MEASURING AGREEMENT BETWEEN ELECTRONIC HEALTH RECORDS AND MEDICAID CLAIMS ACROSS RACE AND FEDERAL POVERTY LEVEL CATEGORIES, IN THE PRESENCE OF MISSING DATA

by

Emile Latour

A THESIS

Presented to the Department of Public Health & Preventive Medicine
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

June 2017

Department of Public Health & Preventive Medicine

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Emile Latour

has been approved

_____

Miguel Marino, PhD (Mentor/Advisor)

_____

Yiyi Chen, PhD (Chair/Committee Member)

_____

John Heintzman, MD, MPH (Committee Member)

TABLE OF CONTENTS

TABLE OF CONTENTS i
LIST OF TABLES iii
LIST OF FIGURES v
ACKNOWLEDGMENTS vi
ABSTRACT vii
Objective vii
Methods vii
Results vii
Discussion and conclusions viii
Keywords viii
INTRODUCTION 1
Electronic Health Records (EHR) 2
Kappa Statistics 3
Missing data, patterns and mechanisms 11
Approaches to missing data 16
Multiple Imputation 19
Multivariate imputation by Chained Equations (MICE) 25
Conclusion 31
METHODS 32
Background and data 32
Examine the missing data 34
Specification of the imputation model 40
Running the imputation models 48
Note on computation time 54
Assessing the imputations 56
Comparing scenarios 59
RESULTS 62
Cholesterol screening 64
Chlamydia screening 69
Colonoscopy 74
DISCUSSION 79
SUMMARY AND CONCLUSIONS 80

REFERENCES 82
ABBREVIATIONS 86
APPENDICES 87
Appendix A. Inspecting the data 87
Appendix B. Data Dictionary 90
Appendix C. Visual diagnostics to assess the imputations for all scenarios 91
Appendix D. Tables for Results 97

LIST OF TABLES

Table 1 – Scale for interpreting kappa 8

Table 2 – Pooling common statistics with Rubin's rules 25

Table 3 – Procedures in this analysis 34

Table 4 – Counts and frequencies of missing values 35

Table 5 – Counts and frequencies by demographic categories 39

Table 6 – Models selected for target variables 41

Table 7 – Variables in Full scenario imputation model 43

Table 8 – Variables selected by software 44

Table 9 – Variables included in Reduced Scenario 45

Table 10– Counts and frequencies of Ethnicity 51

Table 11 – Proportion of useable cases 52

Table 12 – Computation time for multiple imputations 55

Table 13 – Demographic characteristics of study sample 63

Table 14 – Cholesterol screening agreement results *prior to imputation* 66

Table 15 – Cholesterol screening agreement results *after imputation* 67

Table 16 – Chlamydia screening agreement results *prior to imputation* 71

Table 17 – Chlamydia screening agreement results *after imputation* 72

Table 18 – Colonoscopy agreement results *prior to imputation* 76

Table 19 – Colonoscopy agreement results *after imputation* 77

Table 20 – Cholesterol screening results by Race categories 98

Table 21 – Cholesterol screening results by FPL categories 99

Table 22 – Chlamydia screening results by Race categories 100

Table 23 – Chlamydia screening results by FPL categories 101

Table 24 – Colonoscopy results by Race categories 102

Table 25 – Colonoscopy results by FPL categories 103

LIST OF FIGURES

Figure 1 – Example of a 2x2 table 5

Figure 2 – Kappa illustration 7

Figure 3 – Example missing data patterns 13

Figure 4 – Stages of multiple imputation 19

Figure 5 – Missingness map 36

Figure 6 – Missing values histogram and pattern 38

Figure 7 – Convergence plot, Full Scenario, 3 target variables 49

Figure 8 – Convergence plot, Reduced Scenario, 3 target variables 50

Figure 9 – Convergence plot, QPM Scenario, 3 target variables 50

Figure 10 – Convergence plot, Full Scenario, 2 target variables 53

Figure 11 – Convergence plot, Reduced Scenario, 2 target variables 53

Figure 12 – Convergence plot, QPM Scenario, 2 target variables 54

Figure 13 – Distribution comparison: Race 57

Figure 14 – Distribution comparison: Ethnicity 58

Figure 15 – Distribution comparison: FPL 58

Figure 16 – Results from 6 scenarios by Race 60

Figure 17 – Results from 6 scenarios by FPL 61

Figure 18 – Cholesterol screening: Visualization of kappa statistics and 95% CIs 68

Figure 19 – Chlamydia screening: Visualization of kappa statistics and 95% CIs 73

Figure 20 – Colonoscopy: Visualization of kappa statistics and 95% CIs 78

ACKNOWLEDGMENTS

I would first like to thank my thesis advisor Dr. Miguel Marino for the continuous support of my thesis work. We both shared the idea that a thesis serves as an independent learning experience. To that end, he consistently allowed me to explore my curiosities and to make this paper my own, but steered me in the right direction whenever he thought I needed it. His patience and motivation are only exceeded by his kindness and love of statistics. I could not have imagined a better advisor and mentor for this work.

I would also like to thank the other members of my committee: Dr. Yiyi Chen for acting as my committee chair and for sharing her statistical knowledge; and Dr. John Heintzman for authoring the study on which this thesis builds and for his insightful clinical knowledge. Their expert knowledge and support were extremely valuable to this project.

Additional thanks goes to all the biostatics faculty, my fellow students, and colleagues at OHSU. The relationships that I have made have meant as much to me as all the statistics that I have learned.

Finally, special thanks to all of my family and friends for their unending support.

ABSTRACT

Objective

Our objective is to apply multiple imputation methods with statistics of agreement and electronic health records (EHR). These methods used together with this type of data is not well documented. By applying these methods in a novel way, we examine their potential for future use.

As a secondary objective, we assess for discrepancies in screening documentation between EHR and Medicaid claims data, across Race and Federal Poverty Level (FPL) categories which may reveal possible issues with data collection that would need to be addressed.

Methods

Using individual patient data from 43 Oregon community health centers for 13,101 Medicaid-insured adult patients, documentation for screening services were compared using kappa statistics, before and after imputing for missing data. Multivariate imputation by chained equations (MICE) was used to impute missing values due to its flexibility working with large and complex data sets.

Results

We successfully provide a practical example and guidance for combining MICE and kappa statistics. In this work, we determined no differences in documentation of screening services for groups based on Race and FPL.

Discussion and conclusions

We conclude that MICE is a beneficial tool when working with missing EHR data and when measuring agreement using kappa statistics. Though these methods are not well-documented in use together, by following the available literature, these methods were adapted and successfully applied to a non-standard statistic like Cohen's kappa. Ongoing work with this topic would be to examine the effects of transforming kappa prior to pooling. Also, sensitivity analysis would be important to assess the missingness mechanism.

Keywords

Missing data; multiple imputation; chained equations; fully conditional specification; electronic health records; EHR; kappa statistics; agreement.

INTRODUCTION

The inspiration for this thesis came from an interest to gain experience working with electronic health records (EHR) data and to learn about multiple imputation, a statistical technique for analyzing data sets where some values are missing. As a motivating example to investigate these topics, we will build upon the research of a 2014 study by Heintzman, et al. (the "original study") that assessed agreement of EHR with Medicaid claims data for documentation of 11 preventive care procedures in a population of continuously insured adult Medicaid recipients being served by a network of Oregon community health centers (CHCs) during 2011 [1]. The original study was interested agreement for the samples as a whole and utilized kappa statistics to remove agreement due to chance alone. For the purpose of this thesis, we selected three adult health procedures: Cholesterol screening, Chlamydia screening, and Colonoscopy, to assess agreement between the two data sources across demographic categories of Race and Federal Poverty Level. There is missing demographic data in both categories which may impact the stratified results. We have the opportunity here to investigate the potential impact and to compare agreement with the incomplete data and with complete data that has been multiply imputed.

With the Medicare Access and CHIP Reauthorization Act (MACRA) that was signed into law in 2015 and takes effect in 2019, there are changes coming in Medicaid reimbursements that shift the physician focus away from volume and towards better quality of care while avoiding unnecessary costs [2]. How quality is measured will need to be addressed. Consideration must be given to the source of data, EHR or claims data, for measuring quality and how good those sources are. It will also be important to ensure the same quality of care is being provided across demographics groups, and measured with metrics that perform the same in

different contexts. By examining agreement across demographic categories, Race and Federal poverty level (FPL), we can assess any systematic differences among patients to ensure the same quality of care for all.

Through this thesis, we seek to determine:

1. The viability of using multiple imputation techniques with EHR data and kappa statistics. The documentation of these statistical techniques used together is limited to non-existing. Most literature on multiple imputation use statistical models as a basis for discussion. Whereas, multiple imputation for non-standard statistics, like kappa statistics, are not fully developed. By our example, we hope to provide guidance and reference for others working with large electronic health databases to apply these methods.
2. If there are discrepancies in agreement between EHR and Medicaid claims among Race categories and Federal Poverty Level categories. Apparent systematic differences would be cause for concern on a societal level as well as a public health concern.

In the Introduction, we hope to provide a survey of the topics covered here: EHR, kappa statistics, and multiple imputation, and provide sufficient explanation to highlight the topics that shaped the later analysis.

Electronic Health Records (EHR)

Electronic health records (EHR) are growing in use as a data source of healthcare services reporting, for both regulatory and reimbursement purposes. In the past, insurance claims

data have been considered to be the most accurate source of reporting data, but there are many flaws inherent to the claims data. There can be issues and time involved to obtain the information; the data often requires significant work in order to be cleaned and suitable for analysis; and this data work can prove to be costly. By comparison, EHR have desirable advantages of being easier to obtain and in a cleaner format and, thus, less costly. In addition to these advantages, EHR include additional information on unpaid services, services to uninsured person, and those with varied payers. There is potential for EHR to be used in place of or as a proxy data source for insurance claims data. Though EHR have been validated with Medicaid claims data for certain diabetes services, the work is growing to compare the documentation of preventive services [1].

Kappa Statistics

In this thesis, we are interested in the reliability of EHR data to serve as either a proxy or replacement for Medicaid data records. In assessing reliability, we are asking how well do these two sources of data agree with each other. The kappa statistic (or kappa coefficient or kappa), first proposed by Cohen (1960), is the most widely used statistic to measure agreement between two or more observers or "raters".

When discussing reliability and agreement, it should be noted that there are two types that exist and that the kappa coefficient could be used for both. The first, intrarater reliability, assesses agreement between the ratings of the same observer on two or more occasions, i.e. multiple ratings by the same rater. Second, interrater reliability assesses agreement between ratings made by two or more observers. From this distinction, we can see that here we are using the kappa statistic to assess interrater reliability.

The simplest situation where the kappa coefficient can be used is when two observers each provide one rating of the same subject. Kappa can be used when there are more than two possible ratings, i.e. more than one rating per subject by each of the observers. The weighted kappa would be applicable when there are more than two possible ratings and size of the discrepancy between raters matters. Fleiss (1971) has given methods for when more than two raters may rate each subject or when each observer may not rate each subject [3]. In this thesis, for each preventive service, stratified or not, we will be concerned with the simple case where 2 raters (EHR and Medicaid/Claims) give an independent single rating (e.g. received preventive screening or not) for each subject.

To compare the two sources of data and to assess their agreement in the documentation of preventive services, we can summarize the counts of the *n* subjects classified by the two data sources (the two raters) using a 2×2 contingency table. This is customary way to show binary ratings, such as yes or no, by two different raters, i.e. two independent observers are evaluating the same thing. One data source is assigned to the rows and one to the columns. Each of the four cells will represent a specific value for each of the two data sources. An example using Claims data and EHR data sources as the two raters can be seen in Figure 1 where the entries in the table refer to the *number* of subjects.

Figure 1 – Example of a 2x2 table
Example of 2×2 contingency table showing the counts of subjects where the two raters are the EHR data and the Medicaid Claims data; and the ratings for each are binary yes/no response whether a preventive service was recorded in a particular data file. The cells along the main diagonal (*a* and *d*) show where the two raters agree and the off-diagonal cells (*b* and *c*)

	Claims data		
EHR data	Yes	No	Total
Yes	a	b	m_1
No	c	d	m_0
Total	n_1	n_0	n

In Figure 1, cells *a* and *d* show the number of subjects for which the ratings in both the EHR data and the Medicaid claims data are in agreement. Cells *b* and *c* show the number of subjects where the two data sources disagree. The marginal row totals, m_1 and m_0, show the number of subjects rated "Yes" and "No", respectively, for a preventive service in the EHR data only. The marginal column totals, n_1 and n_0, show the number of subjects rated in the Claims data as "Yes" and "No", respectively. *n* is the total number of subjects deemed eligible for a particular preventive service. The 2×2 contingency table of count information can be converted to show the proportions or frequencies by dividing each entry by the total number of eligible subjects, *n*.

The simplest index of agreement that can be determined from this table is the overall proportion of agreement

$$p_O = \frac{a+d}{n}$$

sometimes called the proportion of observed agreement. Note that the values used in the observed agreement come from the concordant cells along the diagonal of the 2×2 table. The interpretation of observed agreement is the proportion of all eligible subjects where the two

data sets agree, i.e. proportion of subjects recorded as "Yes" in both data sources or "No" in both data sources. If there were complete agreement between the two data sets, the cells b and c would be zero and the observed agreement (p_O) would be 1. Conversely, if the two data sets are in complete disagreement, then cells a and d would be zero and p_O would be 0.

Many other indices of agreement have been proposed, but without the realization that, except in extreme circumstances ($m_1 = n_0 = 0$ or $m_0 = n_1 = 0$), there will be some agreement between raters that is due to chance alone [4]. It is important to separate the degree to which the raters agree purely by chance, since, in this case, they would not really be in agreement. Cohen's kappa was introduced as a measure of agreement that could be considered a measure of "true" agreement by adjusting the observed proportional agreement to take account of the amount of agreement which would be expected by chance. To do this adjusting, one must determine the proportion of agreement that is due to chance. This determination is based on the assumption that the assessments are independent between the two raters, i.e. the probability of a subject being in a certain row is independent of the column that they appear in. From the 2×2 table, this proportion of expected agreement is calculated by multiplying the marginal totals that correspond to each cell along the main diagonal, each divided by n

$$p_E = \frac{n_1}{n} \cdot \frac{m_1}{n} + \frac{n_0}{n} \cdot \frac{m_0}{n}$$

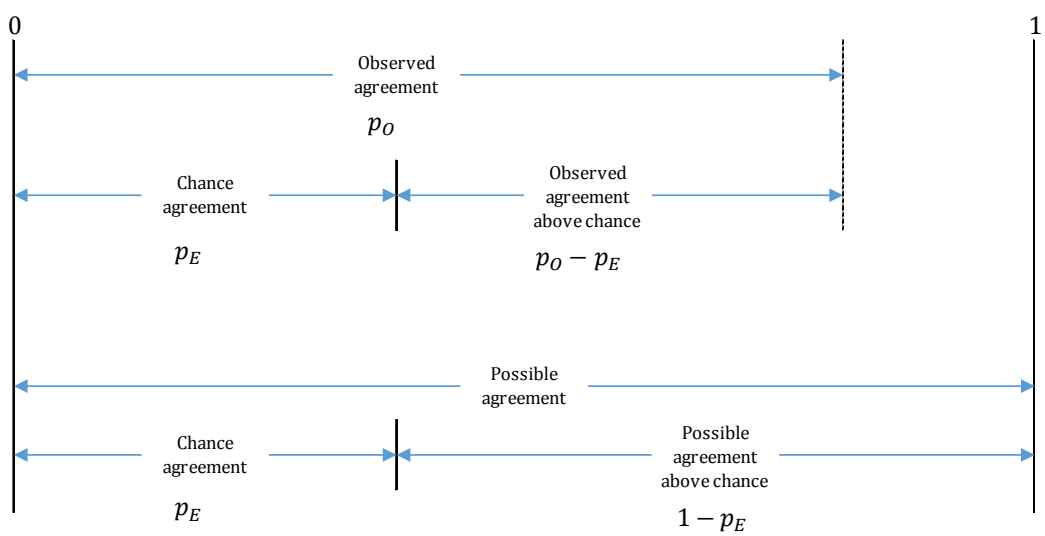
Using the proportion of observed agreement (p_O) and the proportion of expected agreement (p_E), Cohen stated the formula for the kappa statistic as

$$\kappa = \frac{p_O - p_E}{1 - p_E}$$

which he interpreted as "the proportion of agreement *after* chance agreement is removed from consideration" [5].

Figure 2 below shows visually how the components of the kappa statistic reflect chance-adjusted agreement. In the upper half of Figure 2, the proportion of observed agreement is shown to take on values between zero and one, and that chance agreement will be less than observed. The numerator $p_O - p_E$ is the amount of observed agreement above agreement expected by chance. Then in the lower half of Figure 2, we see how the maximum possible observed agreement (1.0) is adjusted for the proportion of agreement expected due to chance to give the denominator, $1 - p_E$. The figure makes it clear that the kappa coefficient is "the observed agreement, corrected for chance, as a fraction of the maximum obtainable agreement, also corrected for chance" [6].

Figure 2 – Kappa illustration
Illustration showing the numerator of kappa as observed agreement adjusted for chance, and the denominator as possible agreement adjusted for chance. Adapted from Rigby and from Sim and Wright [6], [7].



Possible values for kappa range from −1 to 1. Perfect agreement ($p_O = 1$), where the raters agreed in each classification, is indicated by $\kappa = 1$. Observed agreement that is no better than that expected by chance ($p_O = p_E$) is indicated by $\kappa = 0$. In the instance of perfect disagreement ($p_O = 0$), then the lowest possible value of for kappa is $-p_E/1 - p_E$. So depending on the value of p_E, kappa may take on a negative value on the range [−1, 0). Negative kappa would indicate agreement worse than that expected by chance and rarely occurs in clinical contexts [7].

Even though kappa provides an indication of the size and direction of agreement, there are many interpretations of what qualifies as "good" agreement. Thought there is no formal scale, the most widely used standards to assess kappa statistics were originally proposed by Landis and Koch, see Table 1 [8]. They do acknowledge that, though the divisions are arbitrary, it is useful to have benchmarks for discussion.

Table 1 – Scale for interpreting kappa
Interpretation of strength of agreement indicated by the kappa statistic. From Landis and Koch [8].

Kappa statistic	Strength of agreement
< 0.00	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

The original study condensed the cutoff points shown in Table 1 to consider: > 0.60 substantial agreement, 0.41–0.60 moderate agreement, 0.21–0.40 fair agreement [1]. This interpretation is the same as what was proposed by Landis and Koch except that any kappa

greater than 0.60 is considered "substantial"; separate classification of "almost perfect" was not a priority. Since this thesis builds upon the work of the original study, their interpretation will be used. It's important to note that kappa is dependent on the prevalence of a condition, and that care should be taken when comparing kappa values from different studies where the prevalence varies [9].

It won't be examined or discussed in this thesis, but it should be mentioned that there is a hypothesis test where under the null kappa = 0. It does not test the strength of agreement, but only whether agreement is due to chance. Kappa is used to give a quantitative measure of the magnitude of agreement between observers [10]. Thus, the interest lies in strength of agreement not whether or not it is due to chance.

There is debate about interpreting agreement between raters solely based on kappa. Feinstein and Cicchetti discuss the "paradoxes" that can occur and detail the issues that can arise due to prevalence and marginal totals [11]. Their recommendation, in addition to kappa, is to report an index of average positive agreement and an index of average negative agreement (commonly just called proportion of positive and proportion of negative agreement) [12]: $p_{pos} = \frac{2a}{2a+b+c}$ and $p_{neg} = \frac{2d}{2d+b+c}$. Proportion of positive agreement, for example, estimates the conditional probability, given that a randomly selected rater makes a positive rating, then the other will do so also. These indices are closely analogous to the sensitivity and specificity ($sensitivity = \frac{a}{a+c}$ and $specificity = \frac{d}{d+b}$) which are commonly seen and trusted in diagnostic testing, where one rater is considered the "gold standard".

Byrt et al. also sought to resolve the paradoxes associated with kappa and the effect of prevalence on its calculation [9]. They proposed a prevalence index ($PI = \frac{|a-d|}{n}$) to assess the difference between proportions of positive rating and negative rating. The argument is that magnitude of kappa is affected by prevalence, and so kappa must be interpreted taking PI into account. Also, a bias index ($BI = \frac{|(a+b)-(a+c)|}{n} = \frac{|b-c|}{n}$) is discussed that assess the extent to which raters disagree on the proportion of positive or negative cases. If substantial bias exists then it needs to be investigated and an index of agreement may not be appropriate. To adjust kappa for the imbalances caused by these measures, Byrt et al. provide the prevalence and bias adjusted kappa ($PABAK = 2\left[\frac{a+d}{n}\right] - 1$). Some have been critical of $PABAK$ [13] [14], but it has advantages to reporting a single summary measure rather than kappa and several other prevalence and bias indices.

Though there are a range of indices available and some disagreement on which to report. Current literature tend to agree that additional indices provide more information relevant to understanding and improving the interpretation of agreement than if kappa was reported alone [11] [12] [9] [13] [14]. However, we focused this thesis on the kappa statistic because it is commonly used and preferred in the medical literature.

Missing data, patterns and mechanisms

Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data.

Allison (2002) [15]

Missing data are defined as values that are not available and that would be meaningful for analysis had they been observed [16]. Ignoring missing data or editing may make the data seem complete, but it may lead to problems such as:

- Inefficiency – loss of information leading to loss of power,
- Systematic difference – leading to biased results, and
- Unreliable results.

When working with missing data methods, one must first consider both the missing data pattern and the missing data mechanism. The missing data pattern describes which values are observed in the data matrix and which values are missing. The missing data mechanism describes the relationship between missingness and the values of the variables in the data matrix.

Many different patterns can arise in the missingness in a given data set. Some missing data methods will work for general patterns and other methods will apply only to special patterns. Missing data patterns will influence the amount of information that can be transferred between variables. For this thesis and discussion, it important to mention the types that Van Buuren highlights for theoretical and practical reasons [17]:

1. Univariate and multivariate – If only one variable in a data set has missing values, the missing data pattern is univariate. More than one variable with missing values, multivariate.
2. Monotone and non-monotone (or general) – Monotone patterns occur when the missing data can be arranged by observations and variables so that there is a sequential order to the number of missing values by variable. That is, if variable V_j has missing values, the all variables V_k with $k > j$ are also missing. This occurs in longitudinal studies with drop-out occurring; once a subject is missing values they are missing subsequent values. Non-monotone (or general) pattern is not monotone and appears arbitrary.
3. Connected and unconnected – If any observed data point can be reached from any other observed data point through a series of horizontal or vertical moves. Connected patterns are needed in order to identify unknown parameters [17].

Figure 3 – Example missing data patterns
An interpretation of a visual aid provided by Van Buuren [17] that illustrates the differences in missing data patterns listed above.

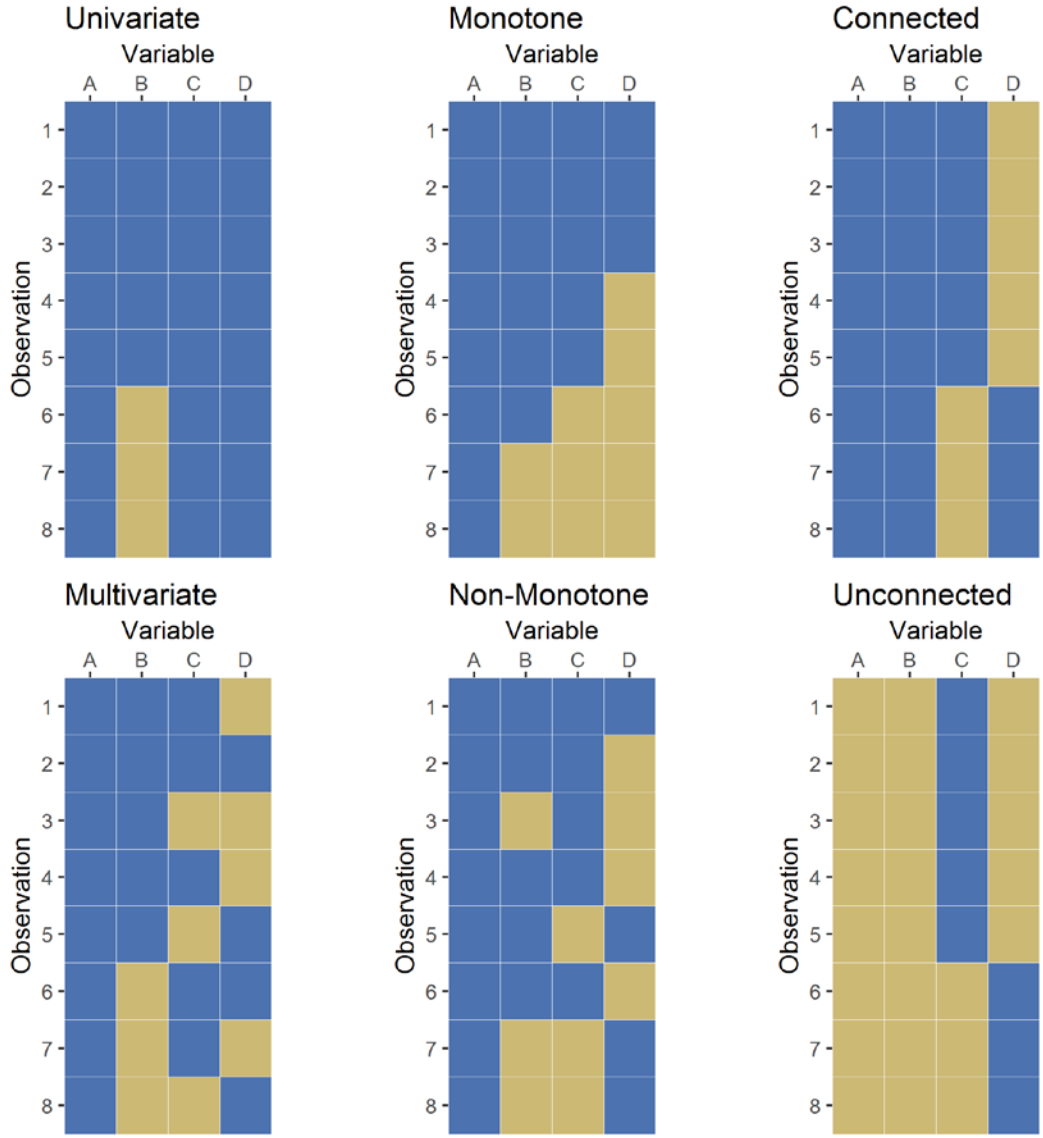


Figure 3 above provides example missing data patterns. The comparative difference between univariate and multivariate should be clear. The monotone example shows how successive variables have more missing values than the one preceding; non-monotone shows a more random looking pattern than sequential. The connected example above is also what is known as *file matching*: an attempt can be made to fill in missing values for variables 3 and 4 by

matching on the basis of variable 1 or 2 and imputing value from the matching units. The unconnected example also shows a case where two variables are never observed jointly; so parameters related to the association between such variables may not be estimable from the data and may lead to misleading results [17] [18].

The mechanism that leads to missing data is a different issue. The mechanism examines the "reason" for missing values and tries to determine whether variables that are missing are related to the underlying values of the variables in the data set.

Rubin (1976) classified missing data mechanisms into three categories [19] [18]. Let Y represent a $n \times p$ matrix with data for n observations (rows) and p variables (columns). To characterize the nature of the missing data, define a response matrix, R, as a $n \times p$ matrix of $0 - 1$ values. The elements of the matrices Y and R are denoted as y_{ij} and r_{ij} where $i = 1, \dots, n$ and $j = 1, \dots, p$. Acting as an indicator for missingness:

$$r_{ij} = \begin{cases} 1, & if\ y_{ij}\ is\ observed \\ 0, & if\ y_{ij}\ is\ missing \end{cases}$$

Note that R is completely observed in the sample. The complete data values Y are made in two parts Y^{obs} and Y^{mis}, where Y^{obs} represents the collective elements in Y that are observed ($r_{ij} = 1$) and Y^{obs} represents the collective elements of Y that are missing ($r_{ij} = 0$). In addressing the problem that Y is observed incompletely, the key assumption is the nature of the mechanism that generates R. Let ψ contain the unknown parameters of the missingness mechanism, then the general expression of the missing data model is

$$Pr(R|Y, \psi) = Pr\left(R \middle| Y^{obs}, Y^{mis}, \psi\right)$$

The data are said to be missing completely at random (MCAR) if no information in Y can predict whether the data are missing or not. The causes of the missing data are unrelated to the data. So data are MCAR if

$$Pr(R = 0|Y, \psi) = Pr(R = 0|\psi)$$

Missing data under MCAR are essentially a random sample of Y, and the probability of being missing is the same for all cases. Though it has convenient properties, it is often an unrealistic assumption for real data. MCAR is the only mechanism that can be tested through a multivariate test developed by Little (1988). Most simple fixes for missing data only work under the MCAR assumption which is often unrealistic and may provide biased results.

The data are said to be missing at random (MAR) if the probability that observations are missing may depend on Y^{obs} but not Y^{mis}. So data are MAR if

$$Pr(R = 0|Y, \psi) = Pr\left(R = 0\middle|Y^{obs}, \psi\right)$$

MAR allows for the possibility that probability of missingness can be predicted from other available, observed data. Van Buuren states that data are MAR if the probability of missingness is the same only within groups defined by the observed data [17]. Some known aspect of the data may influence whether it is missing or observed which does not fit MCAR. However, if the groups are known and MCAR can be assumed within the groups, then the data are MAR.

If data are not MCAR nor are they MAR, then they are classified as missing not at random (MNAR). Some use the label NMAR (for not missing at random), but there we will continue with the same notation used by Van Buuren, MNAR. Here, the probability of missingness varies for reasons that are unknown. This case does not simplify and data are MNAR if

$$Pr(R = 0|Y, \psi)$$

This ends up being the most complex situation to handle. More data must be collected to determine the cause of missingness or scenario analysis can be performed [17] [18] [20].

Note that in each of the mechanisms for missing data there are the unknown parameters for the missing data model, ψ. From Van Burren, in practice, it is important to distinguish between the missingness mechanisms (MCAR, MAR, and MNAR) since it clarifies the conditions where the parameters of interest can be accurately estimated without the need to know ψ [17]. The missing data model is considered "ignorable" if the MCAR or MAR assumption holds (discussed more later in assumptions for multiple imputation). This idea of ignorability is important for imputation since it implies that the distribution of the data Y is the same in the response and non-response groups

$$Pr\left(Y \middle| Y^{obs}, R = 1\right) = Pr\left(Y \middle| Y^{obs}, R = 0\right)$$

When the missing data model is ignorable, one can model the posterior distribution $Pr\left(Y \middle| Y^{obs}, R = 1\right)$ from the observed data and use this model to impute the missing data. For some cases, though, MAR may not be plausible or realistic. An analyst must take it on faith that the observed data are sufficient to correct for the effects of the missing data; this can only be tested against separate validation data. Strategies do exist for determining when the data is nonignorable; the two strategies mentioned by Van Burren are (1) to expand the data and assume ignorability on the expanded data set and (2) formulate a model for the non-response groups different from that of the response [17] [21].

Approaches to missing data

Missing data most often is traditionally handled by deleting cases with missing values. Complete-case analysis (aka listwise deletion) is a default way to handle missing data in many

software packages and confines the data to cases where all variables are present [17]. While a convenient approach to implement, it relies upon the assumption that the data are MCAR and can results in biased estimates and a reduction in statistical power [22]. Pairwise deletion (aka available-case analysis) is typically used when working with a correlation matrix and it tries to fix the data loss problems with listwise deletion. Using the correlation matrix to illustrate this method, all data are taken into account for each *pair* of variables for which data is observed. Though good idea to try to use all available information, the estimates are based on different subsets of cases and may result in biased estimates [22].

Single imputation methods are considered to be an improvement, but they do not reflect the uncertainty in the imputations [23]. Specific details are not examined here, but some common methods include: mean imputation, regression imputation, stochastics regression imputation, hot-deck, cold-deck, last observation carried forward, and baseline observation carried forward [17]. These methods tend to lead to standard errors that are too small and have potential for incorrect conclusions [17] [23].

A method popular in public health and epidemiology is known as the indicator method where a dummy variable (0/1) is included in the statistical model to indicate whether the value for a variables is missing or observed [24]. The advantage is that the method retains the full data set and reduces the loss of statistical power. The method may be suitable for some special cases, but these conditions are difficult to achieve in practice making it a less than optimal general approach [17].

Likelihood-based methods, often referred to as full-information maximum likelihood (FIML) methods, define a model for observed data, and so there is no need to impute missing data or

to omit incomplete cases [17]. FIML assumes that the data are MAR and multivariate normal joint distribution for all the variables [25]. FIML is considered to be more efficient than the methods mentioned above [26]. These methods are only available for certain models, such as longitudinal or structural equations models, and generally require specialized software [23].

Weighting methods, such as inverse probability weighting (IPW), can reduce bias when the probability of being selected in a survey differs between respondents [17] [27]. These methods can be used when individuals vary in the probability of having missing information. It is a relatively simple method to apply and can be useful in certain circumstances.

When available, obtaining alternative sources of information can be an appropriate solution to missing data problems. Prevention of missing data in the first place is the most direct way to avoid problems with missing data. Advice and strategies exist in a variety of sources, but still missing data issues persist.

The field of missing data is extensive. We do not intend to cover all the methods and approaches in depth. This section serves to mention just some of the approaches that are available before moving on to multiple imputation, the missing data method that is the focus of this thesis.

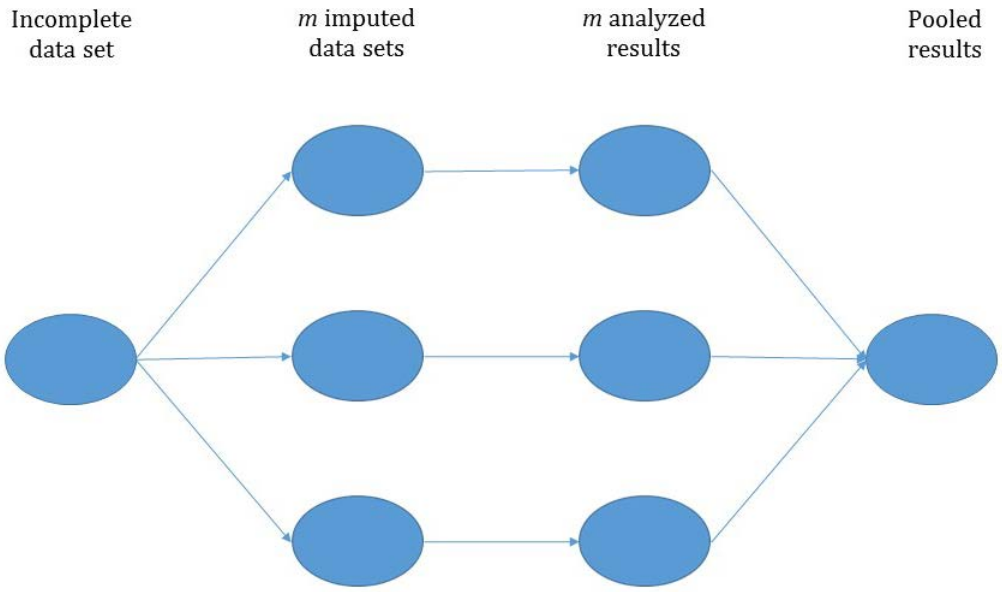
Multiple Imputation

The goal of multiple imputation is to obtain statistically valid inferences from incomplete data.

Van Buuren (2012) [17]

Multiple imputation (MI) is a statistical method for analyzing data sets with missing values. Figure 4 shows the main three stages involved in multiple imputation: generating *m* multiply imputed data sets, analyzing the *m* imputed sets, and pooling the results from the *m* analyses [17]. Starting with observed, incomplete data, MI creates several versions of the data by replacing missing values with plausible ones. The parameters of interest are estimated for each of the imputed data sets. Then the multiple parameter estimates are pooled into one estimate.

Figure 4 – Stages of multiple imputation
The three stages of multiple imputation [17]



Generating m multiply imputed data sets

Figure 4 starts from the left with the observed, incomplete data set (Y^{obs}). Using multiple imputation, *m* complete data sets are created by replacing missing values with plausible ones. It's assumed that the intended data, $Y = (Y^{obs}, Y^{mis})$, follow a distribution $p(Y|\theta)$, where θ is the collection of all the parameters of the model. Then also assuming that the data are MAR, these imputed, plausible values are drawn from the posterior predictive distribution of the Y^{mis} given Y^{obs}, which can be written as

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|Y^{obs}, \theta)p(\theta|Y^{obs})d\theta$$

where $p(Y^{mis}|Y^{obs}, \theta)$ is the conditional predictive distribution of Y^{mis} given Y^{obs} and θ, and where $p(\theta|Y^{obs})$ is the posterior distribution of θ based on the observed data Y^{obs}. This posterior predictive distribution can rarely be expressed in a closed form due to the integration, and it is difficult to draw samples from the distribution directly [28]. A single imputation of Y^{mis} can be produced by

1. Calculating the posterior distribution $p(\theta|Y^{obs})$ of θ based on the observed data Y^{obs},
2. Simulating a random draw of $\tilde{\theta}$ from the observed-data posterior distribution, $p(\theta|Y^{obs})$.
3. Then randomly drawing a value for each element of Y_{mis} from the conditional predictive distribution, $p(Y^{mis}|Y^{obs}, \tilde{\theta})$.

The conditional predictive distribution is usually straightforward once the observed data and values of the parameters θ are given, the first step is not. Though different approaches are possible, Markov chain simulation methods are most often used to do the Bayesian analysis to generate parameter values from the observed-data posterior distribution [29].

Steps 2 and 3 above are repeated to create m ($m > 1$) independent imputations: given Y^{obs}, m values of $\tilde{\theta}$ are independently drawn from the observed-data posterior distribution to get $\tilde{\theta}^{(t)}$ where $t = 1, 2, \ldots, m$. For each $\tilde{\theta}^{(t)}$, one imputed set of values of Y^{mis} is randomly drawn for the corresponding conditional predictive distribution $p\left(Y^{mis} \middle| Y^{obs}, \tilde{\theta}^{(t)}\right)$ [28].

I.E. Steps 2 and 3 are repeated for more imputations.

There are two main assumptions to multiple imputation. First, the missing data should be MAR (i.e. the probability an observation is missing may depend on observed data but not on missing data). MAR allows for the possibility that the probability of missingness can be predicted from the available data [30]. The missing data mechanism is said to *ignorable* if the data are MAR, and the parameters of the data model and missingness parameters are distinct [18]. The MAR requirement is regarded as more important; for practical purposes, the missing data model is ignorable if MAR holds [17]. The second assumption: the imputation model must match the model used for analysis which Rubin termed a "proper" imputation model [31]. Rubin (1987) gives a more precise, technical definition, but if the multiple imputations are proper then the average of the estimators is a consistent, asymptotically normal estimator, and an estimator of its asymptotic variance is given by a simple combination of the average of the complete data variance estimators and the empirical variance of the m estimators (the "between imputation variance") according to "Rubin's rule" (defined later) [32]. From a practical standpoint, it is more important that the chosen imputation model performs well over repeated samples than it is to be technically proper [30]. Last, the algorithm used to generate imputed values must be "correct"; it must allow for and include the necessary variables and their associations [31].

Analyzing the m imputed sets

The second step in the multiple imputation process is to analyze each imputed data set on its own. Here we let Q denote the parameters of scientific interest (e.g. a regression coefficient, or a Kappa statistic). In general though, Q can represent any estimand of scientific interest. So, in this step, $\hat{Q}$ is estimated for each imputed data set, along with their variance-covariance matrices. This is typically done by the intended method had the data been complete, since the missing data have been filled in by the imputations and can now be considered complete data. The results of the m analyses will differ due to differences in the imputed values for each set, i.e. the uncertainty due to the missing observations [17] [33].

Pooling the results from the m analyses

The last step is to pool the m parameter estimates, $\hat{Q}_1, \dots, \hat{Q}_m$ into a single estimate $\bar{Q}$ and to estimate its variance-covariance matrix. For parameters Q that follow an approximately normal distribution, Rubin's rules provides the method to pool estimates [34].

Supposing that $\hat{Q}_l$ is the estimate for the lth imputation that contains k parameters and is represented as a $k \times 1$ column vector. The combined estimate $\bar{Q}$ is equal to the average of the estimates from each of the complete imputed data sets

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l$$

The combined variance-covariance matrix incorporates both within-imputation variability and between-imputation variability. The within-imputation variance reflects the uncertainty about the results from a single imputed data set which is the conventional statistical variance due to the fact that we are taking a sample instead of observing the entire population. Suppose that U_l is the variance-covariance matrix of the estimate $\hat{Q}_l$ from the lth imputation,

then the combined within-imputation variance $\bar{U}$ is equal the average of the complete data variances

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m} U_l$$

The between-imputation variability reflects the uncertainty due to missing information, i.e. the variance between (among) the m complete data estimates

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})^t(\hat{Q}_l - \bar{Q})$$

where the superscript t indicates transpose when Q is a vector.

So the total variance T is given by

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

Note that the total variance T is not given by the simple sum of $\bar{U}$ and B. An additional term B/m is included to reflect the additional variance since $\bar{Q}$ is estimated for finite m. Including the extra term ensures that multiple imputation works at low m; not including it would produce p-values that are too low and confidence intervals that are too narrow [17] [34].

Statistical Inference

For multi-parameter inference, approaches are available such as Wald Test, likelihood ratio test, and χ^2-test. These methods are complex and not utilized in this thesis; further details can be found in references [17]. Most if not all MI software is able to facilitate these calculations.

For single parameter or scalar inference, like kappa statistics and others examined in this thesis, Wald-type significance tests and confidence intervals can be calculated in the usual

way [33]. Since the total variance of T is not known, $\bar{Q}$ follows a t-distribution rather than normal. So univariate tests are based on the approximation

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_v$$

where t_v is Student's t-distribution with v degrees of freedom. Calculation of degrees of freedom is discussed in references [17] [30] [35].

The $100(1 - \alpha)\%$ confidence interval for Q is calculated as [17] [34]

$$\bar{Q} \pm t_{v,1-\alpha/2}\sqrt{T}$$

Pooling non-normal quantities

Rubin's rules for pooling results from m complete data analyses were discussed above and are based on the assumption that the parameter estimates $\hat{Q}$ are normally distributed around the population value Q with a variance of U. When faced with pooling quantities with non-normal distributions (e.g. odds ratios, hazard ratios, etc.), some transformation may be required to ensure that their distribution is close to normal. Statistical inference is improved by first transforming the estimates to approximately normal, then applying Rubin's rules, and back-transforming to the original scale [17]. White et al. provide a summary of common statistics that can and cannot be using Rubin's rules directly [33].

Table 2 – Pooling common statistics with Rubin's rules
Common statistics that can and cannot be combined directly using Rubin's rules.

Can be combined without transformation	May require sensible transformation before combination	Cannot be combined
· mean · proportion · regression coefficient · linear predictor · C-index · area under the ROC curve	· odds ratio · hazard ratio · baseline hazard · survival probability · standard deviation · correlation · proportion of variance explained · skewness · kurtosis	· *p* -value · likelihood ratio test statistic · model chi-squared statistic · goodness-of-fit test statistic

Note that the main statistic of interest for this thesis, kappa statistics, is not mentioned specifically in the table above. Nor did literature review provide any example of how to transform or not transform kappa statistics for combining. For the purposes here, kappa will be treated the same as a proportion and will be combined without transformation; this will be mentioned as a limitation in a later section.

Multivariate imputation by Chained Equations (MICE)

The description above for multiple imputation illustrates the ideas behind univariate imputation, i.e. only one variable in the data set has missing data to be imputed. In practice, missing data are almost always multivariate. Conveniently, the multivariate problem can be broken into a series of univariate problems and solved by univariate imputation [36].

Notation

Some notation is needed to aid in the discussion:

- Let Y_j be one of p variables, where $j = 1, \ \dots, p$. Then the collection of variables to be included in the imputation model are $Y = (Y_1, \dots, Y_p)$.

  The collection of variables Y could be the whole data set or a subset of variables. It would depend on the choices the researcher makes in variable selection for the imputation model.

  These variables included in the imputation model can be complete (no missing) or partially complete (some missing). In the regression models, they could be the dependent variable being predicted or one of the covariates that is used to impute (a predictor).

- Each Y_j consists of observed and missing values, Y_j^{obs} and Y_j^{mis} respectively. Then the observed and missing data in Y are denoted as $Y^{obs} = (Y_1^{obs}, \dots, Y_j^{obs})$ and $Y^{mis} = (Y_1^{mis}, \dots, Y_j^{mis})$.

- The collection of $p - 1$ variables in Y except Y_j is denoted as $Y_{-j} = (Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_p)$. To clarify the distinction, if Y_j is being imputed then Y_{-j} would be the covariates that are used in the regression model.

- Let θ be the unknown parameters of the scientifically interesting model with the multivariate distribution $P(Y|\theta)$ for the hypothetically complete data set.

Fully conditional specification

Two common approaches have emerged to handle multivariate imputation: joint modeling (JM) and fully conditional specification (FCS). The typical model is the JM model developed

by Schafer (1997) using the multivariate normal model [30]. This assumes that variables have a normal distribution, all conditional expectation functions are linear, and all conditional variance functions are homoscedastic [15]. JM techniques exist for other multivariate models also. The general idea applies that the multivariate distribution is specified for the missing data and imputations are drawn from the conditional distributions using Markov chain Monte Carlo (MCMC) methods. The JM method assumes though that the multivariate distribution is a reasonable description of the data [37].

Fully conditional specification (FCS) in contrast does not explicitly assume that a particular form of the multivariate distribution as in JM, though it does assume that a multivariate distribution exists. Instead, FCS implicitly defines a multivariate distribution, $P(Y|\theta)$, by specifying a separate conditional distribution on a variable-by-variable basis. FCS accomplishes this by specifying the multivariate distribution by iteratively sampling a set of conditional distributions [38]

$$P(Y_1|Y_{-1}, \theta_1)$$

$$\vdots$$

$$P\left(Y_p\middle|Y_{-p}, \theta_p\right)$$

The $\theta_1, \dots, \theta_p$ parameters are not of scientific interest and only serve to model the respective conditional densities used for imputation. So they are not intended to be a product of a factorization of the "true" joint distribution of $P(Y|\theta)$ [37]. Starting from a simple draw from the observed marginal distribution, successive draws are to that the tth iteration of the method is as follows

$$\theta_1^{*(t)} \sim P\left(\theta_1 \middle| Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}\right)$$

$$Y_1^{*(t)} \sim P\left(Y_1^{mis} \middle| Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_1^{*(t)}\right)$$

$$\vdots$$

$$\theta_p^{*(t)} \sim P\left(\theta_p \middle| Y_1^{obs}, Y_2^{(t)}, \dots, Y_p^{(t)}\right)$$

$$Y_p^{*(t)} \sim P\left(Y_p^{mis} \middle| Y_p^{obs}, Y_2^{(t)}, \dots, Y_{p-1}^{(t)}, \theta_p^{*(t)}\right)$$

One cycle through all Y_j makes one iteration t. Since no information about Y_j^{mis} is used to draw $\theta_p^{*(t)}$, this approach differs from MCMC method to joint modeling and convergence can be quite fast [38]. Though convergence should be monitored, the suggested number of iterations can be fairly low, 5 to 20 [17] [37], especially compared with other MCMC techniques which can require thousands of iterations. The fast convergence is achieved when there is independence between the imputations themselves. The univariate imputation models create imputations that are already statistically independent for a given value of the regression parameters [37]. The t iterations are executed m times in parallel to generate m multiple imputations.

If the joint distribution defined by the specified conditional distributions exists, then this process is a Gibbs sampler [39], a Bayesian simulation technique that samples from the conditional distributions in order to obtain samples from the joint distribution. FCS is a very flexible method that is adaptable to the data, but the drawback to this flexibility is that the joint distribution may not even exist and convergence criteria are unclear. Two conditional densities are compatible if a joint distribution exists that has the given densities as its conditional densities [38]. The theoretical weakness of FCS is known as *incompatibility of conditionals* where no joint distribution exists for the specification of conditional

distributions. FCS is able to produce imputed data whether the joint distribution exists or not. FCS is guaranteed to work if the conditionals are compatible [39]. The issue of incompatibility of conditionals is still an open topic of research, and not much is known about the impact on the quality of imputations. FCS appears to be robust when the condition is not met, and the issue is minor in practice when the rate of missing data is modest [17]. Simulation work has suggested that the issue is not as serious, but more work in realistic settings is needed. To minimize the issue, Van Buuren suggests to ensure that the order in which variables are imputed is sensible [37].

MICE algorithm

Several implementations of FCS exist, this thesis utilizes the MICE algorithm developed by Van Buuren and applied in his R software package `mice`. The steps of which can be explained generally as follows [17] [23] [33]:

1. An imputation model $P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{-j}, R\right)$ is specified for each variable Y_j with $j = 1, \dots, p$.

   These imputation models are decided by the researcher based on the types of variables (continuous, categorical, ordinal, etc.). To aid in these decisions, the software provides sensible built in defaults

Method	**Variable type**
Predictive mean matching	Numeric
Logistic regression	Binary
Multinomial logit model	Nominal
Ordered logit model	Ordinal

   A more extensive list is available in the `mice` documentation and the defaults can be overridden.

2. For each j, missing values are filled in with starting imputations Y_j^0 by a simple imputation using the observed values Y_j^{obs} (e.g. random sampling of observed value with replacement or mean substitution).

   This initialization is repeated for $t = 1, \dots, T$ and for $j = 1, \dots, p$.

3. The values for the variable to be imputed, Y_j, are set back to missing. The currently complete data except Y_j is defined $Y_{-j}^t = \left(Y_1^t, \dots, Y_{j-1}^t, Y_{j+1}^{t-1}, \dots, Y_p^{t-1}\right)$. So, except for the variable being imputed, the complete data is made of the values already imputed in the current iteration t $\left(Y_1^t, \dots, Y_{j-1}^t\right)$ and the values for the prior iteration $t-1$ for those not imputed yet in the current iteration $\left(Y_{j+1}^{t-1}, \dots, Y_p^{t-1}\right)$.

4. Draw $\phi_j^t \sim P\left(\phi_j^t \middle| Y_j^{obs}, Y_{-j}^t\right)$ where ϕ_j represents the unknown parameters of the imputation model. The observed values Y_j^{obs} are regressed on the other variables in the imputation model Y_{-j}^t in order to obtain estimates of the regression model parameters ϕ_j^t. Here, the type of regression model is chosen based on the type of variable being imputed (continuous, binary, ordered or unordered categorical), so each variable is imputed using its own model.

5. Draw imputations $Y_j^t \sim P\left(Y_j^{mis} \middle| Y_j^{obs}, Y_{-j}^t, \phi_j^t\right)$, the corresponding posterior predictive distribution of Y_j^t. So missing Y_j^{mis} are replaced with predictions (imputations) from the regression model that was fit in step 4. Now, all the values for Y_j^t consist of both observed and imputed values which will be used when Y_j^t is included as a covariate for the regression models for the other variables.

6. End and repeat for the next j. End and repeat for the next t.

Steps 1 through 6 are repeated to create m imputed data sets.

One full pass through all the variables to be imputed is called a cycle or iteration. After one cycle, all of the missing values have been replaced with predictions from regression models that reflect relationships observed in the data [23]. The researcher decides how many cycles to perform so that the results have converged or stabilized (generally 10 to 20) which will produce one imputed data set. This whole process is repeated m times to produce m imputed data sets.

These m imputed data sets are analyzed separately to obtain m separate estimates which then pooled to a single estimate as described in the sections above.

Conclusion

This introduction serves to describe and overview the techniques that will be applied in this thesis. The research into these topics goes far beyond what is presented here. In the following Methods section, we will apply MICE to a data set with missing patient demographic information and assess agreement in data capture between two data sources with kappa statistics before and after multiple imputation.

METHODS

This section will explain the process and steps taken to apply MICE to the data from the original study. This includes describing the data, examining the missing values, setting up the imputation procedures, running and assessing the results of the imputation.

Background and data

Data for the current analysis was provided by OCHIN, a nonprofit community health information network of over 300 CHCs in 13 states. The data set was the same as used for the original study with the exception of some improvement to the patient information for primary language. In the original study, researchers used claims data for 2011 from Oregon's Medicaid program, collected 18 months after the end of the year to allow for processing lag. In the Medicaid data, the researchers identified adult patients aged 19 to 64 during 2011 that were fully covered by Medicaid and had ≥ 1 billing claim. EHR data was extracted from OCHIN's data storage for 43 Oregon CHCs for the year 2011. The researchers matched patients in the EHR data by Medicaid ID and included patients with ≥ 1 primary encounter in at least one of the study clinics during 2011. They excluded patients that had insurance coverage in addition to Medicaid, were pregnant, or died during the study period. The resulting sample contained 13,101 patient records [1].

Except for FPL, data was provided in one file with categories already created for the demographic information. FPL is a measure of income determined every year by the Department of Health and Human Services (HHS) to determine eligibility for certain programs and benefits [40]. For individuals, it is calculated based on household income and

household size. We were given the calculated *percent FPL* in a separate file with more than one observation per subject, i.e. not given income and household size. The records were combined following the method given in the original study: average all 2011 encounters, excluding null values ≥ 1000% (which were considered erroneous) [1]. We matched this information to the main data file based on a subject identifier (ID). FPL was the categorized as ≤ 138% or > 138%. If income is below 138% FPL and the state has expanded Medicaid coverage, then an individual can qualify for Medicaid based only on their income [40].

The original study assessed documentation for 11 adult preventive services. In this analysis, since we are looking at agreement between electronic health records and claims data across Race and FPL, we decided to limit the number of services presented here. All 11 preventive procedures were analyzed as we describe, but only 3 are presented: Cholesterol screening, Chlamydia screening, and Colonoscopy.

From the results of the original study, these procedures were chosen because of how they differed on (1) number of eligible patients, i.e. sample size, (2) level of agreement measured by kappa statistics in the original study, and (3) likelihood that the procedure was performed in clinic. The original study hypothesized post-hoc that there may be a relationship between agreement for procedures and how likely they are to be performed in clinic versus another location [1]. The Table 3 below shows how the three procedures differ across the criteria.

Table 3 – Procedures in this analysis
Three procedures chosen for this analysis and how they differ across the criteria.

Procedure	n	Kappa	Likelihood done in clinic
Cholesterol screening	High (12817)	High (0.80)	High
Chlamydia screening	Low (523)	Medium (0.52)	Medium
Colonoscopy	Medium (3761)	Low (0.26)	Low

Examine the missing data

One of the beginning steps in any data analysis is to examine the data. When missing data is present, additional care must be taken. Particularly with multiple imputation, the type of data and the missingness will drive the decisions in the process. It is helpful to begin by checking the entire data set for missing values. This can be done with visualization, counts, or frequencies. Also, note that in knowing the data one must know how missing values are coded and know how they are treated by the statistical software being used.

Preliminary checks showed that there are 3 variables in the data set that have missing values: ethnicity, race, and FPL. The `mice` package takes many exploratory data steps and combines them into two useful commands for understanding the counts and patterns of missing values: `md.pattern` and `md.pairs`. See Appendix A.1–A.6 for the output of these commands from R. The actual output takes some examination to decipher and one should refer to the vignette for the `mice` package for full details.

The information from the software for All Patients and those eligible for the procedures of interest are presented in Table 4 where we can get a sense of the counts and percentages of missing values in the data.

Table 4 – Counts and frequencies of missing values
Table of counts and frequencies of missing values in the data set for all patients and for those eligible for Cholesterol Screening, Chlamydia Screening, and Colonoscopy.

	All patients		Cholesterol		Chlamydia		Colonoscopy	
	n	%	n	%	n	%	n	%
Total observations	13101		12817		523		3761	
Complete	9706	74%	9506	74%	374	72%	2828	75%
Incomplete	3395	26%	3311	26%	149	28%	933	25%
Missing variable								
Any	4122	31%	4014	31%	162	31%	1111	30%
Ethnicity	531	4%	513	4%	8	2%	127	3%
Race	877	7%	841	7%	46	9%	178	5%
FPL	2714	21%	2660	21%	108	21%	806	21%
No. missing columns								
1	2807	21%	2745	21%	136	26%	796	21%
2	449	3%	429	3%	13	2%	96	3%
3	139	1%	137	1%	0	0%	41	1%

From the information in Table 1, we observe that FPL has the highest amount of missing (21%), then Race, and Ethnicity the lowest. It's useful to see here that there are similar trends of missingness for all patients and across procedures (i.e. there is a similar trend across rows). Of the 41 variables in the working data set only 3 have any missing information. For imputation purposes, it is important to note for later that the percent of incomplete observations (rows with at least one missing value) is 26%.

Visualization is a valuable tool in assessing the pattern of missing data. Below is a plot (Figure 5) that was created to get an overall picture of the amount of missing data. It was limited to just the demographic data in the entire set but could have just as easily included all variables.

Figure 5 – Missingness map
Plot of missingness among the demographic variables in the data set. Note that this plot was inspired by the missingness map in the R package `Amelia` [41].

Right away we can see that the missingness is multivariate and that FPL has the most missing. The pattern is non-monotone which we might expect given that the data is not longitudinal. There is a connected pattern to the data so that there will be available information in the data for imputation.

What is noted in the pattern of missing data is that Race and Ethnicity seem to be missing for many of the same observations. This can be investigated further in the information of missingness in pairs of variables from the `md.pairs` command (Appendix A.5). All 531 observations with missing Ethnicity are also missing Race. Logically there should be a relationship between Race and Ethnicity, but in this data set Race may not be informative of the missingness in Ethnicity.

Another well-developed visualization comes from the R package `VIM` [42], seen in Figure 6. This is another variation on the missingness map where only the three variables with missing values are shown (missing in beige and observed in blue). In the left side of the visualization, a bar plot shows the percent missing in each variable. The right shows the patterns of combinations of missingness along with the frequency. FPL and Race are missing together for 0.4% of observations, all three variables for 1.1%, race only 2.2%, and FPL only 19.2%. Again, here we see that 74% of rows in the data set are completely observed.

Figure 6 – Missing values histogram and pattern
Bar plot of frequency of missing values in FPL, Race, and Ethnicity in the data (left). Pattern and frequency of missingness (right). Yellow and blue colors indicate missing and observed, respectively.

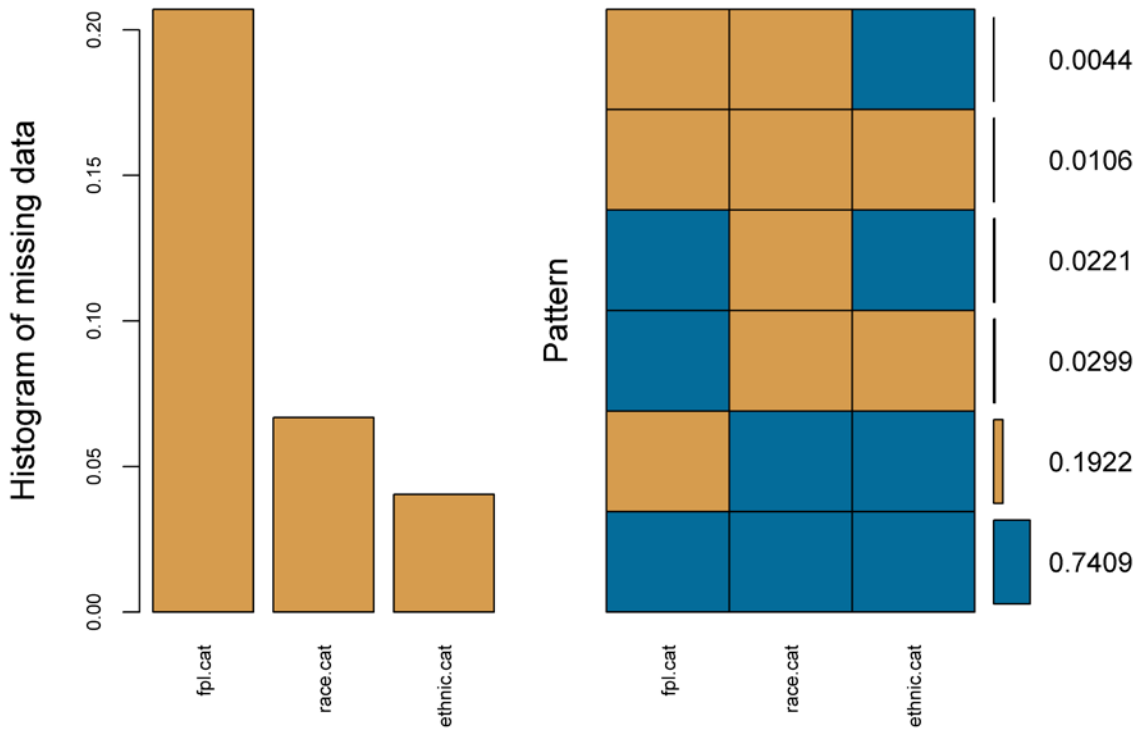


Table 5 below shows the counts and frequencies of all patients and those eligible for each procedure across categories Ethnicity, Race, and FPL. We will be sub-setting the entire data set (n = 13101) for each of the screening procedures; in a way, three separate sub-set analyses. This table let's us confirm that the missingness we see for all patients can be extended to the sub-sets. Looking across rows of the table, we indeed confirm that the distributions and missingness by procedure reflects what is seen when looking at all patients together.

Table 5 – Counts and frequencies by demographic categories
For all patients and for eligible patients for Cholesterol Screening, Chlamydia Screening, and Colonoscopy.

			Patients eligible for procedures					
	All patients		Cholesterol		Chlamydia		Colonoscopy	
	No.	%	No.	%	No.	%	No.	%
n	13101		12817		523		3761	
Race, ethnicity								
Hispanic	1186	9.1	1125	8.8	109	20.8	248	6.6
Non-Hispanic, white	8943	68.3	8782	68.5	300	57.4	2673	71.1
Non-Hispanic, other	2441	18.6	2397	18.7	106	20.3	713	19.0
Missing/Unknown	531	4.1	513	4.0	8	1.5	127	3.4
Total	13101	100.0	12817	100.0	523	100	3761	100.0
Race								
Asian/Pacific Islander	772	5.9	756	5.9	18	3.4	256	6.8
American Indian/Alaskan native	180	1.4	175	1.4	7	1.3	48	1.3
Black	1409	10.8	1388	10.8	70	13.4	397	10.6
White	9720	74.2	9518	74.3	365	69.8	2860	76.0
Multiple Races	143	1.1	139	1.1	17	3.3	22	0.6
Missing/Unknown	877	6.7	841	6.6	46	8.8	178	4.7
Total	13101	100.0	12817	100.0	523	100	3761	100.0
Federal poverty level								
<=138% FPL	10153	77.5	9930	77.5	401	76.7	2906	77.3
>138% FPL	234	1.8	227	1.8	14	2.7	49	1.3
Missing/Unknown	2714	20.7	2660	20.8	108	20.7	806	21.4
Total	13101	100.0	12817	100.0	523	100	3761	100.0

The exploratory data steps cannot be undervalued when dealing with missing data. Depending on what is found here can direct many of the decisions that will be made for the rest of the analysis with imputation whether you impute or not. Whether through counts and frequencies or visualization, there are good tools available to aid the researcher in this step.

Specification of the imputation model

Van Buuren labels the specification of the imputation model as the most challenging step in multiple imputation and that the model should [17]:

- Account for the process that created the missing data,
- Preserve the relations in the data, and
- Preserve the uncertainty about these relations.

The idea is that, by following these principles, the method will yield proper imputations and result in valid statistical inference. Van Burren and Groothuis-Oudshoorn outline 7 ordered choices to make to accomplish this process [37]. We will follow their suggestions to specify the model for the data set. The steps are as follows, using our example for discussion

1. **Decide if the missing at random (MAR) assumption is reasonable**. Missing not at random (MNAR) cannot be determined by looking at observed values; and, for practical reasons, we have not gone back and collected the missing information. Missing completely at random (MCAR) is convenient but can often be unrealistic. Though are tests and methods to check MCAR vs. MAR (e.g. Little's test, tests of association with missingness), MAR is a suitable starting place. From the inspection of the missingness above and the knowledge of the data, there are no strong reasons not to assume MAR.

2. **Decide on the form of the imputation model.** A univariate imputation model needs to be chosen for each incomplete variable. Using `mice`, the software makes default selections based on the variable type. The analyst should review these defaults and make different decisions if needed. The choice is driven by the variable type that is to

be imputed: continuous, categorical, ordinal, etc. For the three variables with missing information, the imputation models are specified as follows based on their variable type:

Table 6 – Models selected for target variables
List of the target variables to be imputed and the form of imputation model selected.

Variable	Variable Type	Imputation model
Ethnicity	Categorical factor with >2 levels	Multinomial logit regression
Race	Categorical factor with >2 levels	Multinomial logit regression
FPL	Categorical factor with 2 levels	Logistic regression

3. **Decide the set of predictors to include in the imputation model**. The general advice cited is to include as many variables as possible [17] [21]. Using all available data results in multiple imputations with minimal bias and maximal certainty. Including as many predictors as possible tend to make the MAR assumption more reasonable [17], i.e. if there are more variables in the model then it is more likely that missingness depends on an observed (included) variable.

The problem with this strategy is that for very large data sets, the models can become unwieldy due to multicollinearity and computational issues. Without derived variables, interaction effects or other complexities, it is reasonable to include all variables for small to medium data sets (20–30 variables). Van Buuren's notes that increase in explained variance in linear regression is negligible after including the best 16 variables. For imputation, with a large data set, his advice is to select a subset of the data that includes not more than 15 to 25 variables [17].

For variable selection with large data sets, he offers additional steps to consider [17] [36]:

1. Include all variables that will be in the model of scientific interest that will be applied post imputation.
2. Include all variables that are known to influence the occurrence of the missing data, i.e. related to nonresponse.
3. Include variables where the distributions differ between response and nonresponse groups which can be found by checking correlations with a missingness indicator (1/0) of the variable to be imputed.
4. Include variables that explain a considerable amount of variance to reduce the uncertainty of the imputations; simply identified by correlation with the variable to be imputed.
5. Remove variables in steps 2–4 with too many missing values in the subgroup of incomplete cases. If the variable to be imputed and the predictor variable are missing on the same cases, then do not include.

To apply this method of predictor selection, the first consideration was the data set is moderate in size. There are 37 usable variables for the imputation model; subject ID is an example of a variable excluded from this list because it does not contain information that would help the multiple imputation. Based on the advice that including more variables makes MAR more likely, it's reasonable to include all the available variables. This imputation model will be referred to as the *Full* scenario.

We also wanted to compare the performance of the imputation model with a reduced list of variables. How would the results differ if less variables were included? We

sought the advice of the primary author for the original study, John Heintzman, MD, MPH for advice on ranking the variables in order of importance. This reflects what would be done in practice: if the statistician questioned which variables may be most informative, they would discuss with the primary investigator (PI) and study team. Below in Table 7 is the Full list of 37 variables ranked.

Table 7 – Variables in Full scenario imputation model
List of variables that were include in the Full scenario, ranked in order of importance based on advice of the primary researcher on the original study. For a description of each variable, see the data dictionary in Appendix B.

1	FPL	11	EHR_COLONOSCOPY	21	EHR_FLU	31	ELIG_CERVICAL
2	RACE	12	DMAP_BREAST	22	EHR_CHLAM	32	ELIG_BREAST
3	ETHNICITY	13	DMAP_COLONOSCOPY	23	EHR_SMOKING	33	ELIG_COLON
4	LANGUAGE	14	EHR_CHOLEST	24	DMAP_CERVICAL	34	ELIG_BMI
5	AGE	15	DMAP_CHOLEST	25	DMAP_COLON	35	ELIG_FLU
6	SEX	16	ELIG_CHOLEST	26	DMAP_FLEXSIG	36	ELIG_CHLAM
7	PRIMARY_DEPT	17	EHR_FLEXSIG	27	DMAP_FOBT	37	ELIG_SMOKING
8	EHR_CERVICAL	18	EHR_FOBT	28	DMAP_BMI		
9	EHR_BREAST	19	EHR_BMI	29	DMAP_FLU		
10	EHR_COLON	20	EHR_WEIGHT	30	DMAP_CHLAM		

The `mice` software contains a useful tool to automate the predictor selection process, `quickpred()`. The function calculated two correlations for each variable pair, an (imputation) target and a predictor, using all available cases per pair. The first correlation uses the values of the variables as they are in the data. The second correlation uses the binary response indicator (1/0, observed/missing) of the target and the observed value of the predictor. If the largest (in absolute value) of these correlations is greater than a minimum value (default = 0.1), then the predictor will be added to the imputation model [37]. You can also specify in the function a minimum proportion of usable cases (observed within a subgroup of incomplete cases), but the default is zero. Table 8 shows the variables selected as predictors of the target variables: Ethnicity, Race, and FPL.

Table 8 – Variables selected by software
Variables selected by `quickpred()` function as predictors of the target variables in the imputation model.

Target	Predictors
ETHNICITY	PRIMARY_DEPT LANGUAGE RACE
RACE	PRIMARY_DEPT ETHNICITY LANGUAGE
FPL	EHR_FLU EHR_SMOKING

We used the combined recommendation of the advice from the investigator and the selections made by the software to cull a list of 21 variables for a *Reduced* scenario imputation model (roughly half of the Full scenario). We also need to make sure and heed the advice of Step #1 of predictor selection above to include all variables that appear in the complete data model. Since our research questions looks at the 3 procedures (Cholesterol screening, Chlamydia screening, and Colonoscopy) and 2 demographic categories (Race and FPL), these should be included in the imputation model.

Keeping variables relevant to the research question in the model and using the ranks from the investigator and using the selections from the software, the variables for the *Reduced* scenario are shown in Table 9.

Table 9 – Variables included in Reduced Scenario
List of variables that were include in the Reduced scenario.

1	FPL	11	EHR_COLONOSCOPY	21	EHR_FLU	31	ELIG_CERVICAL
2	RACE	12	DMAP_BREAST	22	EHR_CHLAM	32	ELIG_BREAST
3	ETHNICITY	13	DMAP_COLONOSCOPY	23	EHR_SMOKING	33	ELIG_COLON
4	LANGUAGE	14	EHR_CHOLEST	24	DMAP_CERVICAL	34	ELIG_BMI
5	AGE	15	DMAP_CHOLEST	25	DMAP_COLON	35	ELIG_FLU
6	SEX	16	ELIG_CHOLEST	26	DMAP_FLEXSIG	36	ELIG_CHLAM
7	PRIMARY_DEPT	17	EHR_FLEXSIG	27	DMAP_FOBT	37	ELIG_SMOKING
8	EHR_CERVICAL	18	EHR_FOBT	28	DMAP_BMI		
9	EHR_BREAST	19	EHR_BMI	29	DMAP_FLU		
10	EHR_COLON	20	EHR_WEIGHT	30	DMAP_CHLAM		

Since the software gave us it's predicted imputation model which only had 3 predictors per target variable, it will be interesting to see how it compares to the other two scenarios that we have planned. We will include another scenario based on the `quickpred()` function: *QPM scenario*.

To summarize, we have 3 planned imputation models: (1) Full scenario with 37 variables, (2) Reduced scenario with 21, and (3) QPM scenario with 3 variables per target.

4. The next choice is to **decide whether to impute variables that are function of other (incomplete) variables**. If the data set contains transformations or sum scores, it can be helpful to include the transformed variable in the imputation model. The variable that we considered here is the categorized FPL (above or below 138%) which is derived from a continuous FPL variable. There is more practical interest in Medicaid studies in whether someone is above or below the limit than the actual amount, so we decided to only include the categorical version of this variable.

5. **Decide the order in which variables should be imputed**. The software allows for the user to indicate the order that variables are imputed (aka the visiting sequence). The visiting sequence may impact the convergence of the algorithm. This is most important when the missing data are monotone or longitudinal, since missing in one variable may have an impact on the ones that follow. The software default is from left to right position in the data set. The imputations were run from highest number of missing to lowest, and from lowest to highest. There was no noticeable difference in convergence or computation time. The final decision was to impute from highest to lowest: FPL, first; then Race; and last, Ethnicity. The idea is to use more observed information to impute variables with high missing.

6. **Decide the number of iterations**. Convergence must be monitored, and the `mice` software offers visual tools to help. Compared to many modern MCMC techniques, which can require thousands of iterations, the MICE algorithm needs a much lower number to converge. The number of iterations needs to be large enough to stabilize the distributions of the regression parameters [37]. Van Buuren suggest that satisfactory convergence comes with just 5 to 10 iterations the software default is 5. With large amounts of missing data convergence can be slower. It cannot hurt to calculate extra iterations, so to assess the convergence over a longer stretch, we decided to use 20.

7. **Decide *m*, the number of multiply imputed data sets.** Advice in the past was to use a low number of imputations, 5 to 10. This idea is based on the statistical efficiency of point estimates [43].

   The true variance of a pooled set of *m* imputed data sets is similar to the estimated quantity given in the INTRODUCTION: *Statistical Inference* section of this thesis:

$$T_m = V + \left(1 + \frac{1}{m}\right) B$$

Where V is the average within-imputation variance and B is the between-imputation variance. Then for infinitely many imputations:

$$T_\infty = V + B$$

Then the relative efficiency of infinite imputations over m imputations is the ratio:

$$\frac{T_m}{T_\infty} = 1 + FMI/m$$

Here $FMI = B/(V + B)$, which is called the fraction of missing information (not to be confused with the percent of complete cases) [34].

With 50% FMI, five imputed data sets would yield point estimates that were 91% (inverse relative efficiency) as efficient as those based on an infinite number. Ten imputed data sets would be 95% as efficient. In the past, the additional resources, computer memory and processing, were not thought to be well spent for the small gains in efficiency [30].

What works for efficiency does not work for reproducible standard error estimates, confidence intervals, and *p*-values [17] [44]. Several sources advocate for a higher number and all tend to suggest the following rule of thumb: "the number of imputations should be similar to the percentage of cases that are incomplete" [33] [43] [17]. Percentage of incomplete cases is used here as a rough approximation for FMI since it requires less calculation. In larger data sets and those with very high missingness (>50%), additional care should be taken and more imputations may be needed. At a minimum, good practice would be to always use at least 5, even if, say, percent of incomplete cases were <5%.

> We followed a suggested good practice by Van Buuren: to use m = 5 as a convenient setting for model building, then increasing m for the final round of imputation. The data set had 26% incomplete cases and m was set to 30 for the final round of imputation.

Van Buuren's advice is that these choices are always needed. His book referred to for deeper details or for a different situation. These steps appear to be good, practical considerations to apply no matter the imputation method so that thoughtful attention is given to the setup.

Running the imputation models

We ran the specified imputation models for the 3 planned scenarios: (1) Full scenario with 37 variables, (2) Reduced scenario with 21, and QPM scenario with 3 variables per target. Variables were imputed in order from highest percent missing (FPL, 21%) to lowest (Ethnicity, 4%). The number of iterations was 20; and the number of imputed data sets was set m = 30.

The models needed to be reviewed for convergence. While there is no one method for determining if the MICE algorithm converged, the advice is to plot the parameters against the iteration number, and the mice package makes this easy. Figures 7, 8, and 9 show the plots of convergence for the three scenarios. For convergence, the lines should be well mixed with each other and show no signs of a trend [37].

Looking at the plots, the streams do seem to be mixing well in the Full and QPM scenarios, but there definitely is noticeable trend in the Ethnicity and Race variables. Though the scale

is so small that if the iterations were run out longer, the trend may go away. The Reduced scenario does not show the same trend as the others but isn't as well mixed. FPL plots look okay for all three scenarios.

Figure 7 – Convergence plot, Full Scenario, 3 target variables
Convergence of MICE algorithm for Ethnicity, Race, and FPL variables for the Full scenario. Observe fair mixing, but noticeable trend for Ethnicity and Race.

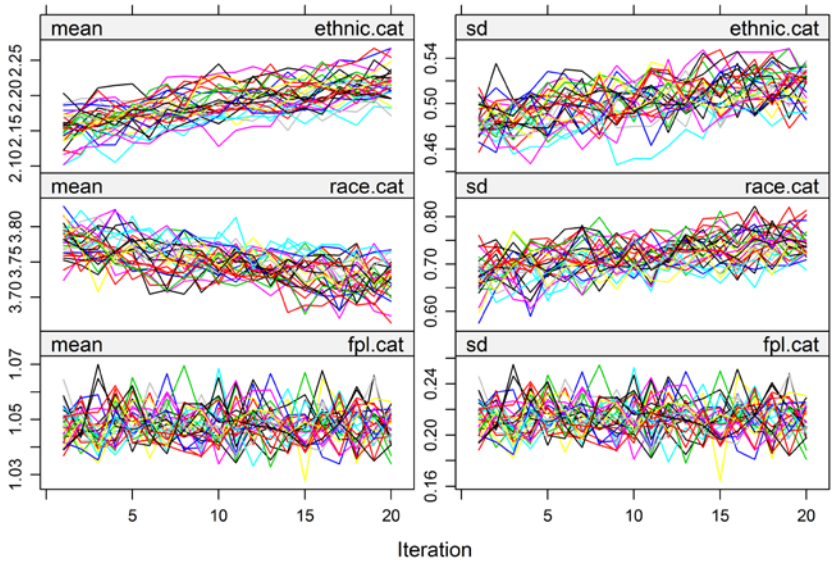


Figure 8 – Convergence plot, Reduced Scenario, 3 target variables
Convergence of MICE algorithm for Ethnicity, Race, and FPL variables for the Reduced scenario. Observe fair mixing, but noticeable trend for Ethnicity and Race.

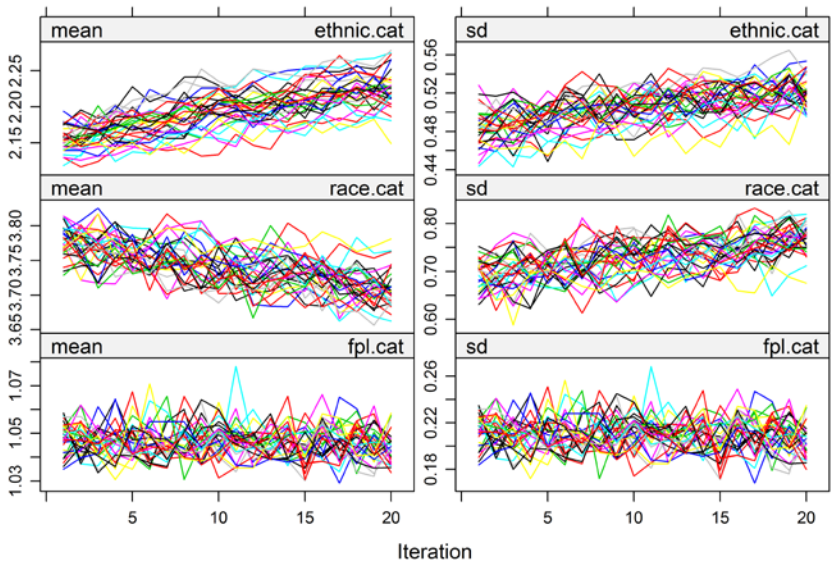
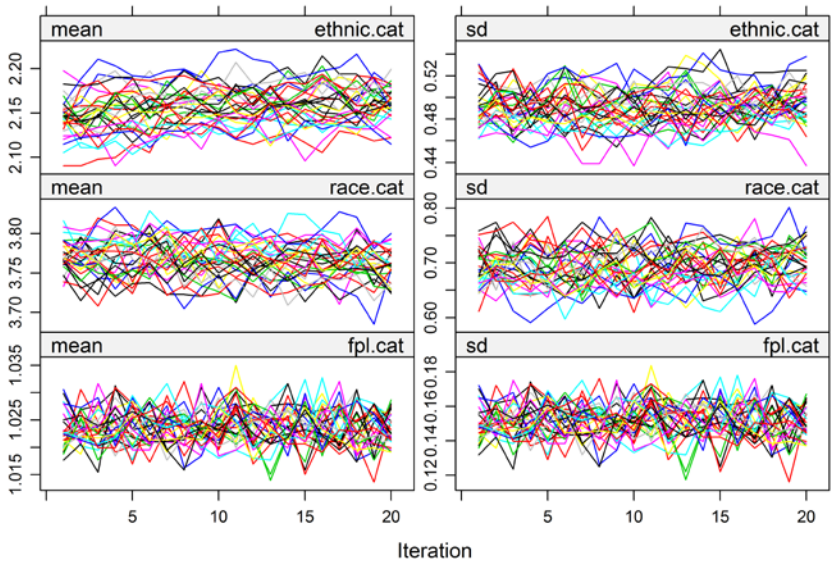


Figure 9 – Convergence plot, QPM Scenario, 3 target variables
Convergence of MICE algorithm for Ethnicity, Race, and FPL variables for the QPM scenario. Less trend than the other two scenarios, but only fair mixing.



The convergence in the three scenarios gave some pause for consideration. There appear to be issues with the Ethnicity and Race variables that results in less than ideal convergence. There is no known direct relationship between the two variables in the data set (they are coded differently and reside separately in the data base). In the real world there is a relationship between Ethnicity and Race. There could be a relationship between the observed data or the missing values for these two variables that is affecting the convergence.

For the purpose of this thesis, we are focused on Race and FPL demographic categories. So Ethnicity is not a variable directly related our interests. It's inclusion in the imputation model would be to aid in the imputation of the other two variables. There are two issues to raise with the Ethnicity variable:

1. There isn't a lot of information in this variable. Mostly it seems to be able to distinguish between "NH White" and others:

   Table 10– Counts and frequencies of Ethnicity.

Ethnicity	n	Percent	Valid Percent
Hispanic	1186	9%	9%
NH White	8943	68%	71%
NH Other	2441	19%	19%
Missing	531	4%	
Total	13101	100%	100%

2. The proportion of usable cases reveals that there may not be a lot of usable information in the variable. Van Buuren discusses this calculation which can be implemented with the `quickpred()` function or `md.pairs()` with some additional calculation (See Appendix A.6)

Table 11 – Proportion of useable cases
The proportion of usable cases in the data set for Ethnicity, Race, and FPL. Labels for target appears on the left and predictor on the top.

	Ethnicity	Race	FPL
Ethnicity	-	-	0.74
Race	0.40	-	0.78
FPL	0.95	0.93	-

Of the records with missing Race information, 40% have observed data on Ethnicity. So Ethnicity may be a poor predictor of Race. But for those missing FPL, 95% have available information for Ethnicity. There is a trade-off here to including or not including it in the imputation model.

We decided to create three more scenarios that did not include Ethnicity in the imputation model: Full_2, Reduced_2, and QPM_2 to distinguish from the scenarios where 3 variables were imputed. The convergence plots are shown in Figures 10, 11, and 12, where we see more healthy convergence before.

Figure 10 – Convergence plot, Full Scenario, 2 target variables
Convergence of MICE algorithm for Race and FPL variables for the Full scenario. Observe good mixing, and no noticeable trend.

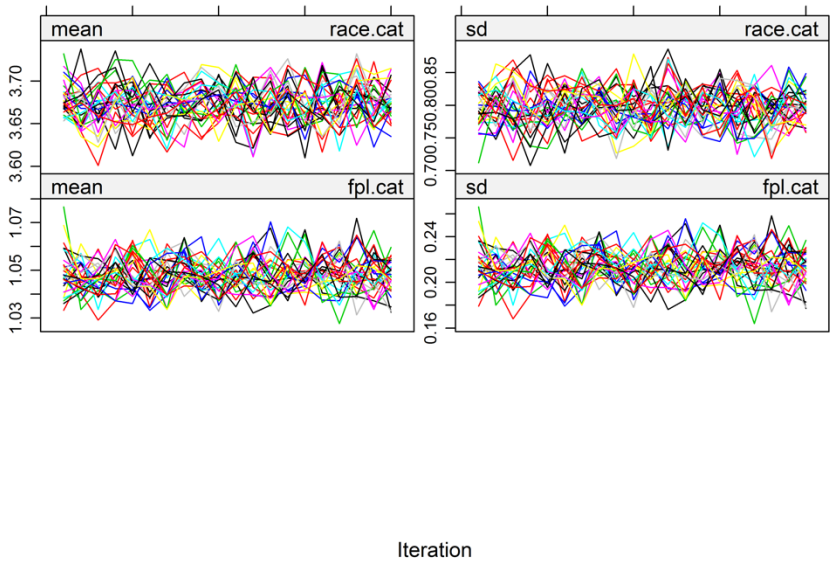


Figure 11 – Convergence plot, Reduced Scenario, 2 target variables
Convergence of MICE algorithm for Race and FPL variables for the Reduced scenario. Good mixing, and no noticeable trend.

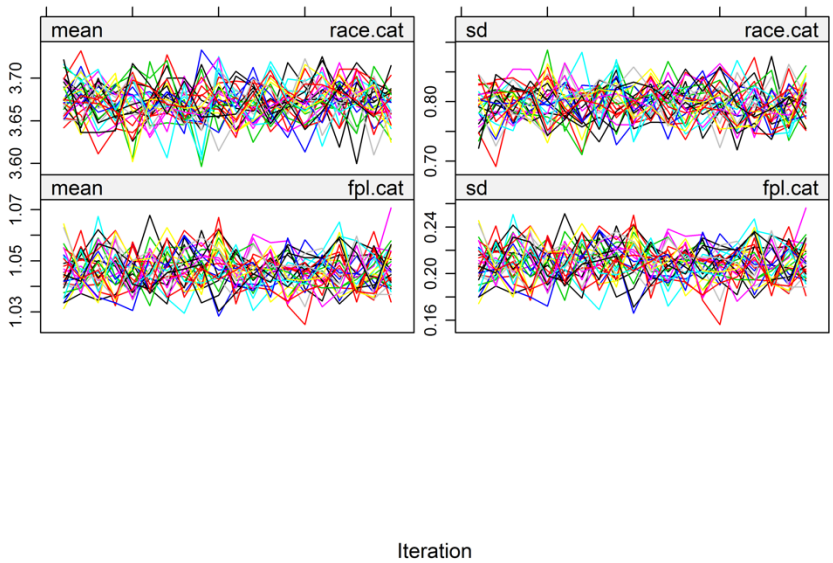
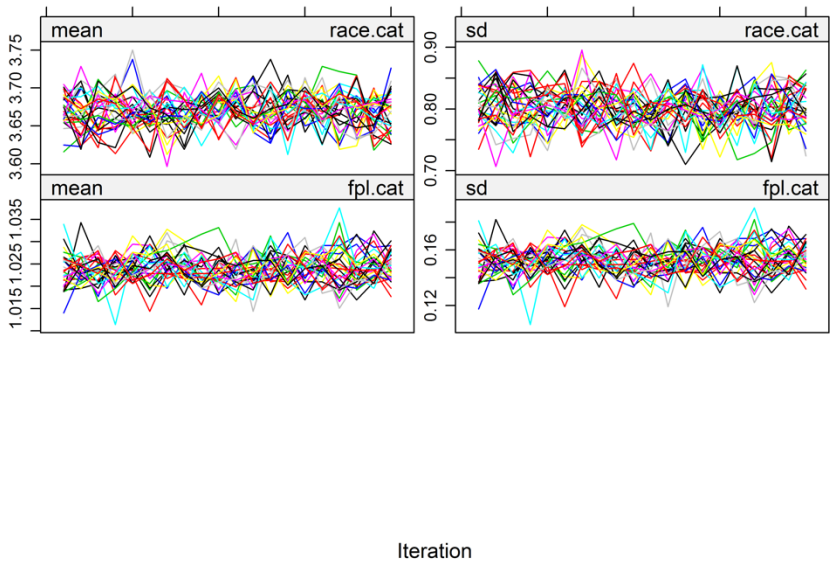


Figure 12 – Convergence plot, QPM Scenario, 2 target variables
Convergence of MICE algorithm for Race and FPL variables for the QPM scenario. Good mixing, and no noticeable trend.



We see improved convergence when Ethnicity is not included in the imputation model. But there may have been some information sacrificed for not including it as a predictor for FPL. If the percent missing for Ethnicity had been high, then it would likely make the decision easier. At least for now, we want to move toward comparing the results between all 6 scenarios.

Note on computation time

It's worth mentioning the amount of time that it took for the software to run the imputations on a desktop computer (Intel® Core™ i7-6700 CPU @ 3.40 GHZ; 230 GB RAM). With large data sets and high missingness, the amount of time to run the software to calculate the imputations could be very intensive. With the data set, 13101 rows and 26% incomplete

cases, it took about an hour on average for all 6 scenarios to calculate 30 imputed data sets. Table 12 shows the actual hours that each scenario took to run.

Table 12 – Computation time for multiple imputations
Hours that the software took to multiply impute 30 data sets for each of the 6 scenarios.

	3 Variables	2 Variables
Full	1.2	0.9
Reduced	1.0	0.7
QPM	0.6	0.4

It took a notably long time to pool the results, but this is due to the fact that we are calculating point estimates (i.e. kappa) and some lacking efficiencies in the code writing. If the results were based on a model, then we would have been able to take advantage of built in functionality of the software to handle statistical models and the pooling process. We were still able to take advantage of the `pool.scalar()` function in the `mice` package. For 6 each scenario we pooled 23 statistics related to agreement for 242 categories (11 procedures times 22 demographic categories). Also, for each of the 6 scenarios we calculated 8 different data sets (23 by 242) of statistics related to the pooled results (point estimates, variances, degrees of freedom, etc.). All of these were written and saved to files so that they could be used in later analysis.

All the work for each scenario took about 12 hours to run, and the complete code finished in about 3 days. It's an important caution that though imputations may happen quickly, the subsequent pooling can take a fair amount of time. With more time and some advice from a more advanced R programmer, some of this computation time could be reduced.

Assessing the imputations

In imputation, diagnostic checking involves assessing whether the imputations are plausible. Good imputations should have a distribution similar to the observed data. To assess the plausibility, it is recommended to focus on the distributional discrepancy, the difference between observed and imputed data [17]. Some things to check for [37]

- The imputed values could have been obtained had they not been missing,
- Imputations should be close to the data, and
- Imputed data should be possible to occur and make common sense (e.g. pregnant fathers).

Data visualization is a helpful tool to identify discrepancies between the observed and imputed data. Unfortunately, most of the graphics discussed in literature as examples use continuous variables: kernel densities, distributional dot plots, and box plots. Here we are working with categorical variables and so the best alternative is likely a box plot which works well when you have only a few imputations (for example, 5 sets from the imputation model building phase). With 30 imputed data sets like we have here, it can be a little cumbersome, so we adapted the idea of the kernel density plot (Van Buuren page 149) to create Figures 13, 14, and 15 using the Full scenario with 3 variables to impute [17].

In the 3 figures, we see a similar shape to the distributions of the imputed sets and the observed data. Figure 13 shows that the imputed sets have a higher percent Black than the observed. Looking at ethnicity in Figure 14, there are differences in the allocations of Hispanic and NH White. FPL in Figure 15 is very close to the same between observed and imputed sets.

Dramatic differences should raise concern, but the resulting figures don't seem too extreme to cause worry. MAR data mechanism may result in systematic differences between distributions. Only under MCAR would they need to be identical. Differences between observed and imputed values may well be appropriate and should be discussed with those with expert knowledge of the data subject matter.

Figure 13 – Distribution comparison: Race
Comparing distributions of observed and imputed data with m = 30 for Race. Original observed percents shown by red dots and imputed values shown in blue. Lines drawn between points are meant to aid in assessment of the distribution not to imply the data points are continuous.

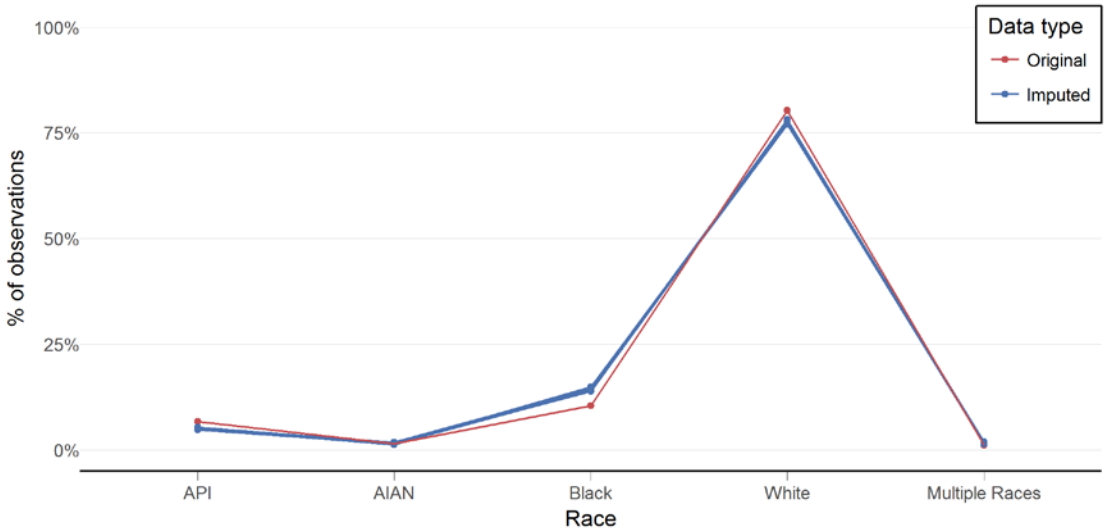


Figure 14 – Distribution comparison: Ethnicity
Comparing distributions of observed and imputed data with *m* = 30 for Ethnicity. Original observed percents shown by red dots and imputed values shown in blue. Lines drawn between points are meant to aid in assessment of the distribution not to imply the data points are continuous.

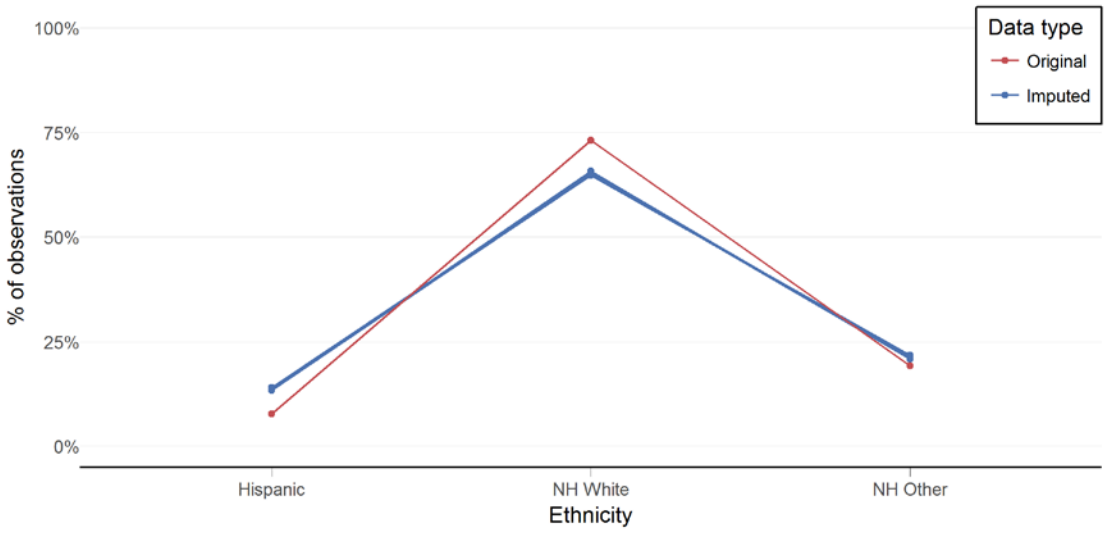
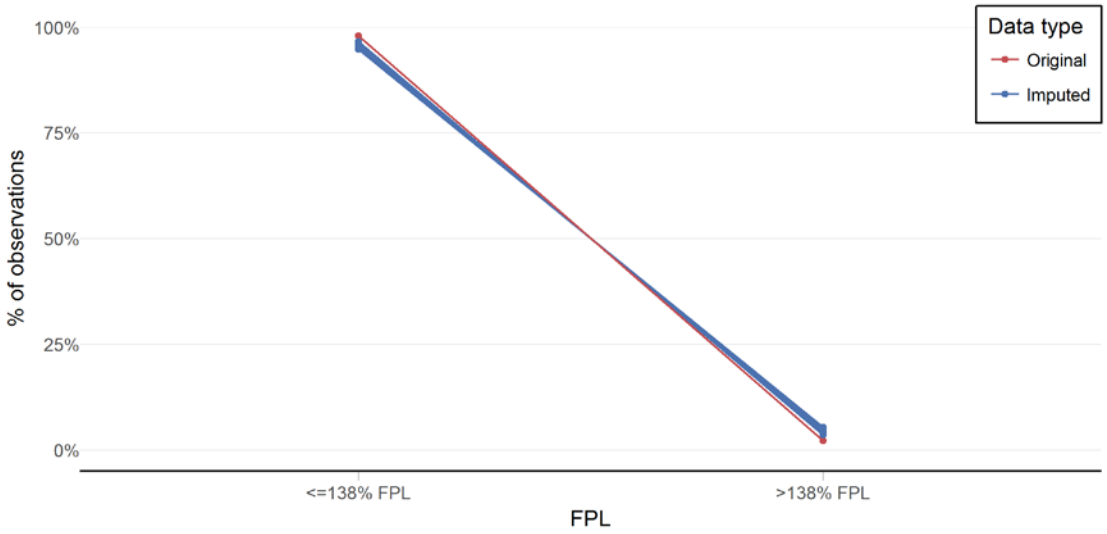


Figure 15 – Distribution comparison: FPL
Comparing distributions of observed and imputed data with *m* = 30 for FPL. Original observed percents shown by red dots and imputed values shown in blue. Lines drawn between points are meant to aid in assessment of the distribution not to imply the data points are continuous.



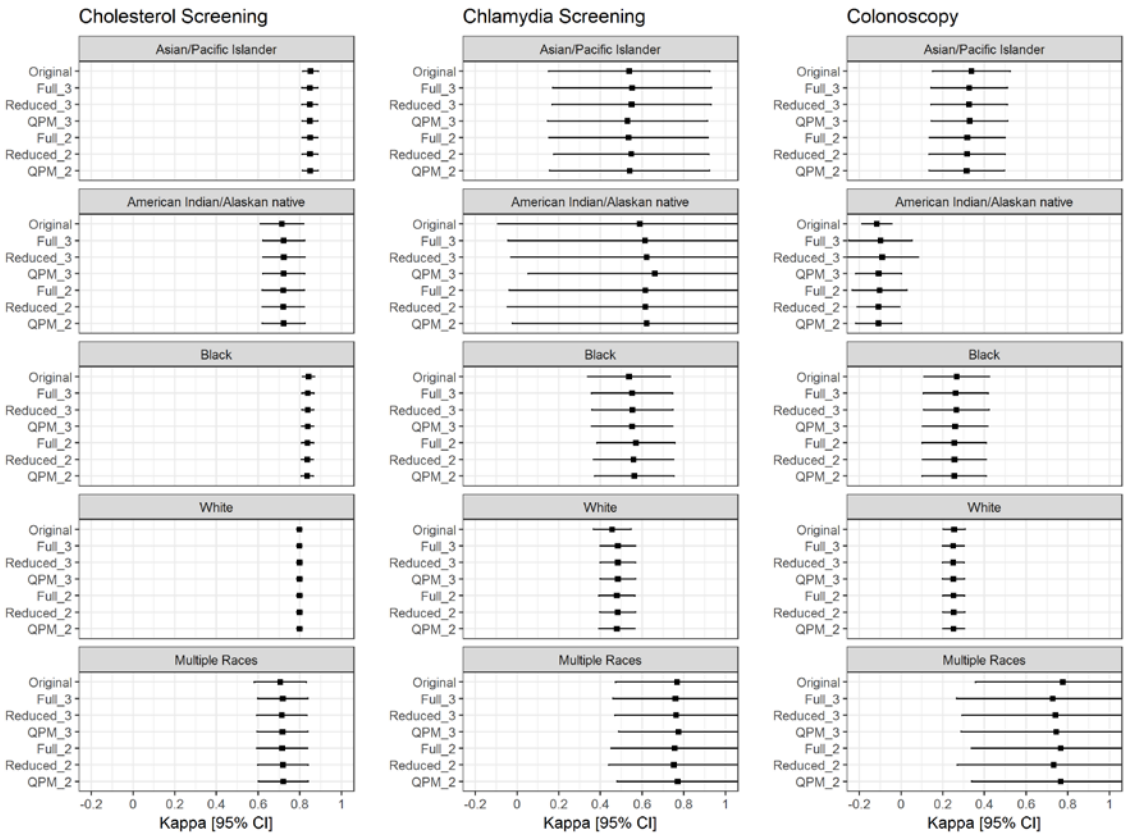
Note that we have only shown the visualizations for the Full scenario with 3 imputed variables. Other scenarios gave similar plots and are included in Appendix C for reference.

Comparing scenarios

At this point in the analysis, we hoped to compare the results for the 6 scenarios and see if there was a noticeable difference due to the different specifications of the imputation models. We will compare the main statistic of interest, kappa and its 95% confidence intervals (CIs), for the scenarios by procedure and by demographic strata.

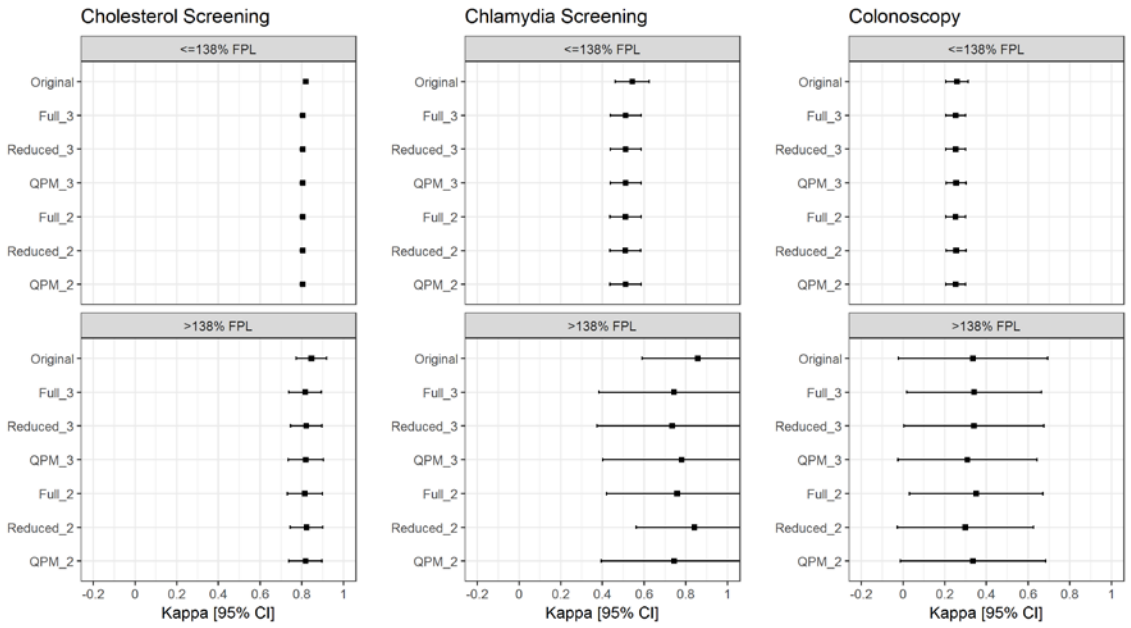
In Figure 16, the kappa statistics and 95% CIs are assessed visually by forest plots. Looking column by column, the scenarios do not seem to differ substantially. It's difficult to say that one performed better than the others. The kappa statistics for the Multiple Race category (Chlamydia and Colonoscopy) seem to change the most from scenario to scenario, but all are between 0.7–0.8 and would all be interpreted as showing substantial agreement.

Figure 16 – Results from 6 scenarios by Race
Kappa statistics and 95% confidence intervals plotted for the pre-imputation data and each of the 6 scenarios. By Procedure and Race.



Next we examine the results of the 6 scenarios for FPL categories, see Figure 17. Comparisons between the scenarios is similar to Race above. There are only noticeable differences between scenarios for >138% FPL for Chlamydia Screening and for Colonoscopy.

Figure 17 – Results from 6 scenarios by FPL
Kappa statistics and 95% confidence intervals plotted for the pre-imputation data and each of the 6 scenarios. By Procedure and FPL.



We suspect that the differences seen in the scenarios for Chlamydia Screening and Colonoscopy are due to low counts in the original data for Multiple Races and >138% FPL categories. The low counts would be sensitive to changes in the number of patients in the categories; the between-imputation variation is not necessarily larger than other categories, but just more noticeable because of scale. For example, assigning 1 or 2 patients to a category would be more visible in a category with 20 patients to begin rather a hundred or a thousand.

Ultimately, we want to select one imputation scenario to represent the results after-imputation. The results proved to be robust in spite of the potential issues with convergence noticed previously. The similar results across scenarios suggest that there are likely a small number of variables that dominate the imputation models. The advice from the literature is

that including more variable in the model makes a stronger argument for the MAR assumption. With this justification, we select the Full scenario with all 3 target variables.

RESULTS

In an analysis that involves missing data, the conclusions drawn from the statistical results are affected by the reporting or not reporting information about the missingness and the procedures used to handle it. Guidelines exist for what should be reported from an analysis with missing data, but they vary in their scope. Van Buuren compiled the recommendations from sources and presented a list of questions that need to be answered when using multiple imputation [17]. His suggestions were the guide for the following paragraph which could serve as a statistical methodology summary. His text should be consulted for further details of the guidelines.

Three of 41 variables in the data set had missing data: race (7% missing), ethnicity (4%), and federal poverty level (21%). Of the total 13,101 observations, 3,395 (26%) were incomplete. Though the percentages are lower, the reasons for missing Race and Ethnicity is not fully known. Full demographics are available in Table 13 below. The original study had reported 2.0% missing Federal Poverty Level (FPL) data, and, from the raw data, we came up with the much higher percent listed before. Normal procedure would be to investigate the discrepancy, but we decided to use the higher percent to enrich the learning goals of this analysis. Multiple imputation [34] was used to create and analyze 30 multiply imputed data sets. Incomplete variables were imputed under fully conditional specification [38]. Calculations were done in R version 3.4.0 (2017-04-21) using the default settings of the `mice`

2.3 package [37] as these were deemed appropriate for our study setting. Kappa and other statistics were estimated based on established formulae and theory applied to each imputed data set separately. The estimates and stand errors were combined using Rubin's rules without transformation. The final results were compared to the analysis done on the subset of cases with complete, observed data.

Table 13 – Demographic characteristics of study sample

	Patients appearing in both EHR and claims (N=13,101)	
	No.	%
Gender		
Female	8,600	65.6
Male	4,501	34.4
Race		
Asian/Pacific Islander	772	5.9
American Indian/Alaskan native	180	1.4
Black	1,409	10.8
White	9,720	74.2
Multiple Races	143	1.1
Unknown	877	6.7
Race, ethnicity		
Hispanic	1,186	9.1
Non-Hispanic, white	8,943	68.3
Non-Hispanic, other	2,441	18.6
Unknown	531	4.1
Primary Language		
English	10,927	83.4
Spanish	589	4.5
Other	1,585	12.1
Federal poverty level		
≤138% FPL	10,153	77.5
≥138% FPL	234	1.8
Missing/Unknown	2,714	20.7
Age in years (as of January 1, 2011)		
19–34	4,632	35.4
35–50	5,033	38.4
51–64	3,436	26.2
Mean (SD)	40.6 (12.3)	

The original study assessed agreement between EHR and Medicaid claims data for 11 preventive procedures. These results focus on three of those procedures (Cholesterol screening, Chlamydia screening, and Colonoscopy) and the agreement by Race and FPL categories.

Cholesterol screening

Cholesterol screening was selected as one of the procedures due to its high number of eligible patients (n = 12,817). Eligible patients were men and women age 20 or older; the screening includes low density lipoprotein, high density lipoprotein, total cholesterol, and triglycerides. The original kappa statistic indicated substantial agreement between EHR and Medicaid claims (κ = 0.80; 95% CI: 0.79 to 0.81). The original study hypothesized that the high agreement could be related to the location the service is provided. Cholesterol screening is usually done in the primary care setting.

Table 14 gives kappa and other statistics, prior to imputation, for Cholesterol screening in total, by race, and by FPL. Even when stratified, the agreement between EHR and Medicaid claims is substantial. Highest agreement is seen among the category for Asian/Pacific Islander (κ = 0.85; 95% CI: 0.81 to 0.89). Lowest agreement was seen in the categories for American Indian/Alaskan native (κ = 0.71; 95% CI: 0.61 to 0.82) and Multiple Races (κ = 0.71; 95% CI: 0.58 to 0.83); these two groups also have the lowest numbers of eligible patients (175 and 139) also seen in the wider confidence intervals for kappa. For these two groups, proportion screened in the EHR data, assessed by EHR (+) p1 and Sensitivity, are lower relative to the other Race categories and contributes to the lower but still substantial

agreement. Looking at FPL, agreement is higher for those with FPL (above or below 138% FPL) than those with Missing/Unknown.

Table 15 provides the same information but after multiple imputation. The two figures are nearly indistinguishable. And agreement in the categories are very similar. The high number of eligible patients and substantial agreement prior to imputation likely contribute to the stable results post-imputation. Figure 18 gives another visual comparison of the kappa statistics and the 95% confidence intervals; there is a slight but noticeable shift lower in agreement for both FPL groups. This seems due to the lower agreement pre-imputation for those with Missing/Unknown FPL.

Overall agreement was already substantial pre-imputation and stayed that way after imputation likely due to the high number of eligible patients. Substantial agreement by category is likely related to the service being provided in clinic. No categories stand out as having vastly different agreement than the others.

See Appendix D for tables of results pre- and post- imputation with additional statistics.

Table 14 – Cholesterol screening agreement results *prior to imputation*

EHR (+) p1 and Claims (+) p2 indicate proportion of patients listed as screening in the respective data sources. EHR (+), Claims (+) p1+p2 shows the combined percent of total eligible patients screened in both data sources. Sensitivity, here, is the proportion of subjects screened in EHR given that they are screened in Medicaid claims. Specificity is the proportion not screened in EHR given not screened in Medicaid claims. Eligible patients: Men and women aged ≥20; screening includes low density lipoprotein, high density lipoprotein, total cholesterol, and triglycerides.

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	12817	39.5	42.1	45.5	0.86	0.94	0.80 [0.79, 0.81]
Race							
Asian/Pacific Islander	756	50.1	47.5	52.5	0.95	0.90	0.85 [0.81, 0.89]
American Indian/Alaskan native	175	33.7	44.0	45.7	0.73	0.97	0.71 [0.61, 0.82]
Black	1388	41.8	40.4	44.9	0.92	0.92	0.84 [0.81, 0.87]
White	9518	38.4	42.1	45.1	0.84	0.95	0.80 [0.79, 0.81]
Multiple Races	139	28.8	36.0	38.9	0.72	0.96	0.71 [0.58, 0.83]
Unknown	841	41.4	41.6	46.2	0.88	0.92	0.80 [0.76, 0.84]
Federal poverty level							
<=138% FPL	9930	41.5	43.5	46.9	0.88	0.94	0.82 [0.81, 0.83]
>138% FPL	227	35.2	35.2	38.8	0.90	0.95	0.85 [0.77, 0.92]
Missing/Unknown	2660	32.2	37.8	40.9	0.77	0.95	0.74 [0.71, 0.77]

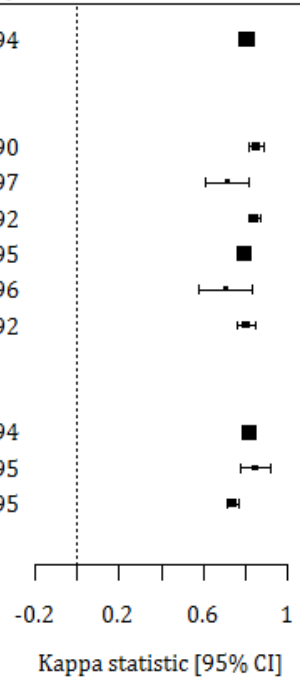


Table 15 – Cholesterol screening agreement results *after imputation*

EHR (+) p1 and Claims (+) p2 indicate proportion of patients listed as screening in the respective data sources. EHR (+), Claims (+) p1+p2 shows the combined percent of total eligible patients screened in both data sources. Sensitivity, here, is the proportion of subjects screened in EHR given that they are screened in Medicaid claims. Specificity is the proportion not screened in EHR given not screened in Medicaid claims. Eligible patients: Men and women aged ≥20; screening includes low density lipoprotein, high density lipoprotein, total cholesterol, and triglycerides.

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	12817	39.5	42.1	45.5	0.86	0.94	0.80 [0.79, 0.81]
Race							
Asian/Pacific Islander	795	49.9	47.5	52.5	0.95	0.90	0.85 [0.81, 0.89]
American Indian/Alaskan native	190	33.9	43.8	45.5	0.73	0.97	0.72 [0.62, 0.82]
Black	1483	41.9	40.6	45.1	0.92	0.92	0.84 [0.81, 0.87]
White	10191	38.6	42.0	45.1	0.84	0.95	0.80 [0.79, 0.81]
Multiple Races	155	29.6	35.8	38.9	0.74	0.95	0.72 [0.60, 0.84]
Federal poverty level							
<=138% FPL	12464	39.7	42.4	45.8	0.86	0.94	0.80 [0.79, 0.81]
>138% FPL	352	32.7	34.0	37.4	0.86	0.95	0.82 [0.74, 0.89]

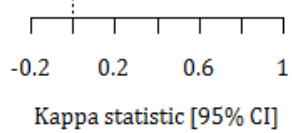
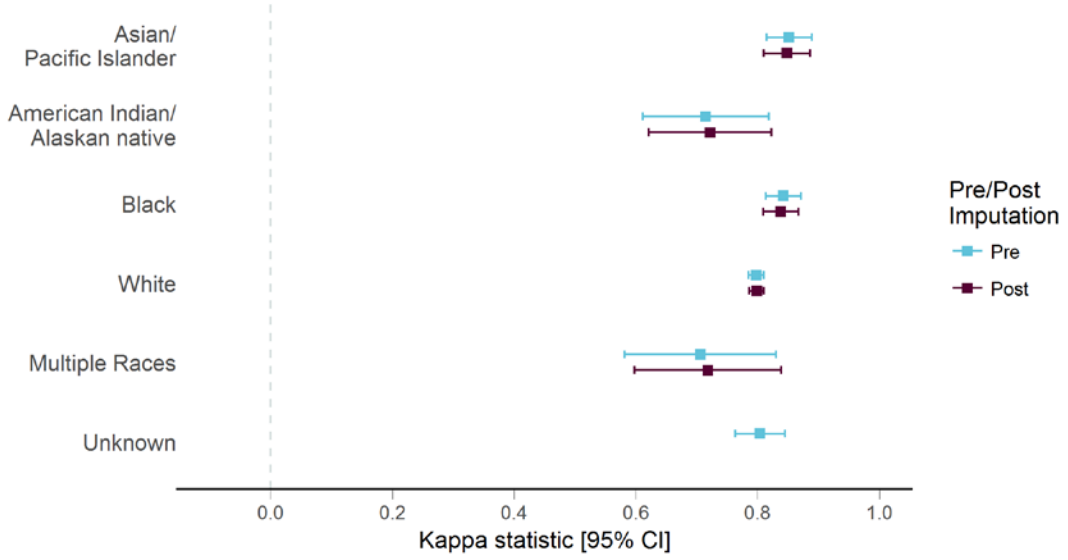
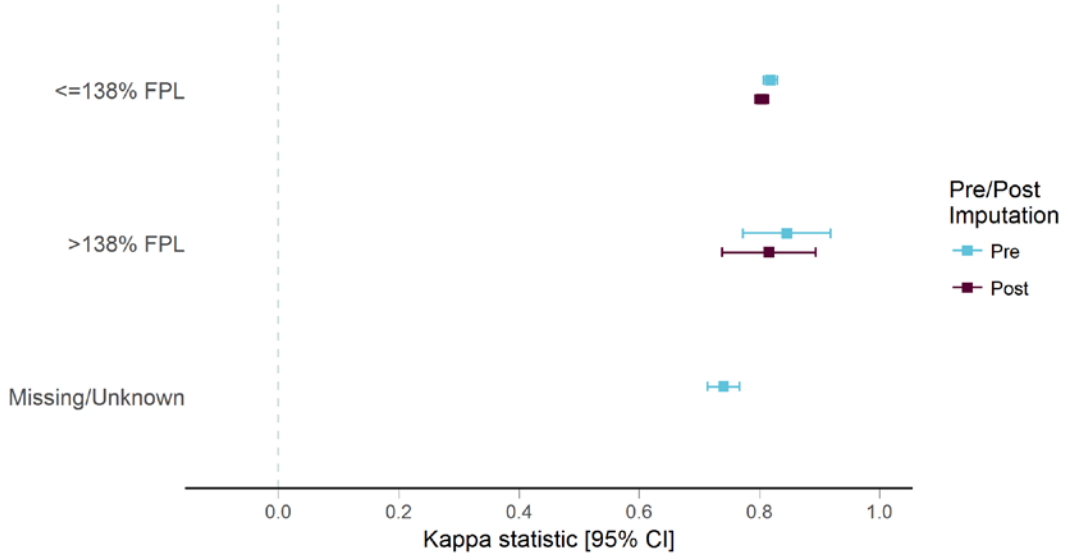


Figure 18 – Cholesterol screening: Visualization of kappa statistics and 95% CIs
Pre-imputation shown in blue; post-imputation in purple.

A. By Race



B. By FPL



Chlamydia screening

Chlamydia screening is included in this analysis since there were a low number patients eligible for this procedure (n = 523). Eligible patients were sexually active women between ages 19 and 24. Overall agreement between EHR and Medicaid claims for this screening was moderate (κ = 0.45; 95% CI: 0.45 to 0.59). The original study had presumed a likelihood of 0.48 that this service is provided in the primary care clinic; near even chance.

Kappa and other statistics prior to imputation are shown in Table 16 by total and by strata. Immediately apparent is the wider confidence intervals for kappa compared to what was seen above for Cholesterol screening. Particularly in the categories with low number of eligible patients: Asian/Pacific Islander, American Indian/Alaskan native, Multiple Races, and >138% FPL.

Table 17 shows post-imputation results where the confidence intervals are still fairly wide. An encouraging sign that multiple imputation did not over assign to Race or FPL categories with lower counts; increasing the numbers artificially would narrow the confidence intervals. Like we saw for Cholesterol screening, results are very similar before and after imputation. Highest agreement (post-imputation) was among Multiple Races (κ = 0.76; 95% CI: 0.54 to 0.93). Lowest agreement was for Whites (κ = 0.48; 95% CI: 0.40 to 0.57), which had the highest allocation of eligible patients for this procedure. Proportions of patients screened according to each data source (EHR (+) p1 and Claims (+) p2) were low indicating a low number of eligible patients in this sample had documentation in either source. Specificity was around 0.85 for each strata while Sensitivity varied quite a bit.

Figure 19 puts the kappa statistics and 95% CIs side by side where we can compare easily pre- and post- imputation visually. For the Race categories, results stayed fairly consistent. Unknown Race had a higher kappa than other categories pre-imputation, and these subjects' agreement did not seem to affect the other categories too much post-imputation. In FPL, we see a different effect where Missing/Unknown FPL had a much lower kappa pre-imputation (κ = 0.39; 95% CI: 0.23 to 0.56). For those with >138% FPL, kappa changed from almost perfect agreement pre-imputation (κ = 0.86; 95% CI: 0.59 to 1.00) to substantial post-imputation (κ = 0.74; 95% CI: 0.38 to 1.00). Though the count for this category is low before and after (14 vs. 20), the addition of other records through imputation will affect the kappa. In this case, the subjects with fair, pre-imputation agreement had the effect to reduce the agreement for the categories to which imputation assigned them.

In general, the moderate pre-imputation agreement persisted across all categories post-imputation despite the lower number of total eligible patients. Though when looking at the individual categories where the counts are divided further, the wide confidence intervals show a lack of precision.

See Appendix D for tables of results pre- and post- imputation with additional statistics.

Table 16 – Chlamydia screening agreement results *prior to imputation*

EHR (+) p1 and Claims (+) p2 indicate proportion of patients listed as screening in the respective data sources. EHR (+), Claims (+) p1+p2 shows the combined percent of total eligible patients screened in both data sources. Sensitivity, here, is the proportion of subjects screened in EHR given that they are screened in Medicaid claims. Specificity is the proportion not screened in EHR given not screened in Medicaid claims. Eligible patients: Sexually active women ages 19 to 24.

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	523	42.8	51.2	59.1	0.68	0.84	0.52 [0.45, 0.59]
Race							
Asian/Pacific Islander	18	55.6	66.7	72.2	0.75	0.83	0.54 [0.15, 0.93]
American Indian/Alaskan native	7	28.6	14.3	28.6	1.00	0.83	0.59 [-0.09, 1.00]
Black	70	58.6	71.4	75.7	0.76	0.85	0.54 [0.34, 0.74]
White	365	38.4	47.1	56.2	0.62	0.83	0.46 [0.37, 0.55]
Multiple Races	17	41.2	52.9	52.9	0.78	1.00	0.77 [0.47, 1.00]
Unknown	46	52.2	52.2	58.7	0.88	0.86	0.74 [0.54, 0.93]
Federal poverty level							
<=138% FPL	401	43.9	51.4	59.1	0.70	0.84	0.54 [0.46, 0.62]
>138% FPL	14	50.0	42.9	50.0	1.00	0.88	0.86 [0.59, 1.00]
Missing/Unknown	108	38.0	51.9	60.2	0.57	0.83	0.39 [0.23, 0.56]

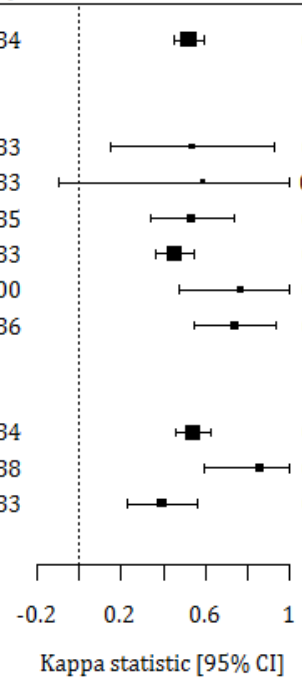


Table 17 – Chlamydia screening agreement results *after imputation*

EHR (+) p1 and Claims (+) p2 indicate proportion of patients listed as screening in the respective data sources. EHR (+), Claims (+) p1+p2 shows the combined percent of total eligible patients screened in both data sources. Sensitivity, here, is the proportion of subjects screened in EHR given that they are screened in Medicaid claims. Specificity is the proportion not screened in EHR given not screened in Medicaid claims. Eligible patients: Sexually active women ages 19 to 24.

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	523	42.8	51.2	59.1	0.68	0.84	0.52 [0.45, 0.59]
Race							
Asian/Pacific Islander	18	55.4	66.2	71.6	0.75	0.84	0.55 [0.17, 0.93]
American Indian/Alaskan native	7	28.7	15.4	28.7	1.00	0.84	0.61 [-0.04, 1.00]
Black	72	58.6	70.9	75.1	0.77	0.85	0.55 [0.36, 0.75]
White	405	39.8	47.7	56.5	0.65	0.83	0.48 [0.40, 0.57]
Multiple Races	18	41.2	52.2	52.8	0.78	0.99	0.76 [0.46, 1.00]
Federal poverty level							
<=138% FPL	502	42.9	51.8	59.6	0.68	0.84	0.51 [0.44, 0.59]
>138% FPL	20	42.3	37.8	46.2	0.90	0.86	0.74 [0.38, 1.00]

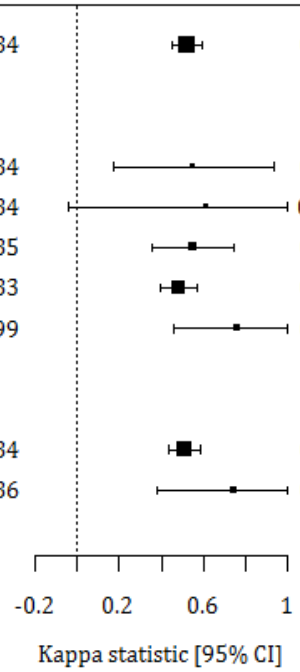
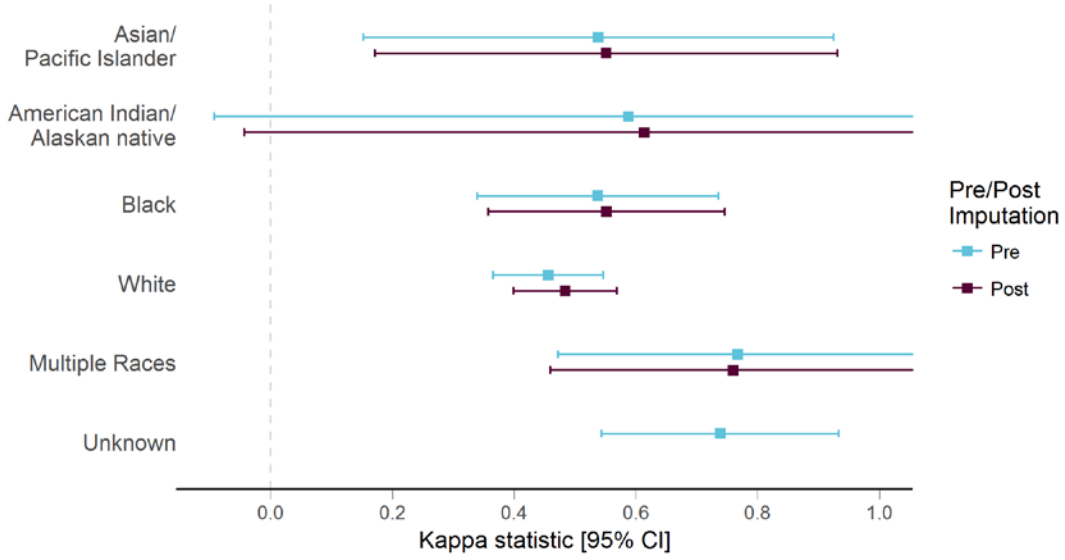
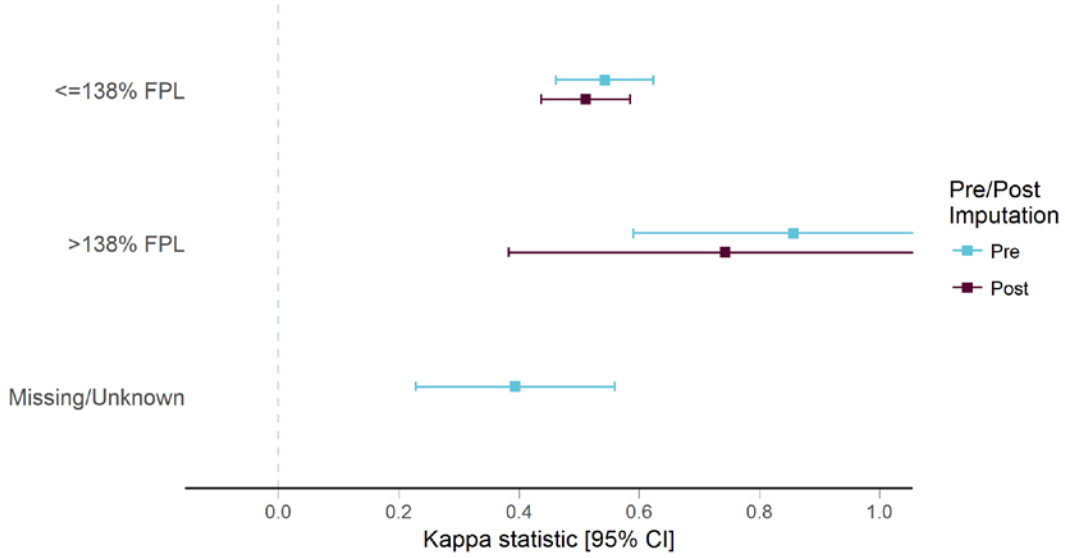


Figure 19 – Chlamydia screening: Visualization of kappa statistics and 95% CIs
Pre-imputation shown in blue; post-imputation in purple.

A. By Race



B. By FPL



Colonoscopy

Colonoscopy was selected as a procedure for the medium number of eligible patients relative to other screenings (n = 3761) but also for its fair (i.e. low) overall agreement (κ = 0.26; 95% CI: 0.21 to 0.30) due to the tendency to be referred out.

Of the three chosen procedures, EHR and Medicaid claims do not agree well in recording Colonoscopy. In Table 18, pre-imputation, the American Indian/Alaskan native strata has estimated agreement worse than chance (κ = –0.12; 95% CI: –0.19 to –0.04) markedly lower than the other groups. One of the groups with lower counts, there agreement between the two data sources (before or after imputation) on patients being screened, though they do agree on "not screened". Specificity is reasonable (0.88) while Sensitivity is zero (0.00). Each data source indicated that 5 patients were screened, but not the same patients. This lack of agreement among screened strongly affects the kappa statistic.

The Multiple Races group stands out as having higher agreement (pre- and post- imputation) than the other categories (κ = 0.78; 95% CI: 0.36 to 1.00). This group has the lowest number of eligible patients (n = 22) and a low number of patients recorded as receiving this procedure (3 in EHR and 2 in Medicaid Claims data). It just so happens that the data sources agreed on 2 of these otherwise the agreement would have looked more similar to that of the American Indian/Alaskan native group.

The rest of the Race groups and FPL categories all show fair agreement, pre- and post-imputation (Table 18, Table 19, and Figure 20). The differences in the American Indian/Alaskan native group and the Multiple Races group highlight the impact that low

counts can have on agreement and do not indicate that these groups are being screened differently than other groups.

Difference pre- and post- imputation are worth discussing for two categories. Figure 20 shows a notably wider confidence interval for the American Indian/Alaskan native group post-imputation. The number of eligible patients only changed slightly (48 vs. 51), as did Sensitivity (0.00 vs. 0.02) and Specificity (0.88 vs. 0.89). The wider confidence interval and less precision is likely due to between imputation variability.

The second category to note is the group >138% FPL. The confidence interval for this group is narrower post-imputation and we don't see this in any other groups for Colonoscopy. The number of eligibly patients in this group nearly double before and after imputation (49 to 81), but the Sensitivity and Specificity did not change (0.38 and 0.93 respectively). Here, agreement didn't change but the increased number of patients lowered the standard error.

With Colonoscopy, we saw some noticeable difference in agreement between categories, though they could be attributed to low number of eligible patients. Most of the groups showed similar agreement and confidence intervals pre- and post-imputation except for the American Indian/Alaskan native group and >138% FPL. Small counts were seen in these groups with the other procedures. These results suggest that multiple imputation may behave differently with kappa statistics when counts are low and agreement is low pre-imputation.

See Appendix D for tables of results pre- and post- imputation with additional statistics.

Table 18 – Colonoscopy agreement results *prior to imputation*

EHR (+) p1 and Claims (+) p2 indicate proportion of patients listed as screening in the respective data sources. EHR (+), Claims (+) p1+p2 shows the combined percent of total eligible patients screened in both data sources. Sensitivity, here, is the proportion of subjects screened in EHR given that they are screened in Medicaid claims. Specificity is the proportion not screened in EHR given not screened in Medicaid claims. Eligible patients: Men and women age 50 or older with no history of colorectal cancer or total colectomy.

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	3761	7.2	11.5	15.7	0.26	0.95	0.26 [0.21, 0.30]
Race							
Asian/Pacific Islander	256	6.2	11.7	14.4	0.30	0.97	0.34 [0.15, 0.52]
American Indian/Alaskan native	48	10.4	10.4	20.8	0.00	0.88	-0.12 [-0.19, -0.04]
Black	397	6.3	9.3	13.1	0.27	0.96	0.27 [0.11, 0.42]
White	2860	7.4	11.8	16.2	0.26	0.95	0.25 [0.20, 0.31]
Multiple Races	22	13.6	9.1	13.6	1.00	0.95	0.78 [0.36, 1.00]
Unknown	178	4.5	11.2	14.0	0.15	0.97	0.16 [-0.05, 0.37]
Federal poverty level							
<=138% FPL	2906	7.5	11.6	16.0	0.27	0.95	0.26 [0.21, 0.31]
>138% FPL	49	12.2	16.3	22.4	0.38	0.93	0.34 [-0.02, 0.69]
Missing/Unknown	806	5.6	10.8	14.0	0.22	0.96	0.23 [0.13, 0.34]

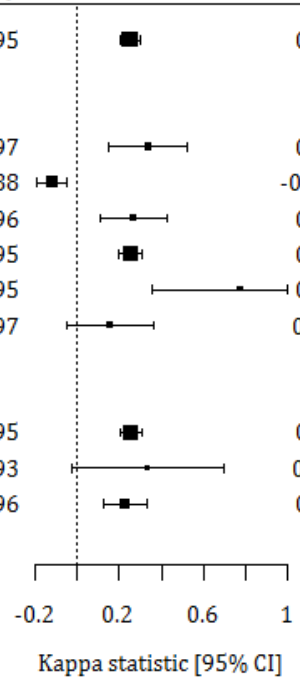


Table 19 – Colonoscopy agreement results *after imputation*

EHR (+) p1 and Claims (+) p2 indicate proportion of patients listed as screening in the respective data sources. EHR (+), Claims (+) p1+p2 shows the combined percent of total eligible patients screened in both data sources. Sensitivity, here, is the proportion of subjects screened in EHR given that they are screened in Medicaid claims. Specificity is the proportion not screened in EHR given not screened in Medicaid claims. Eligible patients: Men and women age 50 or older with no history of colorectal cancer or total colectomy.

	Total eligible patients	EHR (+) p1, %	Claims (+) p2, %	EHR (+), Claims (+) p1+p2, %	Sensitivity	Specificity	Kappa statistic [95% CI]
All eligible patients	3761	7.2	11.5	15.7	0.26	0.95	0.26 [0.21, 0.30]
Race							
Asian/Pacific Islander	267	6.3	11.6	14.5	0.29	0.97	0.33 [0.14, 0.51]
American Indian/Alaskan native	51	10.0	10.9	20.7	0.02	0.89	-0.10 [-0.25, 0.05]
Black	417	6.2	9.3	13.1	0.26	0.96	0.26 [0.11, 0.42]
White	2999	7.3	11.8	16.1	0.26	0.95	0.25 [0.20, 0.30]
Multiple Races	25	12.6	9.5	13.7	0.90	0.95	0.73 [0.27, 1.00]
Federal poverty level							
<=138% FPL	3679	7.1	11.4	15.6	0.26	0.95	0.25 [0.20, 0.30]
>138% FPL	81	11.6	15.2	21.0	0.38	0.93	0.34 [0.02, 0.66]

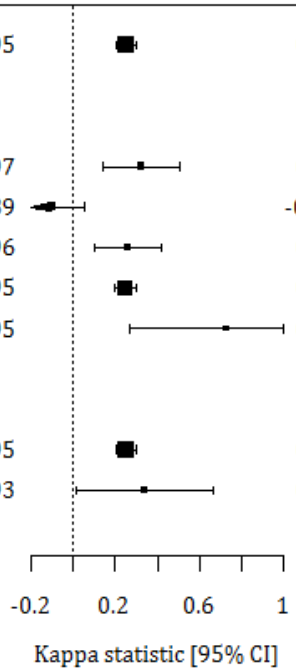
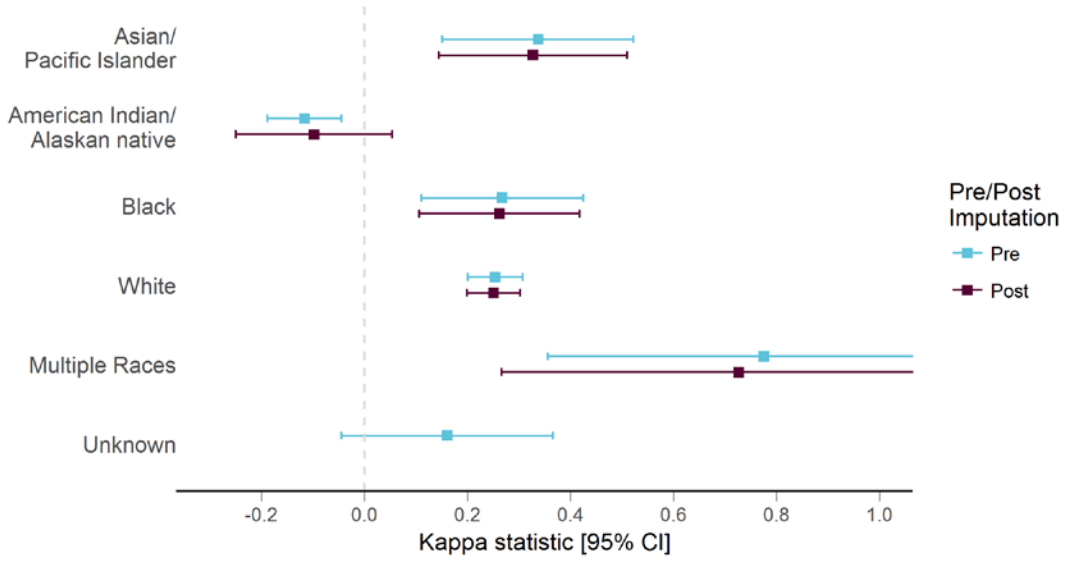
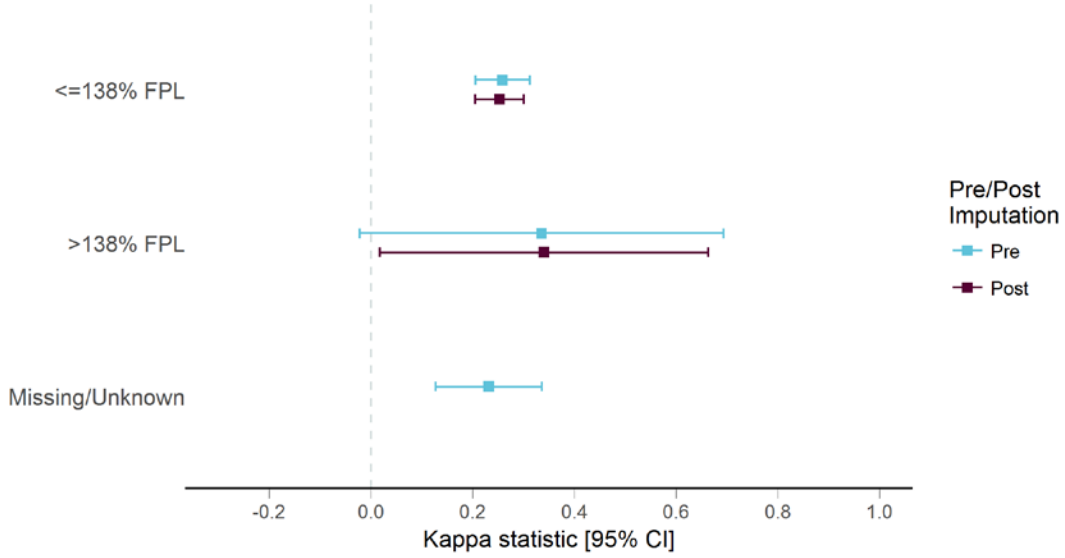


Figure 20 – Colonoscopy: Visualization of kappa statistics and 95% CIs
Pre-imputation shown in blue; post-imputation in purple.

A. By Race



B. By FPL



DISCUSSION

In the previous section, we saw that the post-imputation results were very similar (point estimates and confidence intervals) to the results prior to imputation. This may imply that the assumption about the data mechanism favors MCAR instead of MAR. That is, there is no relationship between whether a data point is missing and any values in the data set, missing or observed. We expect the results of complete case analysis (what was done pre-imputation) to provide the same results as imputation when the data are MCAR. Note, this does not violate the assumptions of multiple imputation since MCAR implies MAR, but not the other way around.

We stated in the introduction that multiple imputation for non-standard statistics like the kappa statistics are not fully developed. We have proceeded with these analyses in earnest effort and due care. Our work in this area has been extensive, and yet we acknowledge that there is still room for further assessment of the methods.

We did not perform sensitivity analysis of the multiply imputed data which is encouraged by Van Buuren [17] to assess the reasonableness of the missing at random (MAR) assumption. The analyses did not reveal any strong reasons to doubt that the data here is MAR. Additionally, our imputation model was carefully constructed using all available data which should be robust to deviations in assumptions.

We performed limited diagnostics on the imputations to assessing the distributions of the imputed data with the observed data. Approaches for future consideration have been

developed by Van Buuren and Groothuis-Oudshoorn [37], by Raghunathan and Bondarenko [45], and by Yucel and Zaslavsky [46].

Rubin's Rules were used to pool the resulting kappa statistics and variances without transformation prior to pooling. There is no guidance provided in literature specifically for kappa statistics. Our results indicate that this approach is reasonable. Future consideration to this would be appropriate when using multiple imputation with kappa statistics.

SUMMARY AND CONCLUSIONS

In this thesis, we sought to explore the use of multiple imputation technique MICE with electronic health records (EHR) and kappa statistics. We also sought to examine the agreement across demographic categories, Race and Federal poverty level (FPL), to assess any systematic differences among patients.

We have shown here a practical execution of the MICE algorithm and associated software to multiply impute missing categorical information in EHR. The flexibility of the approach makes it particularly well suited for EHR due to the many different types of variables that exist in these records. By following the practical guidance developed by Stef Van Buuren, these methods were adapted and successfully applied to a non-standard statistic like Cohen's Kappa.

Both before and after multiple imputation, the examination of agreement stratified between EHR and Medicaid claims by Race and FPL showed no differences that gave cause for concern.

The study results suggest that a similar quality of care is being provided across the demographic categories, judging by agreement between strata.

The 3 procedures (Cholesterol screening, Chlamydia screening, and Colonoscopy) were chosen to provide varying circumstances to assess the application of MICE to kappa statistics across Race and FPL. When number of observations were high and agreement was moderate or better, we saw consistent results before and after multiple imputation. But in the case of Colonoscopy where agreement was less than moderate, in strata with very low counts we saw differences in kappa and confidence intervals pre- and post- imputation. Therefore, we recommend additional scrutiny of the results from multiply imputed data when there is a low number of observations and agreement is low before imputation.

In conclusion, we deem MICE to be an approach worth further investigation and application to obtain statistically valid kappa statistics with incomplete EHR data.

REFERENCES

REFERENCES

[1] J. Heintzman, S. Bailey, M. Hoopes, T. Le, R. Gold, J. P. O'Malley, S. Cowburn, M. Marino, A. Krist and J. E. DeVoe, "Agreement of Medicaid claims and electronic health records for assessing preventive care quality among adults," *J Am Med Inform Assoc,* vol. 24, p. 720–724, 2014.

[2] J. D. Clough and M. McCellan, "Implementing MACRA: Implications for Physicians and for Physician Leadership," *Jounral of the American Medical Association,* vol. 315, no. 22, p. 2397–2398, 2016.

[3] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin,* vol. 76, no. 5, p. 378–382, 1971.

[4] J. L. Fleiss, Statistical Methods for Rates and Proportions, Second ed., John Wiley & sons, Inc., 1981.

[5] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement,* vol. 20, no. 1, p. 37–46, 1960.

[6] A. S. Rigby, "Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient," *Disability and Rehabilitation,* vol. 22, no. 8, p. 339–344, 2000.

[7] J. Sim and C. C. Wright, "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy Journal,* vol. 85, no. 3, p. 257–268, 2005.

[8] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* vol. 33, no. 1, p. 159–174, 1977.

[9] T. Byrt, J. Bishop and J. B. Carlin, "Bias, Prevalence, and Kappa," *Journal of Clinical Epidemiology,* vol. 46, no. 5, p. 423–429, 1993.

[10] A. J. Viera and J. M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic," *Family Medicine,* vol. 37, no. 5, p. 360–363, 2005.

[11] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa I. The problems of two paradoxes," *Journal of Clinical Epidemiology,* vol. 43, no. 6, p. 543–549, 1990.

[12] D. V. Cicchetti and A. R. Feinstein, "High agreement but low kappa: II. Resolving the paradoxes," *Journal of Clinical Epidemiology,* vol. 43, no. 6, p. 551–558, 1990.

[13] F. K. Hoehler, "Bias and prevalence effects of kappa viewed in terms of sensitivity and specificity," *Journal of Clinical Epidemiology,* vol. 53, p. 499–503, 2000.

[14] G. Chen, P. Faris, B. Hemmelgarn, R. L. Walker and H. Quan, "Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa," *BMC Medical Research Methodology,* vol. 9, no. 5, 2009.

[15] P. D. Allison, Missing Data, Thousand Oaks, CA: Sage Publications, 2002.

[16] R. J. Little and e. al., "The Prevention and Treatment of Missing Data in Clinical Trials," *The New England Journal of Medicine,* vol. 367, no. 14, pp. 1355-1360, 2012.

[17] S. van Buuren, Flexible Imputation of Missing Data, Boca Raton, FL: CRC Press, 2012.

[18] R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, Second Edition ed., Hoboken, NJ: Jonh Wiley & Sons, Inc., 2002.

[19] D. B. Rubin, "Inference and Missing Data," *Biometrika,* vol. 63, no. 3, p. 581–592, 1976.

[20] C. K. Enders, "Multiple imputation as a flexible tool for missing data handling in clincal research," *Behaviour Research and Therapy,* 2016.

[21] L. M. Collins, J. L. Schafer and C.-M. Kam, "A comparison of inclusive and restrictive strategies in modern missing data procedures," *Psychological methods,* vol. 6, no. 4, pp. 330-351, 2001.

[22] J. W. Graham, "Missing Data Analysis: Making It Work In the Real World," *Annual Review of Psychology,* vol. 60, p. 549–576, 2009.

[23] M. J. Azur, E. A. Stuart, C. Frangakis and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?," *International Journal of Methods in Psychiatric Research,* vol. 20, no. 1, p. 40–49, 2011.

[24] R. H. Groenwold, I. R. White, A. R. T. Donders, J. R. Carpenter, D. G. Altman and K. G. Moons, "Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis," *Canadian Medical Association Journal,* vol. 184, no. 11, p. 1265–1269, 2012.

[25] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus,* vol. 2, no. 222, 2013.

[26] C. K. Enders and D. L. Bandalos, "The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models," *Educational Psychology Papers and Publications,* no. 64, 2001.

[27] S. R. Seaman and I. R. White, "Review of inverse probability weighting for dealing with missing data," *Statistical Methods in Medical Research,* vol. 22, no. 3, p. 278–295, 2011.

[28] P. Zhang, "Multiple Imputation: Theory and Method," *International Statistical Review,* vol. 71, no. 3, p. 581–592, 2003.

[29] S. Sinharay, H. S. Stern and D. Russell, "The Use of Multiple Imputation for the Analysis of Missing Data," *Psychological Methods,* vol. 6, no. 4, p. 317–329, 2001.

[30] J. L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC, 1997.

[31] P. A. Patrician, "Multiple Imputation for Missing Data," *Research in Nursing & Health,* vol. 25, p. 76–84, 2001.

[32] S. F. Nielsen, "Proper and Improper Multiple Imputation," *International Statistics Review,* vol. 71, no. 3, p. 593–627, 2003.

[33] I. R. White, P. Royston and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Statistics in Medicine,* vol. 30, p. 377–399, 2011.

[34] D. B. Rubin, Mutliple Imputation for Nonreseponse in Surveys, Jonh Wiley & Sons, Inc., 1987, p. 75–77.

[35] J. Barnard and D. B. Rubin, "Small-sample Degrees of Freedom with Multiple Imputation," *Biometrika,* vol. 86, no. 4, p. 948–955, 1999.

[36] S. van Buuren, H. C. Boshuizen and D. L. Knook, "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis," *Statistics in Medicine,* vol. 18, p. 681–694, 1999.

[37] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software,* vol. 45, no. 3, 2011.

[38] S. van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn and D. B. Rubin, "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation,* vol. 76, no. 12, p. 1049–1064, 2006.

[39] S. van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research,* p. 219–242, 2007.

[40] "HealthCare.gov," 2017. [Online]. Available: https://www.healthcare.gov/glossary/federal-poverty-level-FPL/.

[41] J. Honaker, G. King and M. Blackwell, "Amelia II: A Program for Missing Data," *Journal of Statistical Software,* vol. 45, no. 7, p. 1–47, 2011.

[42] A. Kowarik and M. Templ, "Imputation with the R Package VIM," *Journal of Statistical Software,* vol. 74, no. 7, p. 1–16, 2016.

[43] T. E. Bodner, "What improves with increased missing data imputations?," *Structural Equation Modeling: A Multidisciplinary Journal,* vol. 15, no. 4, p. 651–675, 2008.

[44] P. Allison, "Why you probably need more imputations thatn you think.," 2012. [Online]. Available: https://statisticalhorizons.com/more-imputations.

[45] T. Raghunathan and I. Bondarenko, "Diagnostics for Multiple Imputations. Techincal report, Department of Biostatistics, University of Michigan," [Online]. Available: https://ssrn.com/abstract=1031750.

[46] R. M. Yucel, Y. He and A. M. Zaslavsky, "Using Calibration to Improve Rounding in Imputation," *The American Statistician,* vol. 62, no. 2, p. 125–129, 2008.

[47] T. M. Blumenthal D, "The "Meaningful Use" Regulation for Electronic Health Records," *N Engl J Med. 2010 Aug 5;363(6):501-4.*

[48] R. C. Team, *R: A language and environment for statistical computing,* Vienna, Austria: R Foundation for Statistical Computing, 2014.

[49] J. L. Schafer and J. W. Graham, "Missing Data: Our View of the State of the Art," *Psychological Methods,* vol. 7, no. 2, pp. 147-177, 2002.

[50] J. L. Schafer, "Multiple Imputation: a primer," *Statistical Methods in Medical Research,* vol. 8, pp. 3-15, 1999.

ABBREVIATIONS

Abbreviation	Meaning	Page number introduced
EHR	Electronic health records	1
CHC	Community health center	1
MACRA	Medicare Access and CHIP Reauthorization Act	1
MCAR	Missing completely at random	13
MAR	Missing at random	14
MNAR	Missing not at random	14
NMAR	Not missing at random	14
FIML	Full-information maximum likelihood	16
IPW	Inverse probability weighting	17
MI	Multiple imputation	17
MICE	Multivariate imputation by chained equations	24
JM	Joint modeling	25
FCS	Fully conditional specification	25
MCMC	Markov chain Monte Carlo	25
HHS	Department of Health and Human Services	30
ID	Subject identifier	30
PI	Primary investigator	41
QPM	Quick predictor matrix	43
FMI	Fraction of missing information	45
CI	Confidence interval	56

APPENDICES

Appendix A. Inspecting the data

A.1 Output from `mice` command `md.pattern` for full data set

Limited to 3 variables with missing data

```
> cardiac %>%
+   dplyr::select(ethnic.cat, race.cat, fpl.cat) %>%
+   mice::md.pattern(.)
     ethnic.cat race.cat fpl.cat
9706          1        1       1    0
 289          1        0       1    1
2518          1        1       0    1
 392          0        0       1    2
  57          1        0       0    2
 139          0        0       0    3
            531      877    2714 4122
```

There is a lot of information packed into this tables, but for exploring the missing data is does a lot for just a simple command. The 1s in the body of the figure indicate available and zeroes indicate missing. The column on the left shows the number of records with a particular pattern of available/missing data. The last column on the right indicates number of missing columns. The last row tells the number of missing rows by variable and in total.

A.2 Output from `mice` command `md.pattern` for Cholesterol Screening eligible records

```
> cardiac %>%
+   dplyr::filter(ELIG_cholest == 1) %>%
+   dplyr::select(ethnic.cat, race.cat, fpl.cat) %>%
+   mice::md.pattern(.)
     ethnic.cat race.cat fpl.cat
9506          1        1       1    0
 275          1        0       1    1
2470          1        1       0    1
 376          0        0       1    2
  53          1        0       0    2
 137          0        0       0    3
            513      841    2660 4014
```

A.3 Output from `mice` command `md.pattern` for Chlamydia Screening eligible records

```
> cardiac %>%
+   dplyr::filter(ELIG_chlam == 1) %>%
+   dplyr::select(ethnic.cat, race.cat, fpl.cat) %>%
+   mice::md.pattern(.)
    ethnic.cat race.cat fpl.cat
374          1        1       1   0
 33          1        0       1   1
103          1        1       0   1
  8          0        0       1   2
  5          1        0       0   2
             8       46     108 162
```

A.4 Output from `mice` command `md.pattern` for Colonoscopy Screening eligible records

```
> cardiac %>%
+   dplyr::filter(ELIG_colon == 1) %>%
+   dplyr::select(ethnic.cat, race.cat, fpl.cat) %>%
+   mice::md.pattern(.)
     ethnic.cat race.cat fpl.cat
2828          1        1       1    0
  41          1        0       1    1
 755          1        1       0    1
  86          0        0       1    2
  10          1        0       0    2
  41          0        0       0    3
            127      178     806 1111
```

A.5 Output from `mice` command `md.pairs`

Here we are able to see the pairwise relationship of missing and observed between variables. `rr` counts if both are observed; `rm` first variable is observed and the second is missing; `mr` first is missing, second observed; `mm` both missing. So for the pair `race.cat` and `fpl.cat`, there are 9706 completely observed pairs, 2518 pairs where `race.cat` is observed but `fpl.cat` is not, 681 pairs where `race.cat` is missing and `fpl.cat` is observed, and 2714 pairs where both are missing.

```
> cardiac %>%
+   dplyr::select(ethnic.cat, race.cat, fpl.cat) %>%
+   mice::md.pairs(.)
$rr
           ethnic.cat race.cat fpl.cat
ethnic.cat      12570    12224    9995
race.cat        12224    12224    9706
fpl.cat          9995     9706   10387

$rm
           ethnic.cat race.cat fpl.cat
ethnic.cat          0      346    2575
race.cat            0        0    2518
fpl.cat           392      681       0

$mr
           ethnic.cat race.cat fpl.cat
ethnic.cat          0        0     392
race.cat          346        0     681
fpl.cat          2575     2518       0

$mm
           ethnic.cat race.cat fpl.cat
ethnic.cat        531      531     139
race.cat          531      877     196
fpl.cat           139      196    2714
```

A.6 Proportion of usable cases using `mice` command `md.pairs`

Measures how many cases with missing data on the target variable actually have observed values on the predictor. The proportion will be low if both target and predictor are missing on the same cases.

Target on the vertical axis (i.e. left), predictor on the horizontal (i.e. top).

Of the 877 records with missing race, only 39% (346) have observed information on ethnicity.

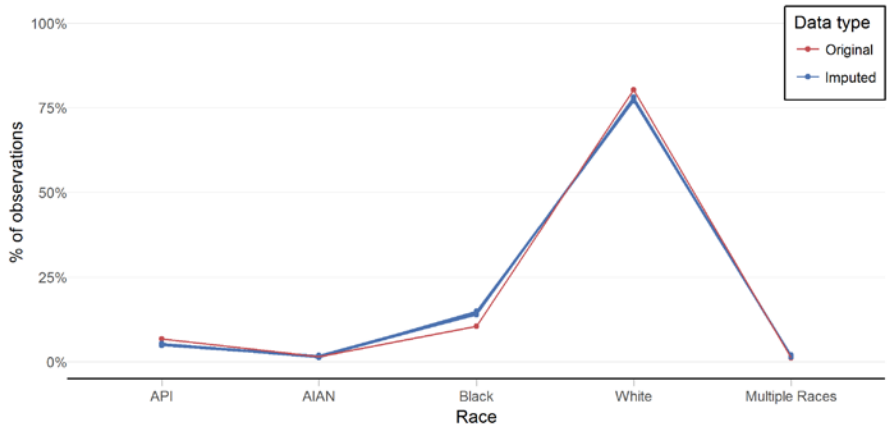
```
> p <- mice::md.pairs(cardiac[, c("ethnic.cat", "race.cat", "fpl.cat")])
>
> round(p$mr / (p$mr + p$mm), digits = 3)
           ethnic.cat race.cat fpl.cat
ethnic.cat      0.000    0.000   0.738
race.cat        0.395    0.000   0.777
fpl.cat         0.949    0.928   0.000
```

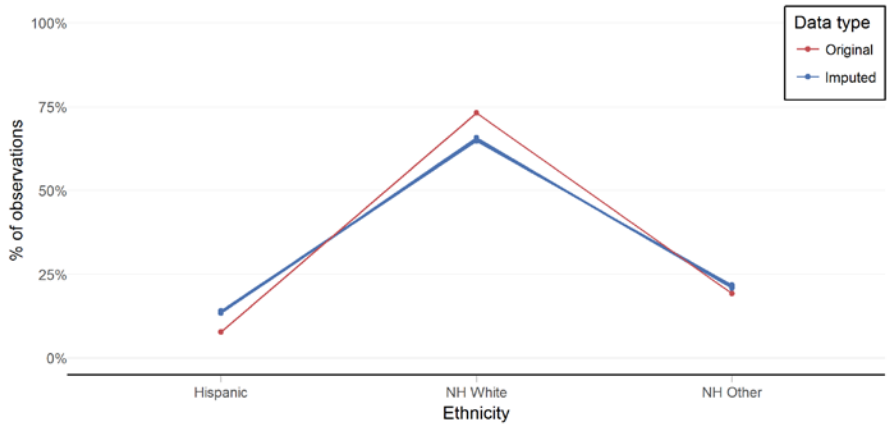
Appendix B. Data Dictionary

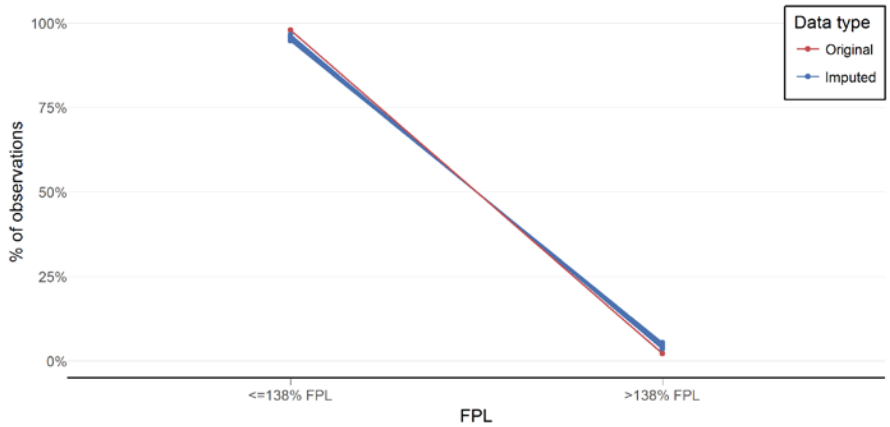
VARIABLE_NAME	DESCRIPTION	VARIABLE TYPE
StudyID	Patient identification	Integer
sex	Sex (M/F)	Categorical, 2 levels
age_start	Age	Integer
FPL_PERCENTAGE	Federal povery level (numerical)	Continuous
PrimaryDept	Primary department/clinic	Categorical, >2 levels
ethnic.cat	Race, ethnicity	Categorical, >2 levels
lang.cat	Language	Categorical, >2 levels
race.cat	Race	Categorical, >2 levels
fpl.cat	Federal povery level (categorical)	Categorical, 2 levels
age.cat	Age (categorical)	Categorical, >2 levels
ELIG_BMI	Eligible for BMI screening (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_breast	Eligible for breast cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_cervical	Eligible for cervical cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_chlam	Eligible for chlamydia screening (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_cholest	Eligible for cholesterol screening (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_colon	Eligible for colon cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_colonoscopy	Eligible for colonoscopy (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_flexsig	Eligible for flexible sigmoidoscopy (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_Flu	Eligible for flu vaccine (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_FOBT	Eligible for FOBT (1 = Y, 0 = N)	Categorical, 2 levels
ELIG_smoking	Eligible for smoking assessment (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_BMI	Medicaid claims, SCREENED BMI screening (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_breast	Medicaid claims, SCREENED breast cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_cervical	Medicaid claims, SCREENED cervical cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_chlam	Medicaid claims, SCREENED chlamydia screening (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_cholest	Medicaid claims, SCREENED cholesterol screening (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_colon	Medicaid claims, SCREENED colon cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_colonoscopy	Medicaid claims, SCREENED colonoscopy (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_flexsig	Medicaid claims, SCREENED flexible sigmoidoscopy (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_Flu	Medicaid claims, SCREENED flu vaccine (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_FOBT	Medicaid claims, SCREENED FOBT (1 = Y, 0 = N)	Categorical, 2 levels
DMAP_smoking	Medicaid claims, SCREENED smoking assessment (1 = Y, 0 = N)	Categorical, 2 levels
EHR_BMI	EHR, SCREENED BMI screening (1 = Y, 0 = N)	Categorical, 2 levels
EHR_breast	EHR, SCREENED breast cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
EHR_cervical	EHR, SCREENED cervical cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
EHR_chlam	EHR, SCREENED chlamydia screening (1 = Y, 0 = N)	Categorical, 2 levels
EHR_cholest	EHR, SCREENED cholesterol screening (1 = Y, 0 = N)	Categorical, 2 levels
EHR_colon	EHR, SCREENED colon cancer screening (1 = Y, 0 = N)	Categorical, 2 levels
EHR_colonoscopy	EHR, SCREENED colonoscopy (1 = Y, 0 = N)	Categorical, 2 levels
EHR_FlexSig	EHR, SCREENED flexible sigmoidoscopy (1 = Y, 0 = N)	Categorical, 2 levels
EHR_Flu	EHR, SCREENED flu vaccine (1 = Y, 0 = N)	Categorical, 2 levels
EHR_FOBT	EHR, SCREENED FOBT (1 = Y, 0 = N)	Categorical, 2 levels
EHR_smoking	EHR, SCREENED smoking assessment (1 = Y, 0 = N)	Categorical, 2 levels
EHR_Weight	EHR, SCREENED weight assessment (1 = Y, 0 = N)	Categorical, 2 levels

Appendix C. Visual diagnostics to assess the imputations for all scenarios

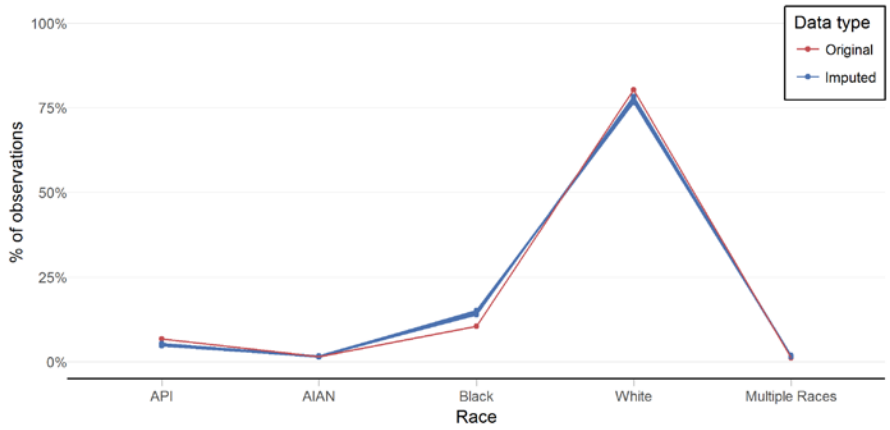
C.1 Full scenario, impute Race, Ethnicity, and FPL

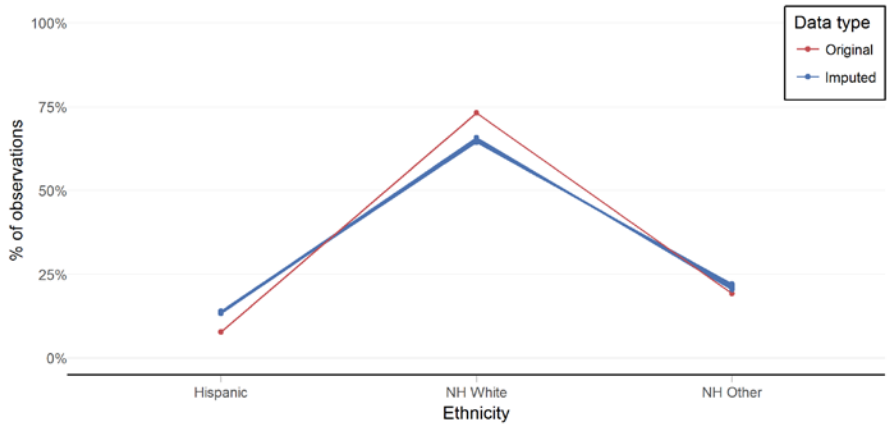


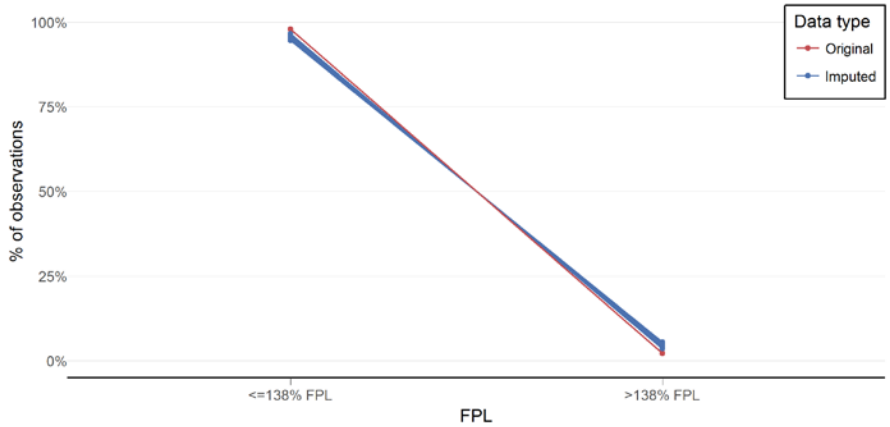




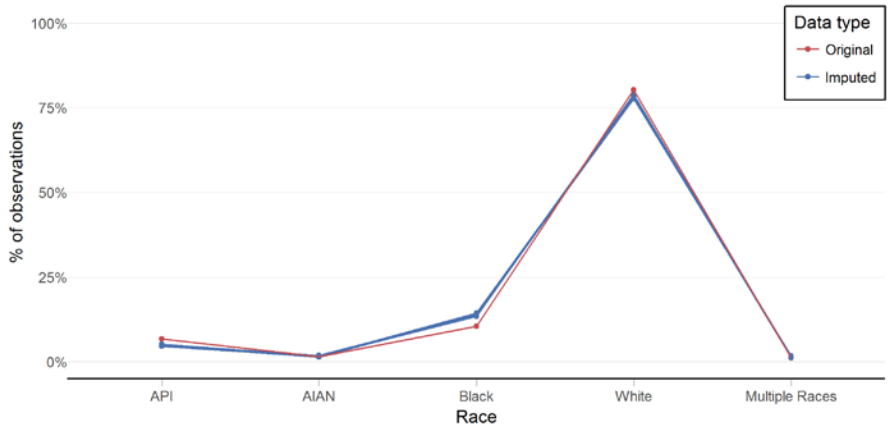
C.2 Reduced scenario, impute Race, Ethnicity, and FPL

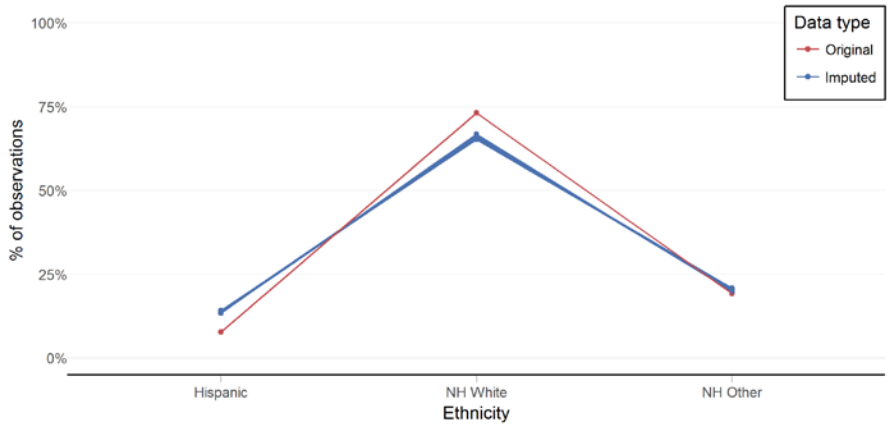


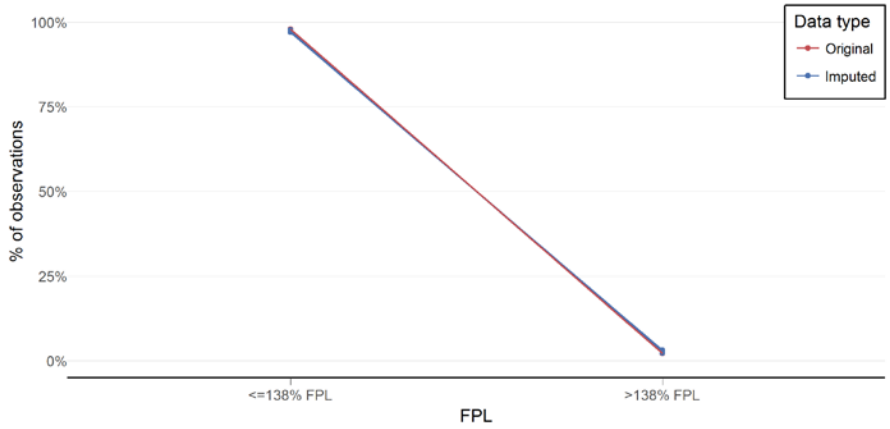




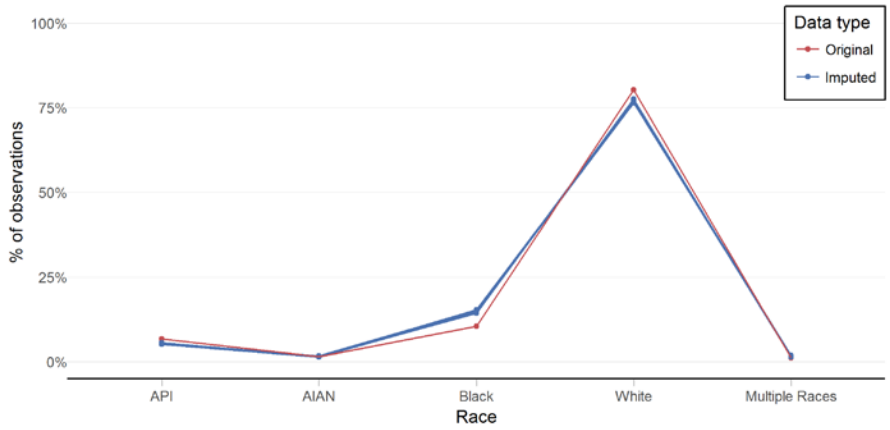
C3 QPM scenario, impute Race, Ethnicity, and FPL

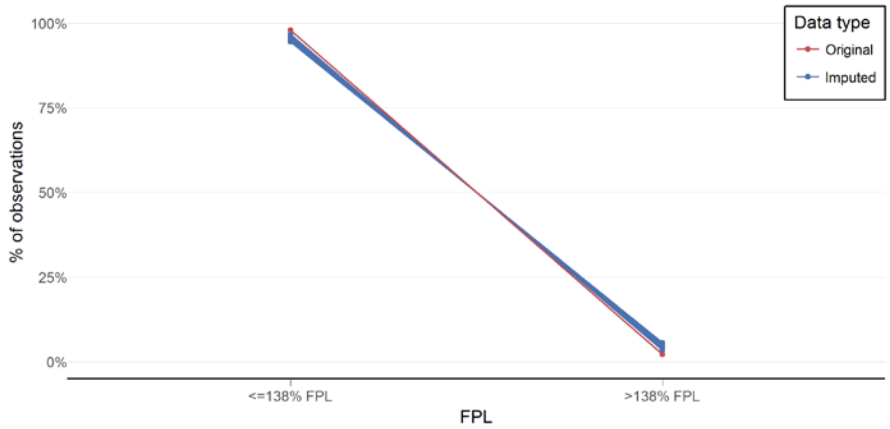




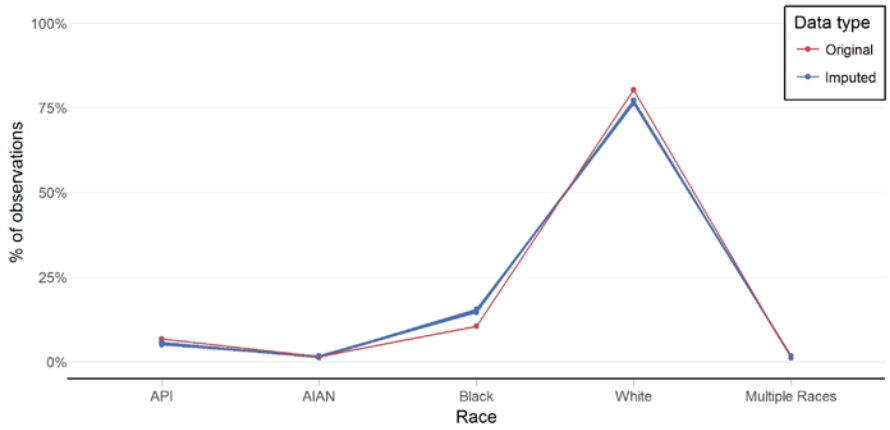


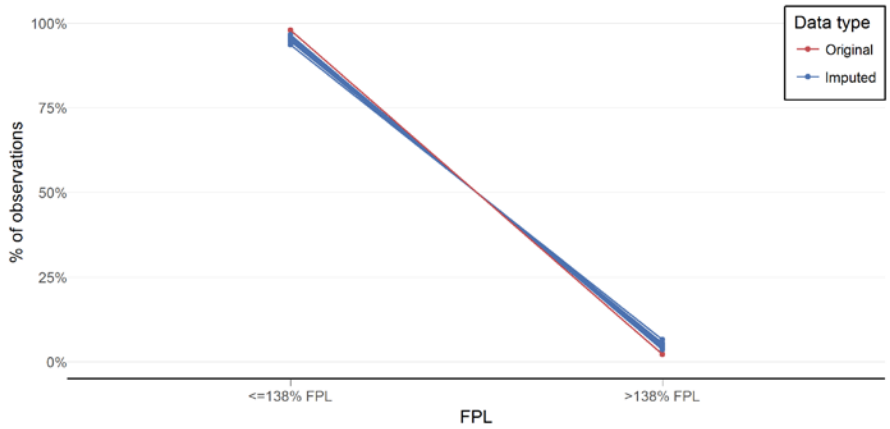
C4 Full scenario, impute Race and FPL



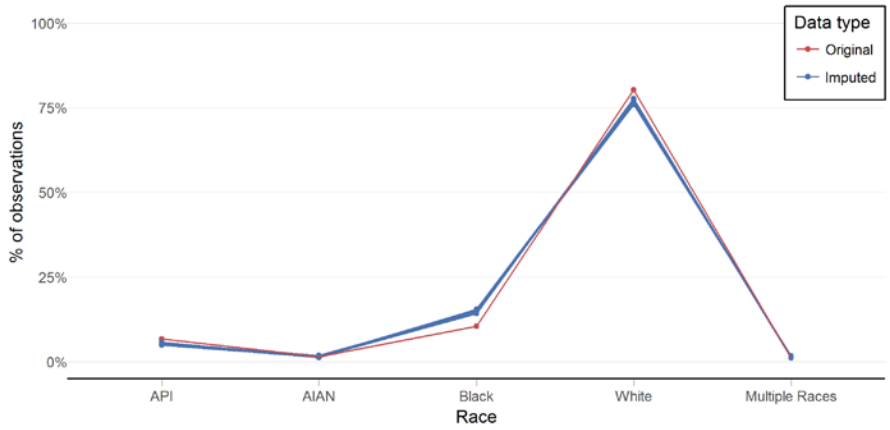


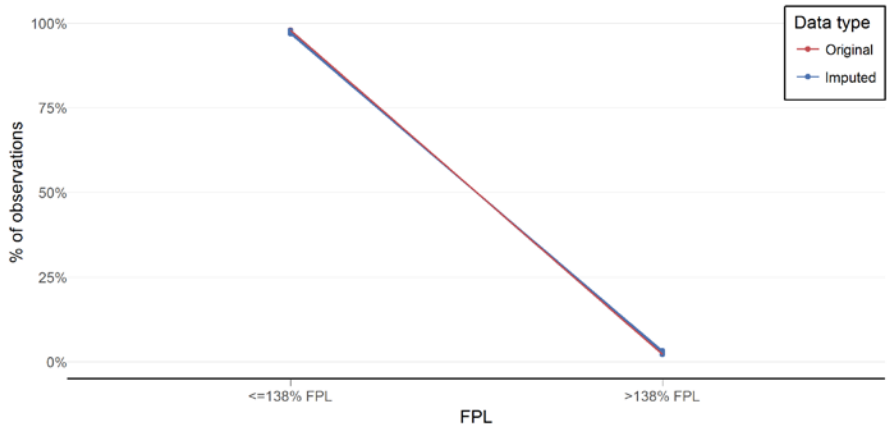
C5 Reduced scenario, impute Race and FPL





C6 QPM scenario, impute Race and FPL





Appendix D. Tables for Results

Starts on next page.

Table 20 – Cholesterol screening results by Race categories

Agreement indices between OCHIN EHR and Medicaid claims data. Post-imputation results shaded gray

	Total eligible patients	a	b	c	d	OCHIN EHR No. (%)	Claims No. (%)	Proportion Correctly Identified by EHR: Screened	Proportion Correctly Identified by EHR: Not Screened	PABAK	κ statistic (95% CI)
All eligible patients	12817	4624	436	776	6981	5060 (39.5)	5400 (42.1)	0.86	0.94	0.81	0.80 (0.79 to 0.81)
Asian / Pacific Islander	756	341	38	18	359	379 (50.1)	359 (47.5)	0.95	0.90	0.85	0.85 (0.81 to 0.89)
	795	357	39	20	377	397 (49.9)	378 (47.5)	0.95	0.90	0.85	0.85 (0.81 to 0.89)
American Indian / Alaskan native	175	56	3	21	95	59 (33.7)	77 (44.0)	0.73	0.97	0.73	0.71 (0.61 to 0.82)
	190	61	3	22	103	65 (33.9)	84 (43.8)	0.73	0.97	0.73	0.72 (0.62 to 0.82)
Black	1388	517	63	43	765	580 (41.8)	560 (40.3)	0.92	0.92	0.85	0.84 (0.81 to 0.87)
	1483	553	67	48	813	621 (41.9)	602 (40.6)	0.92	0.92	0.84	0.84 (0.81 to 0.87)
White	9518	3365	289	639	5225	3654 (38.4)	4004 (42.1)	0.84	0.95	0.81	0.80 (0.79 to 0.81)
	10191	3610	320	670	5590	3931 (38.6)	4281 (42.0)	0.84	0.95	0.81	0.80 (0.79 to 0.81)
Multiple Races	139	36	4	14	85	40 (28.8)	50 (36.0)	0.72	0.96	0.74	0.71 (0.58 to 0.83)
	155	41	4	14	95	46 (29.5)	56 (35.8)	0.74	0.95	0.75	0.72 (0.60 to 0.84)
Unknown	841	309	39	41	452	348 (41.4)	350 (41.6)	0.88	0.92	0.81	0.80 (0.76 to 0.84)
	0	0	0	0	0	0 (NA)	0 (NA)	NA	NA	NA	NA (NA to NA)

Eligible patients: Men and women aged ≥20; cholesterol screening includes low density lipoprotein, high density lipoprotein, total cholesterol, and triglycerides.

Table 21 – Cholesterol screening results by FPL categories

Agreement indices between OCHIN EHR and Medicaid claims data. Post-imputation results shaded gray

	Total eligible patients	a	b	c	d	OCHIN EHR No. (%)	Claims No. (%)	Proportion Correctly Identified by EHR: Screened	Proportion Correctly Identified by EHR: Not Screened	PABAK	κ statistic (95% CI)
All eligible patients	12817	4624	436	776	6981	5060 (39.5)	5400 (42.1)	0.86	0.94	0.81	0.80 (0.79 to 0.81)
<=138% FPL	9930	3779	345	536	5270	4124 (41.5)	4315 (43.5)	0.88	0.94	0.82	0.82 (0.81 to 0.83)
	12464	4521	423	759	6760	4945 (39.7)	5280 (42.4)	0.86	0.94	0.81	0.80 (0.79 to 0.81)
>138% FPL	227	72	8	8	139	80 (35.2)	80 (35.2)	0.90	0.95	0.86	0.85 (0.77 to 0.92)
	352	102	12	16	220	115 (32.7)	120 (34.0)	0.86	0.95	0.84	0.82 (0.74 to 0.89)
Missing/Unknown	2660	773	83	232	1572	856 (32.2)	1005 (37.8)	0.77	0.95	0.76	0.74 (0.71 to 0.77)
	0	0	0	0	0	0 (NA)	0 (NA)	NA	NA	NA	NA (NA to NA)

Eligible patients: Men and women aged ≥20; cholesterol screening includes low density lipoprotein, high density lipoprotein, total cholesterol, and triglycerides.

Table 22 – Chlamydia screening results by Race categories

Agreement indices between OCHIN EHR and Medicaid claims data. Post-imputation results shaded gray

	Total eligible patients	a	b	c	d	OCHIN EHR No. (%)	Claims No. (%)	Proportion Correctly Identified by EHR: Screened	Proportion Correctly Identified by EHR: Not Screened	PABAK	Kappa statistic [95% CI]
All eligible patients	523	183	41	85	214	224 (42.8)	268 (51.2)	0.68	0.84	0.52	0.52 [0.45 to 0.59]
Asian / Pacific Islander	18	9	1	3	5	10 (55.6)	12 (66.7)	0.75	0.83	0.56	0.54 [0.15 to 0.93]
	18	9	1	3	5	10 (55.4)	12 (66.2)	0.75	0.84	0.57	0.55 [0.17 to 0.93]
American Indian / Alaskan native	7	1	1	0	5	2 (28.6)	1 (14.3)	1.00	0.83	0.71	0.59 [-0.09 to 1.00]
	7	1	1	0	5	2 (28.7)	1 (15.4)	1.00	0.84	0.73	0.61 [-0.04 to 1.00]
Black	70	38	3	12	17	41 (58.6)	50 (71.4)	0.76	0.85	0.57	0.54 [0.34 to 0.74]
	72	39	3	12	18	43 (58.6)	52 (70.9)	0.77	0.85	0.58	0.55 [0.36 to 0.75]
White	365	107	33	65	160	140 (38.4)	172 (47.1)	0.62	0.83	0.46	0.46 [0.37 to 0.55]
	405	125	35	67	176	161 (39.8)	193 (47.7)	0.65	0.83	0.49	0.48 [0.40 to 0.57]
Multiple Races	17	7	0	2	8	7 (41.2)	9 (52.9)	0.78	1.00	0.76	0.77 [0.47 to 1.00]
	18	7	0	2	8	8 (41.2)	10 (52.3)	0.78	0.99	0.76	0.76 [0.46 to 1.00]
Unknown	46	21	3	3	19	24 (52.2)	24 (52.2)	0.88	0.86	0.74	0.74 [0.54 to 0.93]
	0	0	0	0	0	0 (NA)	0 (NA)	NA	NA	NA	NA [NA to NA]

Eligible patients: Sexually active women aged 19-24.

Table 23 – Chlamydia screening results by FPL categories

Agreement indices between OCHIN EHR and Medicaid claims data. Post-imputation results shaded gray

	Total eligible patients	a	b	c	d	OCHIN EHR No. (%)	Claims No. (%)	Proportion Correctly Identified by EHR: Screened	Proportion Correctly Identified by EHR: Not Screened	PABAK	Kappa statistic [95% CI]
All eligible patients	523	183	41	85	214	224 (42.8)	268 (51.2)	0.68	0.84	0.52	0.52 [0.45 to 0.59]
<=138% FPL	401	145	31	61	164	176 (43.9)	206 (51.4)	0.70	0.84	0.54	0.54 [0.46 to 0.62]
	502	176	39	84	202	215 (42.9)	260 (51.8)	0.68	0.84	0.51	0.51 [0.44 to 0.59]
>138% FPL	14	6	1	0	7	7 (50.0)	6 (42.9)	1.00	0.88	0.86	0.86 [0.59 to 1.00]
	20	6	1	0	11	9 (42.3)	8 (37.8)	0.90	0.86	0.75	0.74 [0.38 to 1.00]
Missing/Unknown	108	32	9	24	43	41 (38.0)	56 (51.9)	0.57	0.83	0.39	0.39 [0.23 to 0.56]
	0	0	0	0	0	0 (NA)	0 (NA)	NA	NA	NA	NA [NA to NA]

Eligible patients: Sexually active women aged 19-24.

Table 24 – Colonoscopy results by Race categories

Agreement indices between OCHIN EHR and Medicaid claims data. Post-imputation results shaded gray

	Total eligible patients	a	b	c	d	OCHIN EHR No. (%)	Claims No. (%)	Proportion Correctly Identified by EHR: Screened	Proportion Correctly Identified by EHR: Not Screened	PABAK	Kappa statistic [95% CI]
All eligible patients	3761	113	157	320	3171	270 (7.2)	433 (11.5)	0.26	0.95	0.75	0.26 [0.21 to 0.30]
Asian / Pacific Islander	256	9	7	21	219	16 (6.3)	30 (11.7)	0.30	0.97	0.78	0.34 [0.15 to 0.52]
	267	9	7	21	228	17 (6.3)	31 (11.6)	0.29	0.97	0.78	0.33 [0.14 to 0.51]
American Indian / Alaskan native	48	0	5	5	38	5 (10.4)	5 (10.4)	0.00	0.88	0.58	-0.12 [-0.19 to -0.04]
	51	0	5	5	40	5 (10.0)	6 (10.9)	0.02	0.89	0.59	-0.10 [-0.25 to 0.05]
Black	397	10	15	27	345	25 (6.3)	37 (9.3)	0.27	0.96	0.79	0.27 [0.11 to 0.42]
	417	10	15	28	363	26 (6.2)	39 (9.3)	0.26	0.96	0.79	0.26 [0.11 to 0.42]
White	2860	89	124	250	2397	213 (7.4)	339 (11.9)	0.26	0.95	0.74	0.25 [0.20 to 0.31]
	2999	91	127	263	2517	219 (7.3)	355 (11.8)	0.26	0.95	0.74	0.25 [0.20 to 0.30]
Multiple Races	22	2	1	0	19	3 (13.6)	2 (9.1)	1.00	0.95	0.91	0.78 [0.36 to 1.00]
	25	2	1	0	21	3 (12.5)	2 (9.5)	0.90	0.95	0.89	0.73 [0.27 to 1.00]
Unknown	178	3	5	17	153	8 (4.5)	20 (11.2)	0.15	0.97	0.75	0.16 [-0.05 to 0.37]
	0	0	0	0	0	0 (NA)	0 (NA)	NA	NA	NA	NA [NA to NA]

Eligible patients: Men and women aged ≥50 with no history of colorectal cancer or total colectomy.

Table 25 – Colonoscopy results by FPL categories

Agreement indices between OCHIN EHR and Medicaid claims data. Post-imputation results shaded gray

	Total eligible patients	a	b	c	d	OCHIN EHR No. (%)	Claims No. (%)	Proportion Correctly Identified by EHR: Screened	Proportion Correctly Identified by EHR: Not Screened	PABAK	κ statistic (95% CI)
All eligible patients	3761	113	157	320	3171	270 (07.2)	433 (11.5)	0.26	0.95	0.75	0.26 (0.21 to 0.30)
<=138% FPL	2906	91	128	247	2440	219 (07.5)	338 (11.6)	0.27	0.95	0.74	0.26 (0.21 to 0.31)
	3679	108	152	312	3106	261 (07.1)	421 (11.4)	0.26	0.95	0.75	0.25 (0.20 to 0.30)
>138% FPL	49	3	3	5	38	6 (12.2)	8 (16.3)	0.38	0.93	0.67	0.34 (-0.02 to 0.69)
	81	4	4	7	64	10 (11.6)	12 (15.1)	0.38	0.93	0.70	0.34 (0.02 to 0.66)
Missing/Unknown	806	19	26	68	693	45 (05.6)	87 (10.8)	0.22	0.96	0.77	0.23 (0.13 to 0.34)
	0	0	0	0	0	0 (NA)	0 (NA)	NA	NA	NA	NA (NA to NA)

Eligible patients: Men and women aged ≥50 with no history of colorectal cancer or total colectomy.