THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

A Naïve Methodology for Imputing Missing Survey Information due to Survey Skip Conditions

GRANT HOPKINS
SPRING 2022

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Statistics, Mathematics, and Finance
with honors in Statistics

Reviewed and approved* by the following:

Le Bao
Associate Professor of Statistics
Thesis Supervisor

Matthew Beckman
Assistant Research Professor of Statistics
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

Surveying a sample to make an inference upon a population is a fundamental role of statistics. In the simplest cases, a survey is conducted upon respondents who are selected via simple random sampling, all respondents answer all questions with no missing information, and the survey gives meaningful insight into a population of interest. In reality, however, it is often necessary to employ complex sampling designs in order to reach representative respondents and to collect large amounts of information without introducing survey fatigue. Moreover, there are also cases where respondents refuse to answer certain questions, do not know the answer to certain questions, or even provide inaccurate answers. For this reason, it is infeasible and unfavorable to ask respondents questions that they have already answered, cannot answer, would likely decline to answer, or would likely not know the answer.

Such is the premise of the Population-Based HIV Impact Assessment [PHIA] survey, conducted across multiple countries in Sub-Saharan Africa to understand the status of the HIV epidemic in those countries. A particular challenge of analyzing the PHIA survey is that information about a respondent is shrouded in survey skip conditions, prohibiting an analyst from understanding why a respondent does not have an answer to a particular question. Perhaps the respondent's answer can be deduced from an earlier question; perhaps the respondent's answer is impossible due to a logical inconsistency; perhaps the respondent's answer to the question is missing and may be predicted.

This thesis proposes a naïve methodology that researchers can use to probabilistically predict missing information in an indicator variable in the context of large surveys that utilize skip conditions. First, I propose a variable selection method based upon marginal association

with the indicator variable and the proportion of non-skipped values. Second, I discuss the need to impute skipped values among the predictor variables in order to have a fully-specified predictor matrix upon which the response is modelled. Next, I implement the LASSO procedure for subsetting to a sparse set of predictor variables. Finally, I train a logistic regression model on respondents with non-missing indicator values, assess the model performance, and apply the model to respondents with missing indicator values.

In addition to researchers who wish to model with data from surveys with skip conditions, designers of such surveys may take interest in the discussion surrounding data encoding. Surveys with skip conditions have the great potential to discover niche behavioral patterns and risk factors by targeting questions based upon preceding responses. Improving data encodings will shed light into what subpopulations a particular pattern holds for, and will also provide clarity into the reasons for missing information throughout the survey. Ignoring this missing information may bias sample estimates for population parameters.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many exceptional individuals who have inspired my interest in research, but I am especially grateful for the mentorship of Dr. Le Bao and Dr. Michael Daniels as I developed as a researcher. I have benefited tremendously from weekly meetings with Dr. Le Bao, Dr. Michael Daniels, and David Lindberg. I am also thankful for Dr. Matthew Beckman for his constant support towards my undergraduate education, and his engaging instruction in Introduction to R, which solidified my interest in applied statistics. Finally, I would like to thank my family and friends for their unwavering support during my undergraduate studies.

This work was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI136664. We appreciate the PHIA teams making the data publicly available for research purposes.

Chapter 1

Introduction

Since the first reported case of the HIV/AIDS epidemic in 1981, the epidemic has disproportionately impacted numerous subpopulations, including men who have sex with men [MSM], female sex workers [FSW], and sex worker clients [SWC]. At the United Nations General Assembly High-Level Meeting on AIDS in June 2021, the United Nations announced a goal to end all inequalities facing subpopulations who are disproportionately impacted by the epidemic (UNAIDS, 2021). It is therefore essential to understand the state of the epidemic within key subpopulations, so as to understand what groups would benefit from targeted public health interventions, and to subsequently track the success of those interventions over time.

With that being said, there is no single way to measure the state of the HIV epidemic. Researchers may have interest in understanding the prevalence of HIV infection, either in the entire population or a subpopulation. However, HIV infection is a chronic disease that has only ever been cured in 3 patients at the time of this writing (Mandavilli, 2022), but has effective treatments to reduce spread and prevent onset of AIDS (Deeks et al., 2013). For this reason, the prevalence of the HIV infection will not change substantially from year-to-year. Another way to measure the state of the epidemic is then to estimate the proportion of cases that are new cases, which is known as infection incidence. An HIV positive individual's probability of infection within the past year can be estimated using biomarker information, i.e. physical attributes that are associated with the virus, including CD4 lymphocyte count, viral load, and optical density. While there are numerous ways to estimate HIV incidence, this thesis applies the methodology

from “Probabilistic HIV Recency Classification -- A Logistic Regression without Individual Level Training Data” (Sheng et al., 2021). These individual-level probabilities can then be aggregated to produce mean infection recency estimates for a key subpopulation.

One important source of data for these metrics is the Population-Based HIV Impact Assessment [PHIA], managed by the U.S. Centers for Disease Control and Prevention [CDC]. The PHIA surveys are conducted across numerous years and countries in Sub-Saharan Africa, sampling thousands of individuals in each country. The PHIA surveys collect demographic, behavioral, and biomarker information from every consenting participant, allowing researchers to understand the development of the HIV epidemic.

However, complex survey designs are infamous for challenges that prohibit simple statistical modelling choices. For example, it would be illogical to ask biological men if they have ever been pregnant, as biological men cannot birth children. For this reason, an efficient survey would elect to skip this question for biological men. Beyond this single example, one can consider that surveys seeking to find complex behavioral associations will ask narrow follow-up questions only when applicable. For example, “Did you ever XYZ?” followed by “When is the last time that you XYZ?” given an affirmative answer. This is what is meant by a survey skip condition: when the answer, or lack-there-of, to a prior question causes certain follow-up questions to never be asked.

Often times, behavioral questions in the survey are used to identify subpopulations of interest. For example, researchers may be interested to test whether active FSW have a significant difference in mean HIV infection recency, which would give insight into how the virus continues to be transmitted. However, the variable pertaining to whether or not respondents have sold sex in the past 12 months, ‘sellsx12mo’, is preceded by skip conditions according to

prior responses about sexual activity. For example, if a respondent described that they have had 0 partners in the past 12 months, then ‘sellsx12mo’ question is skipped entirely and instead labeled with a period (“.”). If a respondent declines to answer how many partners in the past 12 months, then ‘sellsx12mo’ question is also skipped entirely and labeled with a period (“.”). While these encodings are identical, the information is entirely distinct. In the first case, we have certain knowledge that the respondent is not an active FSW, whereas in the second case, we cannot deduce the respondent’s status. Moreover, respondents who have had no partners in the past 12 months are not sexually active; this likely means that the respondents were not infected with HIV in the past year, however it is important to note that sexual intercourse is not the only way HIV can be transmitted. Hence, researchers need to take special care to subset their sample according to their population of interest. There are two primary challenges that this introduces: 1) researchers must carefully investigate the questionnaire to understand what any given variable encodes in the context of preceding variables, and 2) immediately subsetting to the population of interest ignores missing information throughout the survey.

This thesis delves into the missingness pattern that arises in surveys with skip conditions, discusses a naïve methodology for imputing missing values, and offers a preliminary method to encode the missingness mechanism to ensure that all survey variables may be easily used in analysis.

Chapter 2

Data Overview

The PHIA surveys were conducted in several countries in Sub-Saharan Africa to understand the state of the HIV epidemic. The surveys were performed in 12 countries, each with some distinctions from the other surveys, but with substantial overlap in theme and phrasing of questions. Nevertheless, since each country's survey has a different design, this thesis will focus on the Malawi Population-Based Impact Assessment [MPHIA], conducted from 2015 to 2016. MPHIA consists of thousands of individuals who live in Malawi and contains hundreds of variables for each respondent, corresponding to the sequentially asked questions in the survey.

Survey Design

MPHIA is a household survey, which has a key advantage in underserviced communities for reaching representative individuals from the population in a probabilistic and systematic manner. In particular, MPHIA employs a two-stage, stratified cluster sample design, with clusters stratified by geography and subsequently sampled via probability proportional to size, then households in each cluster sampled via equal probability (PHIA Data Use Manual, 2021, pg. 22). Due to the complex nature of this survey design, estimator variance is estimated via respondent reweighting using a form of the jackknife replication method (PHIA Data Use Manual, 2021, pg. 28), as opposed to a closed-form formula for estimator variance. Existing estimates are constructed using analytic variables that logically deduce a respondent's status for a population of interest using all available survey information. However, when the population status cannot be deduced for any reason, the analytic variable is coded as ("99"), meaning the

status is missing. This (“99”) does not offer insight as to why a respondent has no label for the population status of interest, yet the survey design under skip conditions inherently means that the missingness is not Missing Completely At Random [MCAR]. That is to say, estimates are not necessarily unbiased with ignorable missingness. In the best-case scenario, these missing data are Missing At Random [MAR], meaning they are viable for valid imputation conditioning on available information. In the worse-case scenario, these missing data are Missing Not At Random [MNAR], meaning an imputation procedure will not eliminate bias in the estimates (Berdikulov, 2019).

Whether the data are MAR or MNAR, both cases benefit from a researcher being able to understand why pertinent questions were skipped. A researcher who is optimistic of the MAR assumption could more easily consider imputation mechanisms. A researcher who is pessimistic of the MAR assumption could more easily perform sensitivity analysis for specific skip conditions. For example, perhaps active FSW are more likely to decline to answer if they have sold sex in the past 12 months compared to non-active FSW. In this case, the researcher may want to consider two ad hoc estimates, one where those respondents are included in the sample as active FSW and one as non-active FSW. Similar sensitivity analysis could be performed for other skip conditions.

Data processing

Prior to modelling with the data, it was first necessary to perform some rudimentary data processing. Foremost, we subset to only respondents who are representative of the population of interest, as explicitly stated in the survey instructions.

There are a small number of paired variables that encode time since an event, where one variable encodes an integer value and another variable encodes the corresponding units. We instead recode those integers to be on the same scale, i.e., all integers in terms of days. There are some questions where respondents are instructed to "select all that apply". In those instances, we code non-affirmative answers to mean "not selected"; it is worth noting that the answer may not be selected due to the question set being non-applicable. We also encode "select all that apply" questions to be represented as a single variable, where multiple selections are encoded as a single "interaction" level. We remove all variables that were skipped by every female participant. As a technical detail, we remove white space trailing every entry in the dataset.

Finally, it is important to describe the logical deductions used to identify active FSW status for as many respondents as possible. Our construction bears some similarity to the 'paidsex12months' variable constructed by the PHIA team, provided in Appendix A. However our deductions have some small differences. To construct the response variable, active FSW status, we perform the following logical evaluations using 9 variables in the survey:

- If the respondent had 0 partners in the past 12 months ('part12monum' == 0), then the respondent is not an active FSW.
- If the respondent never had vaginal sex and never had anal sex ('firstsxagedk' == 96 AND 'ansxyn' == 2), then the respondent is not an active FSW.
- If the respondent has never sold sex ('sellsxever' == 2), then the respondent is not an active FSW.
- If the respondent has not sold sex in the past 12 months ('sellsx12mo' == 2), then the respondent is not an active FSW.

- If the one of respondent's most recent three partners in the past 12 months includes a sex worker client ('partrelation1' == 6 OR 'partrelation2' == 6 OR 'partrelation3' == 6), then the respondent is an active FSW.
- If the respondent's most recent partner in the past 12 months was a sex worker client ('lastpartnerrelation12months' == 6), then the respondent is an active FSW.
- If the respondent has sold sex in the past 12 months ('sellsx12mo' == 1), then the respondent is an active FSW.
- Otherwise, the active FSW status for the respondent cannot be determined.

Ultimately, while it is true that respondents who are not sexually active are particularly not active FSW, it is important to note that the population of interest for testing mean HIV infection recency is HIV positive individuals who were sexually active in the past 12 months. The reason we must encode these respondents as not active FSW is because training a prediction model on only sexually active respondents could bias predictions. Therefore, once the final active FSW status predictions are made, it is important to manually subset out respondents who are not sexually active before aggregating the individual-level probabilities.

Chapter 3

Methods

Step 1: Marginal Association and Skip Proportion Screening

Noting that MPHIA has many variables that were answered by only a small subset of respondents, the immediate challenge is realizing how one can predict missing information using covariates that themselves have skipped data. There are two conditions that predictive variables must satisfy: 1) the variables are predictive of active FSW status, and 2) the variables have sufficient non-skipped and non-missing information, both in the observed and unobserved cases of active FSW status. Therefore, the methodology first begins with a screening of all variables in the survey to check which variables satisfy these two conditions. However, a challenge of variable selection in a complex survey design is that there are many types of variables, including both categorical and numerical variables. Therefore, our screening for marginal association of variables with the response, active FSW status, must take into consideration the variable type. Fortunately, the variable codebook conveniently describes nearly every variable as one of {select_one, select_multiple, integer, continuous}.

For variables of type select_one or select_multiple, we perform a Chi-Square test for independence with active FSW status to test if active FSW status is approximately proportional in each level of the variable or not. As a rule of thumb, the Chi-Square test should not be used when there are small expected counts among the contingency tables (McHugh, 2013). In order to maintain a statistically rigorous methodology without manually screening each contingency table, we simulate the p-value for independence using 2000 Monte Carlo simulations as opposed

to using the Chi-Square test statistics itself. The simulation is performed via the chisq.test function in the ‘stats’ R package.

For variables of type integer or continuous, we perform an ANOVA test for a difference in means of the predictor variable between active FSW and non-active FSW. The ANOVA test is more generalizable than a t-test and produces an equivalent result under the equal variance assumption. The p-value is directly extracted from the F-statistic of the ANOVA model.

For each predictor, we calculate the proportion of cases that are non-skipped for respondents that have observed active FSW status. This proportion is useful to understand the number of samples used to estimate the test statistics. For each predictor, we also calculate the proportion of cases that are non-skipped for respondents that have missing active FSW status. This proportion is useful to understand how many respondents with missing active FSW status the variable will be able to directly predict.

Finally, we decided to subset to only predictors that have a p-value below 0.05 and at least 90% of samples for both non-missing and missing active FSW status have a non-skipped and non-missing values for the predictor.

Step 2: Skipped Value Imputation among the Predictors

Once the predictor variables have been selected, the next step is to predict the skipped values among the predictors. This is because the final logistic regression model requires a fully-specified predictor matrix. While there are numerous options for the imputation procedure, we elected to use the Multivariate Imputation by Chained Equations [MICE] procedure using the ‘mice’ R package. The user is given considerable flexibility to model, but for sake of simplicity,

we proceeded using predictive mean matching [PMM] regardless of variable type, as it is simple to implement and only returns values present in the observed data. This helps to ensure that non-sensical imputations, like negative counts, are avoided (Buuren, 2018). However, the imputations among several covariates may still have inconsistencies and logical errors. For example, consider the variables ‘age’ and ‘agemar’. Note that the variable ‘agemar’ is not selected as a predictor due to having a large proportion of skipped or missing answers, however it serves as a clear example. We know that the age someone was married must be less than or equal to their current age, however the MICE imputation procedure under PMM does not impose such a constraint. In principle, imputations should be performed in the order in which the questions were asked, as subsequent questions depend upon preceding ones. This helps to find a suitable match for PMM, but still does not ensure the logical condition is met.

A related challenge is that the predictors suffer from similar skip conditions as the response variable. For example, the ‘agemar’ variable was only asked of respondents who previously identified that they were married. Imputing a value for a respondent who is not married would be contextually unsuitable, and could even prevent proper identification of active FSW status. So there is a tradeoff between comprehensive predictor screening and information compression. A researcher could carefully investigate the survey skip commands for a few covariates of interest to encode information-dense, complete, accurate variables. Alternatively, a researcher could ignore the survey skip commands in the covariates and only look at all covariates that already meet desired benchmarks for association and non-missingness. This will effectively remove focused behavioral questions from consideration, leaving only general variables that can be asked of all respondents.

Ultimately, our method is the naïve approach: to ignore survey skip commands in the covariates. Future analysis may consider variable transformations, such as coding skipped levels as 0 in numerical variables, to allow for detection of niche behavioral dependencies. The primary challenge in this approach is that the current presentation of data does not easily allow us to discern when a skipped variable is viable for imputation or not (i.e., if it is logically impossible or redundant given an earlier answer). Encoding truly missing values as 0 (i.e., no effect) could similarly prevent identification of active FSW status.

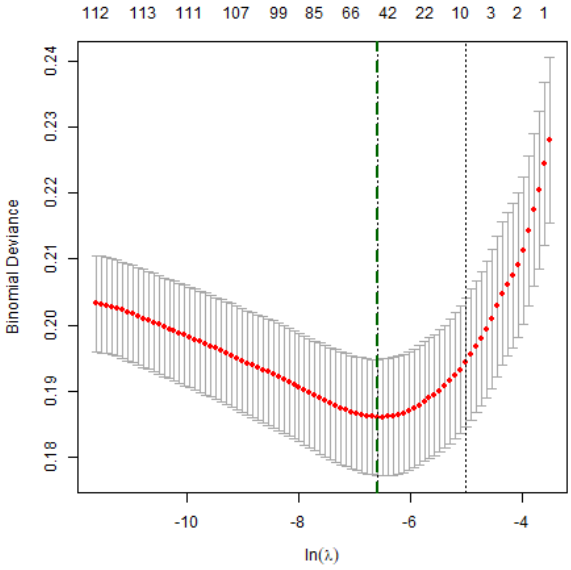
Step 3: LASSO Regression to Increase Sparsity

Once skipped values in the predictors have been imputed using the MICE procedure, we use LASSO regression to find a subset of covariates that are predictive of the response variable with minimal increase in prediction error. We utilize the 'glmnet' R package to implement LASSO regression. Prior to fitting the LASSO regression model, it is critical to ensure that categorical variables are first converted into dummy variables, with a dummy variable corresponding to each level and one level included in the intercept. We utilize the dummyVars function in the 'caret' R package to construct the dummy variables.

In the LASSO regression model, we first subset only to respondents who have had sex and have had at least one partner in the past 12 months. This is because earlier data processing already logically deduced that respondents who have never had sex or had 0 partners in the past 12 months could not be active FSW. Therefore, the primary predictive objective is to predict status for respondents who could meet these two conditions, so the model is trained only on representative respondents.

We use 10-fold cross-validation to find the lambda that is one-standard error away from the cross-validated errors of the optimal lambda [1se.lambda], helping to increase sparsity in the covariates. Using a small number of informative predictors may be considered desirable in surveys with skip commands because it allows researchers to manually review the predictors to understand the missingness assumptions underlying the model. Figure 1 shows the optimal lambda as a green dotted line and the 1se.lambda as a black dotted line for one run of 10-fold cross-validation. We repeated this cross-validation 20 times, as opposed to increasing the number of folds, because active FSW are in small proportion compared to non-active FSW. This helps to ensure that the random folds have sufficient samples from both groups. Finally, we averaged the optimal 1se.lambda from each of 10 cross-validation procedures to yield the final lambda used in the full-data model.

Figure 1 – LASSO Regression Cross-Validation for Lambda



Once the final lambda is used in the full-data model, we keep only numerical variables that have a non-zero regression coefficient and categorical variables that have at least one level with a non-zero regression coefficient.

Step 4: Logistic Regression Model for the Response

Finally, we fit a logistic regression model using only the variables with non-zero coefficients after LASSO shrinkage. As described previously, the logistic regression model is constructed over the covariate data following MICE imputation. Finally, predictions for active FSW status are made for all respondents; predictions are subsequently compared to true status for respondents with known active FSW status.

Chapter 4

Results

Because the proportion of active FSW in the sample is approximately 1.8% of respondents with known status, that means our prediction model could achieve 98.2% accuracy by simply predicting that all respondents are not active FSW. Importantly, this sample statistic is not an estimate for the population proportion of active FSW because it does not consider the survey sample weighting. This sample statistic demonstrates that we cannot rely on accuracy to assess model performance, as it is sensitive to class imbalance. Instead, we compare the predicted values to the observed values for the subset of respondents with known status to yield the Receiver Operating Characteristic [ROC] Curve and corresponding Area Under the Curve [AUC], shown in Figure 2 constructed using the 'pROC' R package. A key advantage in referring to ROC AUC is that the metric is insensitive to class imbalance (Fawcett, 2006), meaning the true proportion of active FSW does not affect the ROC curve that is produced.

Figure 2 – ROC Curve for Prediction of Active FSW Status on Observed Respondents

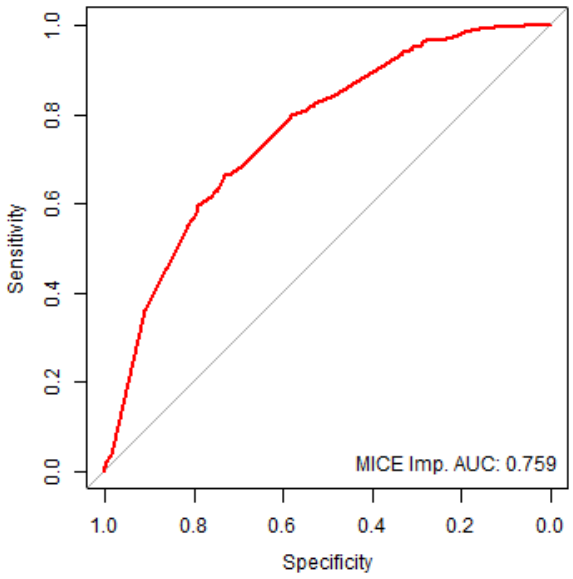


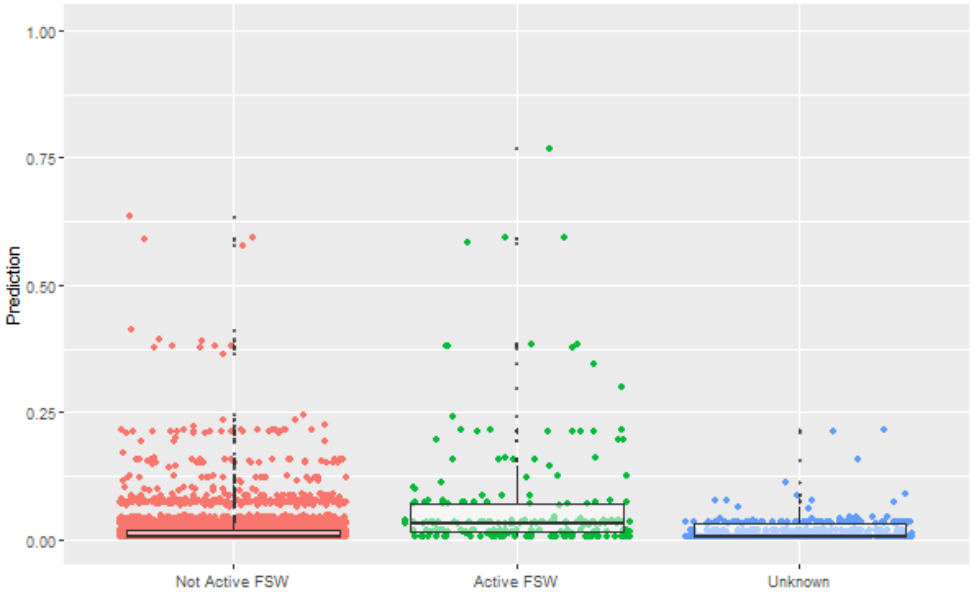
Figure 2 shows an ROC AUC of 0.759 after training the logistic regression model on the dataset with observed active FSW status, showing a notable improvement in the prediction model compared to a null guess, corresponding to the diagonal line with an ROC AUC of 0.5. Note that the mean 10-fold cross-validated ROC AUC was 0.750, which is supports that the overall model performance is not due to overfitting on the full dataset. There is no clear benchmark to assess the predictive performance of this model, in particular because a target ROC AUC inherently depends upon the predictability of the response. In this case, where the response relates to human behavior and the predictors are not all related to sexual activity, the model performance is expectedly lower than typical scientific benchmarks for ROC AUC, like 0.9.

An essential consideration of this approach is that the predictor variables cannot consist of questions that were used to deduce active FSW status. For example, if the variable ‘sellsxever’ was used to predict ‘sellsx12mo’, then the ROC AUC would be much closer to 1, as it is likely that someone who has sold sex in the past has particularly sold sex in the past 12 months. However, since the response was already deduced using all available information, this predictive information is particularly only available for respondents with known active FSW status. Therefore, the variable ‘sellsxever’ will offer almost no predictive information for respondents with unknown active FSW status. Researchers must be careful to ensure that the AUC ROC is not mistakenly inflated by ignoring the missing data issue that motivates the prediction model.

Figure 3, constructed using the ‘ggplot2’ and ‘cowplot’ R packages, shows the distribution of predictions among each of the three groups: active FSW, non-active FSW, and unknown status. Observe that the boxplot for the active FSW group shows a strong difference in distribution, and in particular, the median predicted probability. Observe that the unknown status

group has a third-quartile of predicted probabilities between the non-active FSW and active FSW groups; this suggests that the distribution of predicted probabilities is indeed a mixture of the other two distributions. However, the absence of large predicted probabilities in the unknown group suggests that perhaps the naïve methodology is unsuccessful in predicting truly missing data, if there are truly active FSW in the unknown sample.

Figure 3 – Distribution of Predictions of Active FSW Status for Observed and Unobserved Respondents



Chapter 5

Discussion

In the imputation procedure, a primary challenge is determining which variables are significantly associated with active FSW status among hundreds of variables. Shrinkage models, like LASSO regression, require a fully-specified predictor matrix, meaning there can be no missing values. However, imputation methods, like MICE, cannot handle hundreds of variables collected upon thousands of respondents—especially when imputations themselves may be illogical due to survey skip conditions under the survey design. So there is a balancing act: researchers want to subset to a small number of variables prior to using the MICE imputation procedure, but effective methods for subsetting variables are not possible until missing data are imputed. The p-values of marginal association, from either the Chi-Square test for independence or the ANOVA F statistic, give researchers some leeway to select an arbitrary number p of highly-associated predictors by proceeding with the variables corresponding to the p smallest p-values. Researchers may choose p based upon expected information gain and ad-hoc benchmarks that support convergence of the MICE imputation procedure. However, because our research made use of 2000 Monte Carlo simulations for the Chi-Square test, the corresponding p-values were bounded by a simulated minimum of $\frac{1}{2001}$, so such a cutoff could not be flexibly implemented. Moreover, the p-values inherently correspond to different test statistics, so their immediate comparison may be unwarranted. Nevertheless, the marginal association step is pivotal to the naïve methodology, as it allows researchers to eventually subset to a small number of variables among hundreds for the final prediction model.

Beyond a naïve methodology for imputing missing data, my research revealed that there is room for improvement in how to display the reason for missing data. Looking at the raw data, it is impossible to see—at a glance—why a respondent is missing a value for a particular variable. To guide in transparency for reasons of missingness, surveys with skip conditions might consider providing three layers of data encoding the survey: 1) the raw data, 2) the skip conditions corresponding to periods, and 3) the skip reasons corresponding to skip conditions. While these layers have some challenges of their own (for example, sometimes a skip condition is a complex logical evaluation of several conditions), this nonetheless would help researchers to understand the extent to which missing information could impact estimates.

Table 1 – Original Data

Raw Data

	V1	V2	V3	V4
Person 1	F	-9	.	.
Person 2	M	.	2	.
Person 3	M	.	0	.
Person 4	F	.	-8	A

Table 1 shows a toy demonstration of how data are currently presented, where each row represents respondent labels and each column represents variable labels. Then the raw data are presented in the matrix, with a period for skipped answers. Table 2 shows a demonstration of how missing data can be labeled according to a particular skip condition. For example, Person 1's missing value in Variable 2 is attributable to Condition 2, representing a skip condition that the survey placed on Variable 2.

Table 2 – Proposed Skip Condition Labels

Skip Condition				
	V1	V2	V3	V4
Person 1	.	C2	C2	C2
Person 2	.	C1	.	C4
Person 3	.	C1	.	C4
Person 4	.	.	C3	.

Then Table 3 shows that the reason the skip condition was activated is R2(C2). The Proposed Skip Condition Codebook shows that this means the respondent refused to answer Variable 2. Similarly, V3 and V4 were skipped under the same condition. The periods in this table represent non-missing information in the original data.

Table 3 – Proposed Skip Activation Labels

Skip Reason				
	V1	V2	V3	V4
Person 1	.	R2	R2	R2
Person 2	.	R1	.	R1
Person 3	.	R1	.	R3
Person 4	.	.	R1	.

Table 4 – Proposed Skip Condition Codebook

Skip Condition			Skip Reason		
Label	**Skip**	**If**	**Label**	**Skip trigger**	**Interpretation**
C1	V2	V1=M	R1	V1=M	Respondent is male
C2	V2,V3,V4	V2=-8|V2=-9	R1	V2=-8	Respondent did not know the answer to V2
C2	V2,V3,V4	V2=-8|V2=-9	R2	V2=-9	Respondent refused to answer V2
C3	V3	V3=-8|V3=-9	R1	V3=-8	Respondent did not know the answer to V3
C3	V3	V3=-8|V3=-9	R2	V3=-9	Respondent refused to answer V3
C4	V4	V1=M|V3=0	R1	V1=M	Respondent is male
C4	V4	V1=M|V3=0	R2	V3=0	The answer to V3 is 0
C4	V4	V1=M|V3=0	R3	V1=M&V3=0	The respondent is male and the answer to V3 is 0

A key advantage in using the Skip Condition Codebook is that the answers to sequentially asked questions can be immediately deduced. For example, suppose a respondent has a skipped value for the ‘sellsx12mo’ variable. If the Proposed Skip Condition Labels and

Proposed Skip Reason Labels reveal that the value was skipped due to the respondent having 0 partners in the past 12 months, then we can immediately deduce that the respondent is in the non-active FSW group. Contrarily, if the labels reveal that the value was skipped due to the respondent refusing to answer the number of partners in the past 12 months, then the respondent has unknown active FSW status. As long as the skip conditions are logical filters, then every variable in the survey can be immediately processed and analyzed, as opposed to needing to carefully read the questionnaire to understand which respondents were asked which questions. This also has the aforementioned advantage of allowing insight into the reasons for missing data, which the CONSORT diagrams do not currently offer.

There are also procedural advantages to this presentation. First, encoding the missing data skip conditions in a separate layer of data will allow existing research done on PHIA surveys to be replicated for future surveys, since the presentation of the original data will remain the same. Second, the logic for skip conditions is already performed in the Open Data Kit software through which the survey is administered, so storing the skip conditions would not require substantial changes to existing code, except to find which of several OR conditions are met.

One of the greatest procedural issues in the methodology presented above is in imputing skipped values among the covariates; based upon the reason the values were skipped, this imputation could introduce a logical inconsistency. One solution to this problem would be for the PHIA survey to process a Deduction-Complete Dataset [DCD], where all redundantly skipped variables are deduced as far as logically possible using the variables composing the skip conditions. For demonstration, consider the Sexual Activity Module. Respondents who have never had sex skip past all remaining questions in the module. Instead of labeling each of these skipped values as (“.”), they could instead be answered as if the respondent had been asked the

question. This would mean that ‘part12monum’ would be coded as 0 partners in the past 12 months, for example. This solution is not suitable for questions that are only logical to ask conditioning on a prior variable. For example, the variable ‘agemar’ has no valid value for respondents who have never been married. In this case, an alternative special character such as (“<”) could be used to signify non-applicability due to logical exclusion. Variables skipped due to an unknown response may be coded as (“?”) and values skipped due to a refused response may be coded as (“!”). With these data augmentations made, screening for significantly associated covariates would not have the possibility for logical inconsistency. Future work may consider DCD encodings in more detail to ensure compatibility with existing data mining methods.

Appendix A

Extract from MPHIA 2015-2016 Data Use Manual Supplement

Figure 4 – CONSORT Diagram of Analytic Variable 'sexever' from Page 110 of MPHIA Data Use Manual Supplement

Variable: sexever
Found in MPHIA 2015-2016 dataset:
Adult Interview

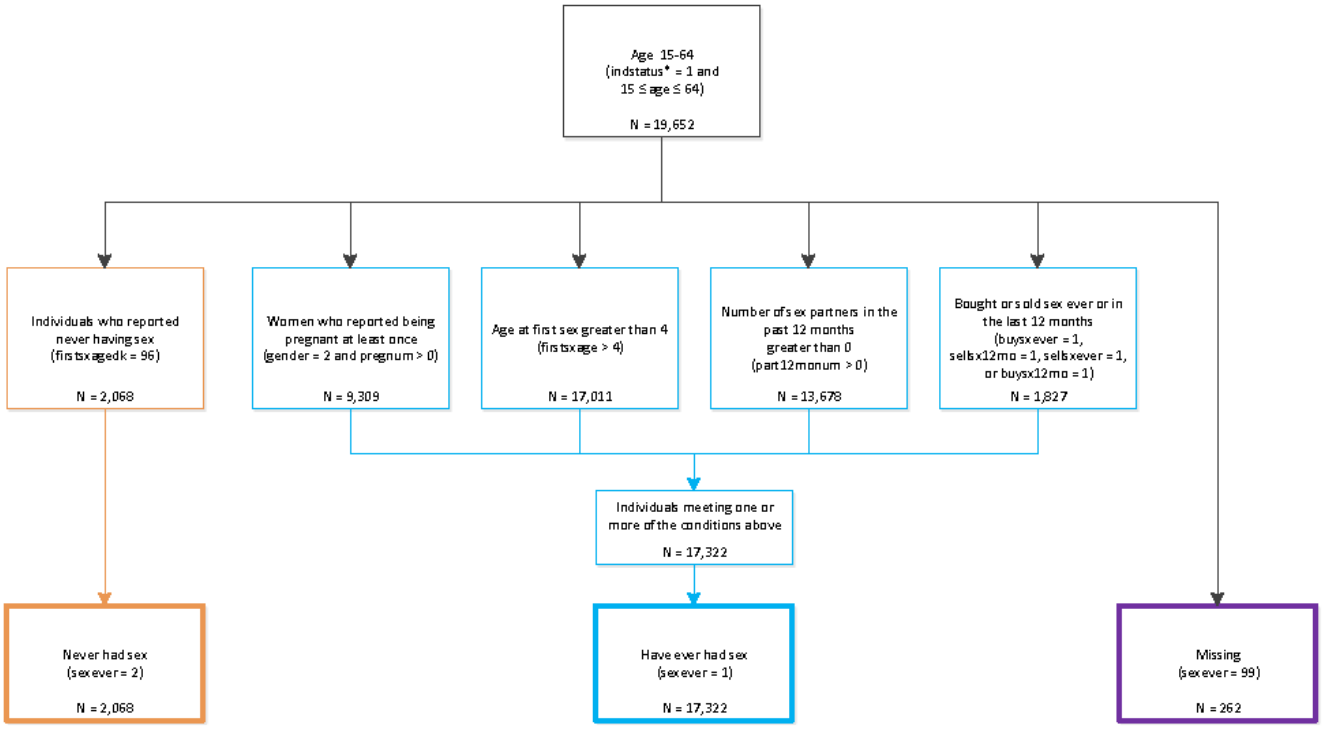
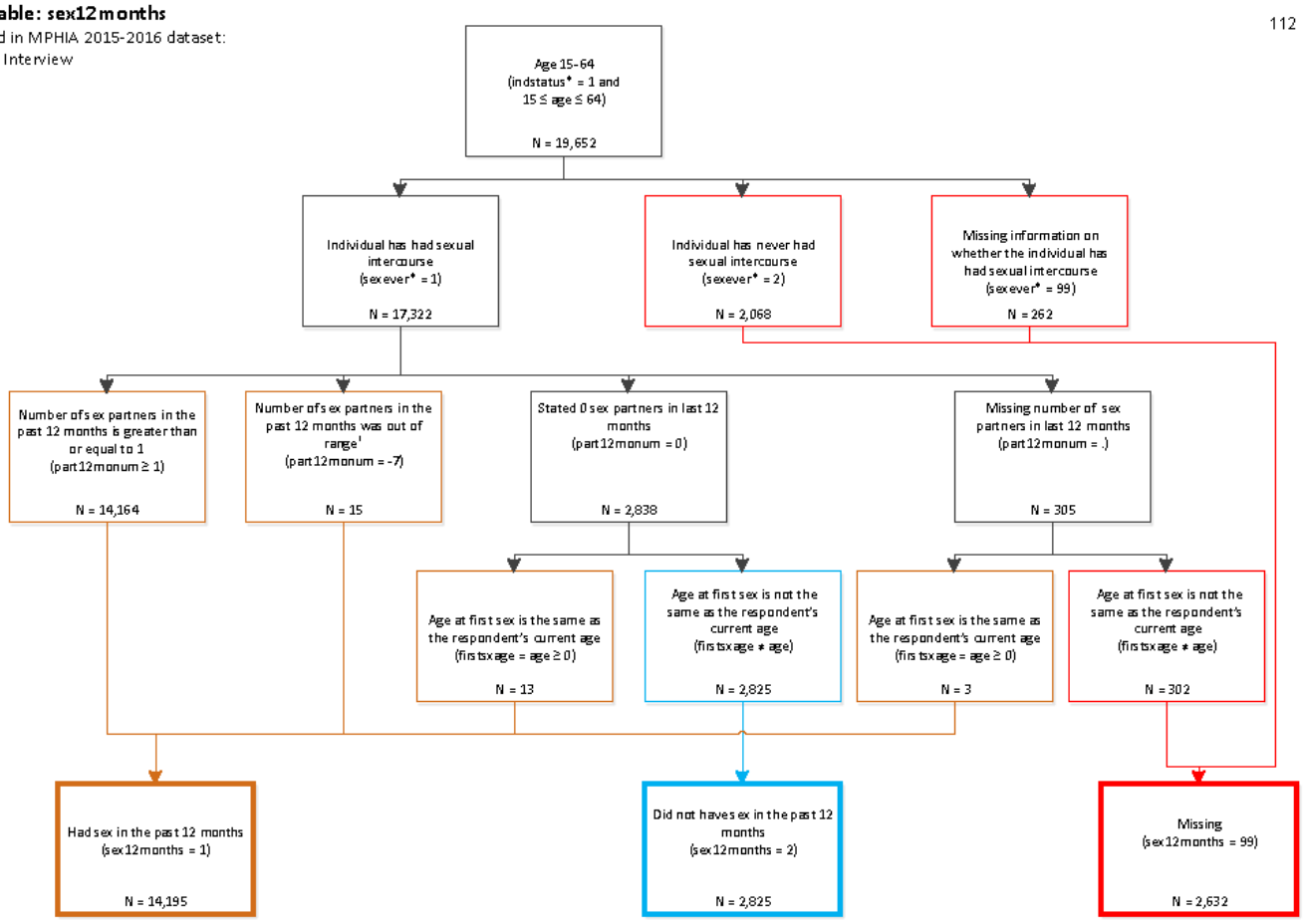


Figure 4 is Page 110 exactly extracted from the MPHIA Data Use Manual Supplement. Note that the variable 'firstsxage' corresponds to the vaginal intercourse only, so the 'sexever' variable is only an indicator for vaginal intercourse. This is an important note, as subsequent analytic variables are constructed upon this assumption.

Figure 5 – CONSORT Diagram of Analytic Variable 'sex12months' from Page 112 of MPHIA Data Use Manual Supplement

Variable: sex12months
Found in MPHIA 2015-2016 dataset:
Adult Interview



1. Number of sex partners in the past 12 months is out of range for women who had a child in the past 12 weeks and respondents who reported having bought or sold sex in the past 12 months that also reported having zero sex partners in the past 12 months.

Figure 5 is Page 112 exactly extracted from the MPHIA Data Use Manual Supplement.

Figure 6 – CONSORT Diagram of Analytic Variable 'paidsex12months' from Page 118 of MPHIA Data Use Manual Supplement

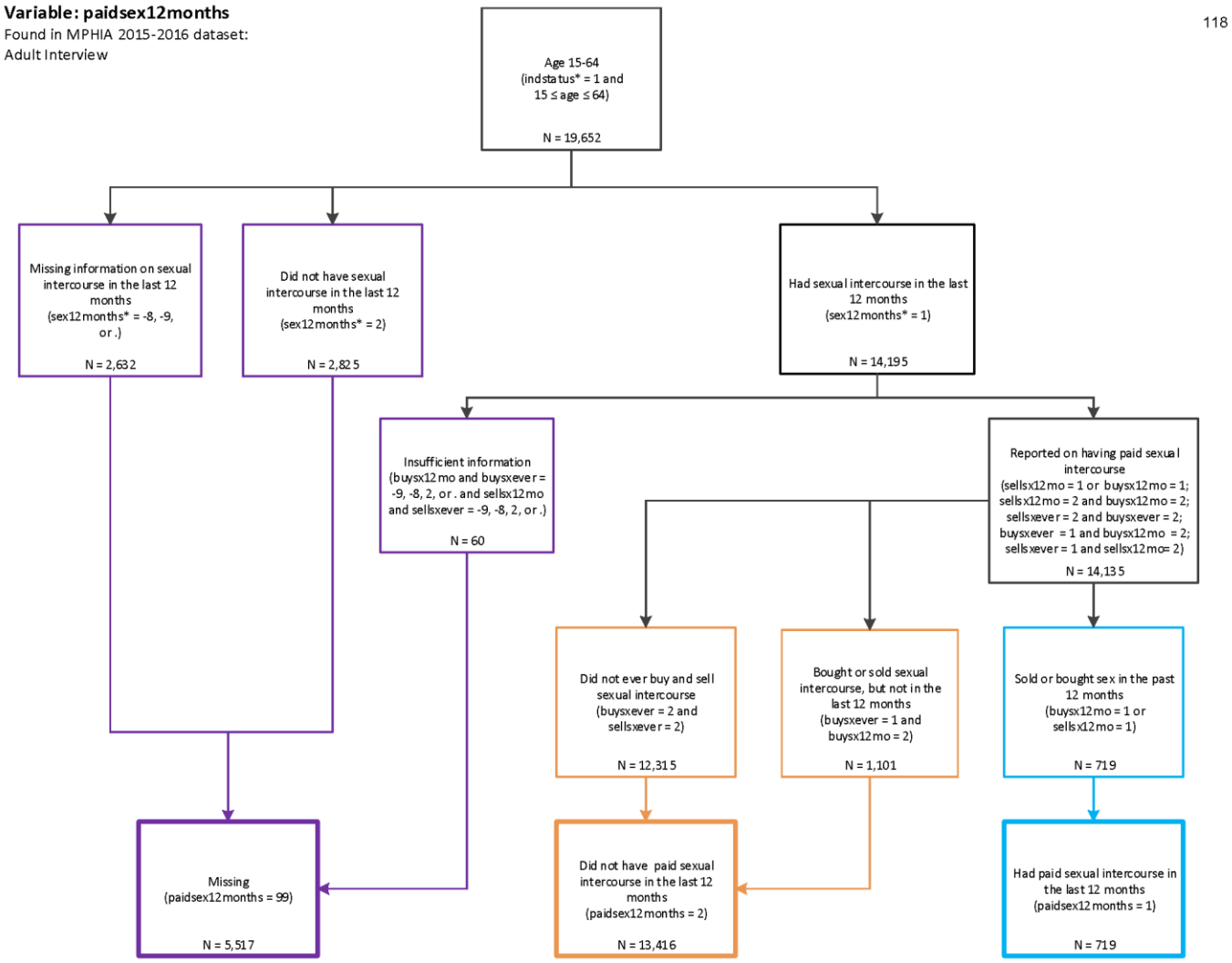


Figure 6 is Page 118 exactly extracted from the MPHIA Data Use Manual Supplement. This shows that CONSORT diagram does not display the reason for missing information in the indicator variable of interest. It also displays small differences in encodings, such as identifying respondents who have either sold or paid for sex in the past 12 months.

One issue to take note of is that people who have not sexual intercourse in the past 12 months are counted a missing, as opposed to “Did not have paid sexual intercourse in the past 12

months". This may be to ensure the variable only identifies sexually active individuals in these groups, however non-sexually active individuals should not be labeled as "Missing", as suggests that the missingness problem is larger than it truly is. Rather, if such respondents were labelled "Not sexually active", then researchers would still know to not mistakenly include them in the population of interest, but their true status would be well-described.

Appendix B

Extract from PHIA Data Use Manual

Figure 7 – Discription of Missing Data Types from Page 19 of PHIA Data Use Manual

3.3. Data management and cleaning

3.3.1. Missing data and other exceptions

PHIA surveys are administered using open data kit (ODK)-based software on electronic tablets, which enables forced responses. As a result, missing data for survey variables are minimal, except where participants explicitly responded "don't know" (generally coded as "-8", with some exceptions where "don't know" is a valid response), refused to answer ("-9"), or responses were out of range ("-7", e.g., when a woman who has been pregnant says she has never had sex), or where a question does not apply (".", e.g., number of prior pregnancies does not apply for men). Variables from check-all-that-apply questions are coded as character variables, with "don't know" and refusal responses coded as "Y" and "Z", respectively. Missing data for analytic variables (see *Section 3.6. Analytic variables*) are coded as "99" without distinguishing the reason for missingness (don't know, refused or not applicable). For biomarker data, missing values (".") indicate that participants were not tested for the biomarker.

Users should take care when conducting analyses to check for and determine appropriate treatment for missing responses. Consult each PHIA survey's ***Survey Questionnaires***, ***Codebook***, and ***Variable Frequencies*** for further information on each variable.

Users are also strongly advised to take caution and refer to each PHIA survey's supporting documentation when conducting analyses using variables pertaining to sexual partnerships. Data on sexual partnerships are captured in PHIA surveys as part of the Sexual Activity module of the Individual Interview. It should be noted that in some countries (e.g., Lesotho, Namibia, Swaziland, Uganda, and Zambia), the total number of unique sexual partners reported in the last 12 months (`part12mo`) may exceed the total number of reported lifetime sexual partners (`partlifetm`). Consult each PHIA survey's ***Survey Questionnaires*** and ***Codebook*** for specific information on each variable.

3.3.2. Age and date variables

Several age and date variables are provided in the PHIA datasets. Age variables include, but are not limited to, ages of participants and their children at the time of their interviews, and self-reported age at first sex and of recent sexual partners. Ages of all household members are first reported by the household head, which are subsequently confirmed by the individual if they participated in the individual interview. In cases of discrepancies, the participant's confirmed age takes precedence. Household member ages that are above a maximum age threshold (typically 80 years) are "top-coded" to the maximum age to preserve confidentiality.

Other age data may be cleaned, though this varies on a variable-by-variable basis. Mother's age, for example, is cleaned to missing (".") when mother and child ages are less than 10 years apart since this is considered to be biologically implausible (see *Section 3.5.3 Mother-to-child linking*). Conversely, ages of recent sexual partners are not cleaned; for example, if a participant reports that a recent sexual partner was >100 years old, this is considered unlikely but not implausible. PHIA data users are strongly advised to check distributions of variables to ensure appropriateness to the intended analytical purpose.

Date variables include prior pregnancies, HIV diagnosis/testing, and ART use, and birth year (birth day and month data are redacted to protect participant privacy) among others. Tablet-generated dates, such as start date for survey procedures at a given household, are retained as one date variable. However, dates based on responses to survey questions generally allow

Figure 7 is Page 19 exactly extracted from the PHIA Data Use Manual Supplement. This describes the types of missing data in the PHIA survey and how the missing data are coded. It also displays that the PHIA data is carefully screened for logical consistency in respondent answers, speaking to the high quality of the data.

Appendix C

Extract from MPHIA 2015-2016 Adult Questionnaire (Ministry of Health, 2018)

Figure 8 – Sexual Acivity Module from Pages 227 to 234 of MPHIA 2015-2016 Adult Interview Questionnaire

407	Who did the circumcision?	DOCTOR, CLINICAL OFFICER, OR NURSE = 1 TRADITIONAL PRACTITIONER / CIRCUMCISER =2 MIDWIFE = 3 OTHER = 96 DON'T KNOW = 8 REFUSED=9	

MODULE 5: SEXUAL ACTIVITY			
Interviewer says: "In this part of the interview, I will be asking questions about your sexual relationships and practices. These questions will help us have a better understanding of how they may affect your life and risk for HIV. **Let me assure you again that your answers are completely confidential and will not be shared with anyone. If there are questions that you do not want to answer, we can go to the next question."**			
501	If you wanted a condom, would it be easy for you to get one?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	YES, DK, REFUSED → 503
502	Why is it not easy for you to get a condom? SELECT ALL THAT APPLY.	CONDOMS NOT AVAILABLE/TOO FAR = A NOT CONVENIENT = B COSTS TOO MUCH = C EMBARASSED TO GET CONDOMS = D DO NOT WANT OTHERS TO KNOW = E DO NOT KNOW WHERE TO GET CONDOMS = F OTHER = X DON'T KNOW = Y REFUSED = Z	

MODULE 5: SEXUAL ACTIVITY			
503	How old were you when you had vaginal sex for the very first time? Vaginal sex is when a penis enters a vagina.	AGE IN YEARS _ _ NEVER HAD VAGINAL SEX = 96 DON'T KNOW = -8 REFUSED = -9	
504	People have sex in different ways. Some have vaginal sex. Some have anal sex. Anal sex is when a penis enters a person's anus. Have you ever had anal sex?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	NO, DK, REFUSED →506 NEVER VAGINAL OR ANAL SEX→NEXT MODULE
505	How old were you when you had anal sex for the very first time?	AGE IN YEARS _ _DON'T KNOW = -8 REFUSED = -9	
506	The first time you had vaginal or anal sex, was a condom used?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	
507	The first time you had vaginal or anal sex, was it because you wanted to or because you were forced to?	WANTED TO = 1 FORCED TO = 2 DON'T KNOW = 8 REFUSED = 9	WANTED, DK, REFUSED →509
508	The first time you had vaginal or anal sex, were you physically forced or were you pressured into having sex through harassment, threats or tricks?	PHYSICALLY FORCED = 1 PRESSURED = 2 DON'T KNOW = 8 REFUSED = 9	
509	People often have sex with different partners over their lifetime. In total, with how many different people have you had sex in the last 12 months? IF NONE CODE '00'.	NUMBER OF SEXUAL PARTNERS IN LAST 12 MONTHS __ __ __ DON'T KNOW = -98 REFUSED = -99	IF 00 PARTNERS IN LAST 12 MONTHS → 532

MODULE 5: SEXUAL ACTIVITY			
	IF NUMBER OF SEXUAL PARTNERS IS GREATER THAN 100, WRITE ' 100'.		

Interviewer says: "Now I would like to ask you some questions about the partners you have had sex with in the last 12 months. Let me assure you again that your answers are completely confidential and will not be told to anyone. I will first ask you about your most recent partner."

510	I would like to ask you for the initials of your partner so I can keep track. They do not have to be the actual initials of your partner.	INITIALS ____ ____	
511	Does (INITIALS) live in this household?	YES = 1 NO = 2 NO→513	
512	HOUSEHOLD LINE NO. for (INITIALS) CODE '00' IF NOT LISTED IN HOUSEHOLD ROSTER.	LINE NO_____	
513	What is your relationship with (INITIALS)?	HUSBAND/WIFE = 1 LIVE-IN PARTNER = 2 PARTNER, NOT LIVING WITH RESPONDENT = 3 EX-SPOUSE/PARTNER = 4 FRIEND/ACQUAINTANCE = 5 SEX WORKER = 6 SEX WORKER CLIENT =7 STRANGER = 8 OTHER = 96 DON'T KNOW = -8 REFUSED = -9	

514	How long has it been since you last had sex with (INITIALS)? IF LESS THAN ONE WEEK RECORD IN DAYS, IF LESS THAN ONE MONTH, RECORD IN WEEKS, OTHERWISE RECORD IN MONTHS.	DAYS _ _ WEEKS _ _ MONTHS _ _ DON'T KNOW = -8 REFUSED = -9	
515	How long has it been since you first had sex with (INITIALS)? IF LESS THAN ONE WEEK RECORD IN DAYS, IF LESS THAN ONE MONTH, RECORD IN WEEKS. IF LESS THAN ONE YEAR, RECORD IN MONTHS. OTHERWISE RECORD IN YEARS.	DAYS = _ _ WEEKS = _ _ MONTHS = _ _ YEARS = _ _ DON'T KNOW = -8 REFUSED = -9	
516	Is **(INITIALS)** male or female?	MALE = 1 FEMALE = 2 DON'T KNOW = 8 REFUSED = 9	
517	How old is **(INITIALS)**? Please give your best guess.	AGE IN YEARS ____ DON'T KNOW =98 REFUSED = -9	
518	The last time you had sex with (INITIALS) was a condom used?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	
519	The last time you had sex with (INITIALS) did either of you drink alcohol beforehand?	ONLY I WAS DRINKING = 1 ONLY PARTNER WAS DRINKING= 2 BOTH WERE DRINKING= 3 NEITHER = 4 DON'T KNOW = 8 REFUSED = 9	

520	In the last 12 months, how often did you use condoms with (INITIALS) when having vaginal sex? Was it always, most of the time, sometimes, rarely or never?	ALWAYS = 1 MOST OF THE TIME = 2 SOMETIMES = 3 RARELY = 4 NEVER = 5 NO VAGINAL SEX IN THE LAST 12 MONTHS = 3 DON'T KNOW = 8 REFUSED = 9 SKIP IF NEVER HAD VAGINAL SEX.	
521	In the last 12 months, how often did you use condoms with (INITIALS) when having anal sex? Was it always, most of the time, sometimes, rarely or never?	ALWAYS = 1 MOST OF THE TIME = 2 SOMETIMES = 3 RARELY = 4 NEVER = 5 NO ANAL SEX IN THE LAST 12 MONTHS = 6 DON'T KNOW = 8 REFUSED = 9 SKIP IF NEVER HAD ANAL SEX.	.
522	In the last 12 months, when you had sex with (INITIALS), did the condom you were using ever break, leak or slip off during sex or while pulling out?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9 SKIP IF NEVER USED CONDOM	
523	Did you enter into a sexual relationship with (INITIALS) because (INITIALS) provided you with or you expected that (INITIALS) would provide you with material support or help you in other ways? Material support means helping you to pay for things, or giving you	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9 NO, DK, REFUSED → 525,	

	gifts or other items you needed or requested.	SKIP IF SEX WORKER OR CLIENT	
524	In the last 12 months, what all did you receive? SELECT ALL THAT APPLY.	DID NOT RECEIVE ANYTHING = A MONEY = B FOOD = C SCHOOL FEES = D EMPLOYMENT = E GIFTS/FAVORS = F TRANSPORT = G SHELTER/RENT = H PROTECTION = I OTHER = X DON'T KNOW = Y REFUSED = Z SKIP IF SPOUSE, LIVE-IN PARTNER SEX WORKER OR CLIENT	
525	Was (INITIALS) circumcised?	YES =1 NO =2 DON'T KNOW = 8 REFUSED = 9 SKIP IF PARTNER NOT MALE.	.
526	Do you expect to have sex with (INITIALS) again?	YES =1 NO =2 DON'T KNOW = 8 REFUSED = 9	
527	Have you ever taken an HIV test with (INITIALS)?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9 YES, DK, REFUSED → 529	

528	What is the main reason you have never tested for HIV with (INITIALS) as a couple? READ RESPONSES ALOUD.	NOT A PARTNER/COUPLE= 1 NEVER DISCUSSED = 2 WE ARE NOT AT RISK FOR HIV = 3 PARTNER REFUSED = 4 I REFUSED = 5 WE KNOW OUR STATUS = 6 OTHER = 96 DON'T KNOW = -8 REFUSED = -9	
529	Does (INITIALS) know your HIV status? HIV status could mean you are HIV negative or HIV positive.	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	
530	What is the HIV status of (INITIALS)? READ RESPONSES ALOUD.	THINK (INITIALS) IS POSITIVE = 1 (INTIALS) TOLD ME HE/SHE IS POSITIVE = 2 POSITIVE, TESTED TOGETHER = 3 THINK (INITIALS) IS NEGATIVE = 4 (INITIALS) TOLD ME HE/SHE IS NEGATIVE = 5 NEGATIVE, TESTED TOGETHER=6 DON'T KNOW STATUS = 7 REFUSED = 9	
531	DOES THE RESPONDENT HAVE ANOTHER PARTNER IN THE LAST 12 MONTHS?	YES = 1 NO = 2 YES → RETURN TO 510 I will now ask about your second to last partner.	.

Interviewer says: "Now I am going to ask you some additional questions about your sexual activities. Again, I am asking that you answer these questions honestly. Let me assure you again that your answers are completely confidential and will not be shared with anyone."

532	Have you ever sold sex for money?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	NO, DK, REFUSED → 535
533	In the last 12 months, have you sold sex for money?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	NO, DK, REFUSED →535
534	The last time you sold sex for money, was a condom used?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	
535	Have you ever paid money for sex?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	NO, DK, REFUSED → NEXT MODULE
536	In the last 12-months, have you paid money for sex?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	NO, DK, REFUSED →NEXT MODULE
537	The last time you paid money for sex, was a condom used?	YES = 1 NO = 2 DON'T KNOW = 8 REFUSED = 9	

MODULE 6: HIV/AIDS KNOWLEDGE AND ATTITUDES			
Interviewer says: "Now I will ask you questions on your knowledge of HIV."			
601	Can the risk of HIV transmission be reduced by having sex with only one uninfected partner who has no other partners?	YES = 1 NO = 2 DON'T KNOW = 3 REFUSED = 9	

Figure 8 is Pages 227 to 234 exactly extracted from the MPHIA 2015-2016 Adult Interview Questionnaire.

BIBLIOGRAPHY

BBC News. (2021, May 12). Homosexuality: *The countries where it is illegal to be gay*. BBC News. https://www.bbc.com/news/world-43822234.

Berdikulov, D. (2019, April 3). *Types of missing data*. Medium. Retrieved from https://medium.com/@danberdov/types-of-missing-data-902120fa4248

Buuren, S. van. (2018). Flexible Imputation of Missing Data, Second Edition. Chapman & Hall/CRC.

ChartsBin statistics collector team 2010, *The Legal Status of Prostitution by Country*, ChartsBin.com, viewed 9th July, 2021, http://chartsbin.com/view/snb.

Deeks, S. G., Lewin, S. R., & Havlir, D. V. (2013). The end of AIDS: HIV infection as a chronic disease. *Lancet (London, England)*, 382(9903), 1525–1533. https://doi.org/10.1016/S0140-6736(13)61809-7

Fawcett, Tom. "An Introduction to ROC Analysis." *Pattern Recognition Letters*, vol. 27, no. 8, 2006, pp. 861–874., https://doi.org/10.1016/j.patrec.2005.10.010.

Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Science Review: Vol. 1 : No. 3 , Article 9. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/9.

Malawi Population-based HIV Impact Assessment (MPHIA) 2015-2016 Data Use Manual Supplement. New York, NY. Revised September 2021.

Mandavilli, Apoorva. "A Woman Is Cured of H.I.V. Using a Novel Treatment." *The New York Times*, The New York Times, 15 Feb. 2022, https://www.nytimes.com/2022/02/15/health/hiv-cure-cord-blood.html.

McHugh M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2), 143–149. https://doi.org/10.11613/bm.2013.018

Ministry of Health, Malawi. Malawi Population-Based HIV Impact Assessment (MPHIA) 2015-2016: Final Report. Lilongwe, Ministry of Health. October 2018.

Population-based HIV Impact Assessment (PHIA) Data Use Manual. New York, NY. April 2021.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." *BMC Bioinformatics*, 12, 77.

Sheng B., Li C., Bao L., Li R. (2021). "Probabilistic HIV Recency Classification -- A Logistic Regression without Individual Level Training Data"

UNAIDS. (2021, June 8). *New global pledge to end all inequalities faced by communities and people affected by HIV towards ending AIDS*. unaids.org. https://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2021/june/20210608_hlm-opens.

Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.

ACADEMIC VITA

GRANT T. HOPKINS

EDUCATION

The Pennsylvania State University | Schreyer Honors College — **University Park, PA**
Eberly College of Science | Master of Applied Statistics — *Class of 2022*
Eberly College of Science | B.S. in Statistics and B.S. in Mathematics
Smeal College of Business | B.S. in Finance

UNDERGRADUATE RESEARCH

Dr. Le Bao's Statistical Research Lab at Penn State University — **University Park, PA**
Undergraduate Research Assistant — *Jan 2021 – May 2022*

PROFESSIONAL WORK EXPERIENCE

Unisys — **Blue Bell, PA**
Corporate Accounting Intern — *Jun 2020 – Aug 2020*

Block Renovation — **Brooklyn, NY**
Contractor Operations Intern — *Jun 2019 – Aug 2019*

High School Work Experiences
Print & Marketing Associate at Staples, *Summer Lifeguard* at American Pool, *Food Runner* at Bluestone Country Club

LEADERSHIP & INVOLVEMENT

Presidential Leadership Academy — **University Park, PA**
Member of 2022 Cohort — *Apr 2019 – May 2022*

The Sapphire Leadership Academic Program — **University Park, PA**
Associate Sapphire Leader; Former Professional and Leadership Development Chair — *Jan 2019 – May 2022*

Alpha Kappa Psi Co-Ed Professional Business Fraternity — **University Park, PA**
Judicial Review Board Member; Former Bylaws Chair — *Jan 2019 – May 2022*

HONORS & SKILLS

Honors: Evan Pugh Scholar Senior Award, President Sparks Award, President Freshman Award, Robert W. Koehler Award, Schreyer Scholarship, Statistics Department Student Marshal, Finance Department Student Marshal, Phi Beta Kappa, Mu Sigma Rho

Skills: R, SAS, Python, Microsoft Office including Excel, LaTeX, Minitab, Mathematica, statistical analysis, communication