.8	57.7	66.2
	PHIA	72.7	69.1	56.3	65.1
	Joint-class	72.3*	68.7*	56.3	65.0
	Two-step	72.5	66.4*	56.3	64.9 *

Percentage of HIV positive adults 15-49 years old who received Sustained Antiretroviral Therapy and achieved Viral Load Suppression

Sex	*Method*	*A*	*B*	*C*	*D*
Male	Unadjusted	52.6	49.6	43.4	46.6
	PHIA	51.9	49.5	42.8	45.4
	Joint-class	51.6	47.7*	42.4	45.4
	Two-step	51.2*	47.9*	42.0*	45.2
Female	Unadjusted	64.3	64.1	51.4	57.4
	PHIA	63.3	63.4	50.0	56.4
	Joint-class	63.0*	63.0	50.1	56.4
	Two-step	63.1	61.0*	50.1	56.3

*Difference to PHIA estimate is statistically significant at α=0.05% level

5.2 Design Effects for Selected Estimates

In this section, we compare the design effects of the estimates of the survey outcomes for the three sets of nonresponse-adjusted weights. We expected smaller values of the design effects for the alternative weighting methods since these methods targeted only the nonresponse bias of the key survey outcomes, thus reducing the variability of the weights. Table 3 shows the design effect of three survey outcomes by gender, weighting method, and country. The average reduction of design effect of HIV prevalence rate with respect to the joint-classification estimates to the PHIA estimates was 0.05 for males and females across countries. Although with smaller reductions in design effects, the same pattern held for the other survey outcomes for the joint-classification method. These results matched our expectation. However, the extent of improvement in efficiency among the survey outcomes were not the same. This was because a weight that is efficient for one variable is not necessary efficient for another since the efficiency depends on the correlation of the weights and the outcome. Targeting weights for one survey outcome improves the efficiency of that estimate, but it may decrease the efficiency of the other outcome if the two survey outcomes are not correlated.

The design effect of the two-step method for HIV prevalence rate of males showed a gain in efficiency. However, this pattern did not hold for females and for other survey outcomes presented in the table. These results suggested that the two-step method produced less efficient estimates than the joint-classification method (although the differences in estimates is not significant). Additional research is needed to understand the role of the models in the two-step method to explain the loss of efficiency.

Table 3: Design Effects for Selected Estimates for Blood Test Nonresponse Adjusted Weights by Country, Gender, and Weighting Method

HIV Prevalence Rate of adults 15 to 49 years old

Sex	*Method*	*A*	*B*	*C*	*D*
Male	PHIA	1.07	1.25	1.29	1.28
	Joint-class	1.02	1.19	1.20	1.28
	Two-step	0.99	1.16	1.22	1.32
Female	PHIA	1.16	1.76	1.46	1.19
	Joint-class	1.13	1.75	1.43	1.20
	Two-step	1.13	1.78	1.44	1.21

Percentage of HIV positive adults 15-49 years old who are aware of their HIV status

Sex	*Method*	*A*	*B*	*C*	*D*
Male	PHIA	0.91	1.38	1.32	1.37
	Joint-class	0.90	1.34	1.35	1.36
	Two-step	0.91	1.37	1.37	1.42
Female	PHIA	1.10	1.09	1.34	1.15
	Joint-class	1.10	1.05	1.34	1.16
	Two-step	1.11	1.16	1.36	1.18

Percentage of HIV positive adults 15-49 years old who received Sustained Antiretroviral Therapy

Sex	*Method*	*A*	*B*	*C*	*D*
Male	PHIA	0.94	1.36	1.35	1.18
	Joint-class	0.93	1.34	1.36	1.17
	Two-step	0.92	1.38	1.38	1.23
Female	PHIA	1.18	1.25	1.47	1.23
	Joint-class	1.20	1.18	1.40	1.18
	Two-step	1.20	1.27	1.45	1.22

6. Conclusion

In this paper, we explored the notion of producing efficient estimates (i.e., estimates smaller variances) by including key survey outcomes in response propensity models. We implemented two alternative weighting methods on data from four countries of the PHIA surveys. The first method was an expansion on Vartivarian and Little's joint-classification by response propensity and predictive mean stratification method. The second method is an application of a machine learning algorithm studied by several researchers.

The results of our analyses showed that all three methods adjust the estimates downward compared to the unadjusted estimates, and that there was little difference among estimates produced by the alternative weighting methods and the PHIA estimates. In terms of design effects of the estimates, the joint-classification method produced more efficient estimates compared to the PHIA method. This observation does not hold for the two-step method.

Additional research is needed to understand the role of the models and the algorithm in this approach.

There are some limitations when developing nonresponse-adjusted weights from the alternative methods. The joint-classification method is more time-consuming than the PHIA method since in addition to modeling response propensity, it also requires modeling of survey outcomes and cluster analysis to create weighting adjustment cells. On large-scale multi-country studies such as PHIA where time and budget are of the essence, this can be an important driving factor.

The `xgboost` package of the two-step approach included numerous parameters that needs to be "tuned." Therefore, the weighting adjustments may not be robust to the parameters. However, the main drawback of the two-step method is that the algorithm is a black box and there is no easy way to understand the role of the selected variables in the models.

The weights produced by the alternative weighting methods were useful as evaluation tools of the public-use weights (i.e., PHIA weights). As an evaluation of the PHIA weights, the results presented in this paper suggest that the PHIA weights, which do not take into account the outcome variables, perform well compared to weights derived for key survey outcomes. Note that we do not advocate the use of weights developed for specific outcomes. These weights would produce efficient estimates for those variables correlated to the outcome but they are inefficient for those that are not. As a multi-purpose weight, the efficiency of the estimates produced by the PHIA weighting method closely resembles those specifically targeted at key survey outcomes.

References

Avert. (2016). *Funding for HIV and AIDS*. Retrieved September 26, 2017, from https://www.avert.org/professionals/hiv-around-world/global-response/funding.

Brick, J., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., . . . Li, Y. (2018). xgboost: Extreme Gradient Boosting, R package version 0.71.2. Retrieved from https://cran.r-project.org/web/packages/xgboost/index.html.

Fay, R. E., and Riddles, M. K. (2017). "One- Versus Two-Step Approaches to Survey Nonresponse Adjustments." *JSM Proceedings* (pp. 953-964). Baltimore, MD: American Statistical Association.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). "Boosting and Additive Trees." In *The Elements of Statistical Learning* (2nd ed., pp. 337-384). New York: Springer.

Lin, T.-H., Weil, N., Flores Cervantes, I., and Saito, S. (2017). "Developing Nonresponse Weighting Adjustments for Population-Based HIV Impact Assessments Surveys in Three African Countries." *JSM Proceedings* (pp. 965-982). Baltimore, MD: American Statistical Association.

Little, R. J., and Vartivarian, S. (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Statistics Canada*, *31*(2), 161-168.

Magidson, J. (2005). SI-CHAID 4.0 user's guide. Retrieved September 2017, from https://www.statisticalinnovations.com/wp-content/uploads/SICHAIDusersguide.pdf

Morral, A. R., Gore, K. L., and Schell, T. E. (2014). "Sexual Assault and Sexual Harassment in the U.S. Military: Volume 1. "Design of the 2014 RAND Military

Workplace Study. Santa Monica, California: RAND Corporation. Retrieved from www.rand.org/t/RR870z1

Morrison, D. (1976). “Multivariate Statistical Methods” (2nd ed.). New York: McGraw-Hill.

Pearson, K. (1901). “On Lines and Planes of Closest Fit to Systems of Points in Space.” *Philosophical Magazine*, *6*(2), 559-572.

Rao, C. R. (1964). “The Use and Interpretation of Principal Component Analysis in Applied Research.” *Sankhya A*, 26, 329-358.

Tibshirani, R. (1996). “Regression Shrinkage and Selection Via the LASSO.” *Journal of the Royal Statistical Society*, *58*(1), 267-288.

UNAIDS. (2016). “Global AIDS Update.” Geneva: Joint United Nations Programme on HIV/AIDS. Retrieved September 26, 2017, from http://www.unaids.org/sites/default/files/media_asset/global-AIDS-update-2016_en.pdf

Vartivarian, S., and Little, R. J. (2002). “On the Formation of Weighting Adjustment Cells for Unit Nonresponse.” The University of Michigan Department of Biostatistics Working Paper Series, Working Paper 10.